

NASA CR - 167,780

NASA-CR-167780
19830007503

A Reproduced Copy

OF

NASA CR - 167,780

Reproduced for NASA

by the

NASA Scientific and Technical Information Facility

LIBRARY COPY

MAY 4 - 1983

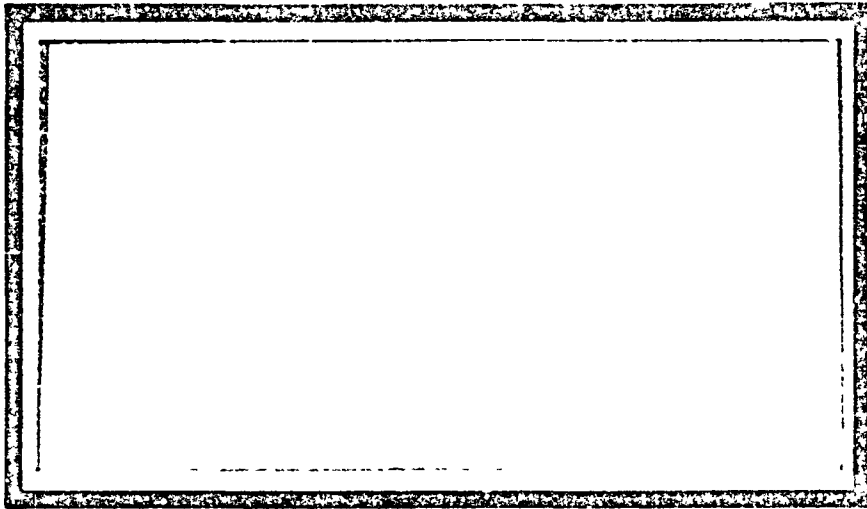
LANGLEY RESEARCH CENTER
LIBRARY, NASA
HAMPTON, VIRGINIA

FFNo 672 Aug 65



NF01438

AKA 11. 11. 1980



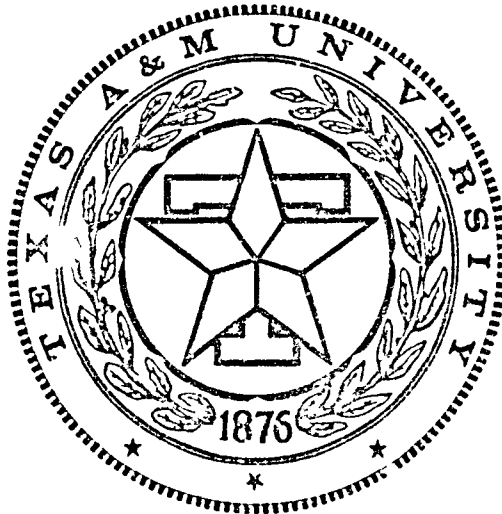
"Made available under NASA sponsorship
in the interest of early and wide dis-
semination of Earth Resources Survey
Program information and without liability
for any use made thereof."

(E83-10119) PROCEEDINGS OF THE NASA
WORKSHOP ON DENSITY ESTIMATION AND FUNCTION
SMOOTHING (Texas A&M Univ.) 491 p
HC A21/MF A01

CSCI. 05B

N83-15774
THRU
N83-15789
Unclas
00119

63/43



DEPARTMENT OF MATHEMATICS

TEXAS A&M UNIVERSITY
COLLEGE STATION TEXAS

N83-15774 #
+HAW
N83-15789 #

PROCEEDINGS OF THE
NASA WORKSHOP ON DENSITY ESTIMATION
AND FUNCTION SMOOTHING

Texas A&M University
College Station, Texas
March 11-13, 1982

Prepared for

Earth Resources Research Division
NASA/Johnson Space Center
Houston, Texas 77058

by

L. F. Guseman, Jr.
Principal Investigator
Department of Mathematics
Texas A&M University
College Station, Texas 77843

under

NASA Contract NAS 9-16447

"Studies in Mathematical Pattern Recognition
and Image Analysis"

TABLE OF CONTENTS

Introduction--Emanuel Parzen	1
Agenda	5
Speakers	7
Other Participants	8
Papers:	
Topics in Global Convergence of Density Estimates--Luc Devroye	9
Cross-Validation for Densities and Regressions--Stuart Geman	20
Estimation of Planar Sets from Poisson Projections--Donald E. McClure	32
Characterization of a Maximum-Likelihood Nonparametric Density Estimator of Kernel Type--Stuart Geman and Donald McClure	38
Remote Sensing of Temperature Profiles in the Atmosphere--Finbarr O'Sullivan	48
Quantiles, Parametric-Select Density Estimation, and Bi-Information Parameter Estimators--Emanuel Parzen	60
Consistency and Other Large Sample Properties of Maximum Likelihood Estimates of Mixture Parameters--Charles Peters	85
Smoothing Splines: Regression, Derivatives and Deconvolution--John Rice and Murray Rosenblatt	102
Considerations in Cross-Validation Type Density Smoothing With a Look at Some Data--Eugene F. Schuster	142
Review of Some Results in Bivariate Density Estimation--David W. Scott	165
A Bootstrap Approach to Bump Hunting--B. W. Silverman	195
A Data Based Random Number Generator for a Multivariate Distribution (Using Stochastic Interpolation)--James R. Thompson and Malcolm S. Taylor	214
Mixture Densities, Maximum Likelihood, and the EM Algorithm--Richard A. Redner and Homer F. Walker	226

Regression Methods for Spatial Data--S. J. Yakowitz and F. Szidarovszky	343
Estimation of Divergence and Vorticity Using Multidimensional Smoothing Splines--James G. Wendelberger	386
The Computation of Laplacian Smoothing Splines with Examples--James G. Wendelberger	407
Some New Mathematical Methods for Variational Objective Analysis Using Splines and Cross Validation--Grace Wahba and James Wendelberger	475



by
Emanuel Parzen
Texas A&M University

The workshop on "Density Estimation and Function Smoothing" held at Texas A&M University on March 11-13, 1982 under the sponsorship of NASA, provided the occasion for a cross-section of mathematical scientists involved in this field to meet for an intensive sharing of results and viewpoints. All participants regarded the workshop as an unusually warm, stimulating, and productive experience. The papers collected in this volume provide written versions of the papers presented, enabling a wide audience to enjoy the excitement experienced at the workshop in being able to learn about the diverse research directions that constitute the current state of the art in the statistical discipline of density estimation and function smoothing.

One conclusion to be drawn from these papers is that solutions to problems of density estimation and function smoothing involve aspects of theoretical and applied mathematics, probability and statistics, numerical analysis and computer science, information theory and approximation theory, as well as the scientific fields such as meteorology and remote sensing. I believe this field of mathematical science merits a name of its own, and I propose "statistical functional inference." I believe that statistical model identification techniques are required to develop and implement workable practical solutions to problems in density estimation and function smoothing. There is reason to believe that the techniques being developed by the workshop participants will ultimately prove to be of great value in accomplishing the objectives of NASA.

The papers collected here are extremely rich in content, and it is impossible to convey their importance in a few summary sentences. Nevertheless, to help the reader obtain an overview of each paper I have written a short description of each.

Devroy takes a critical look at mathematical results on the convergence of estimators of a probability density f on R^d from a random sample x_1, \dots, x_n .

Geman provides insight about the problem of choosing a smoothing parameter by cross-validation.

McClure discusses estimation of a planar convex region from projections of counts of events which are Poisson distributed at different rates inside and outside the region.

Geman and McClure relate kernel type density estimators to maximum likelihood density estimators calculated by the method of sieves.

O'Sullivan discusses how methods of regularized and generalized cross-validation can be used to estimate the atmosphere's temperature, moisture, and wind structure from a finite number of noisy measurements by meteorological satellites on the intensity of upwelling radiation in selected channel frequencies.

Parzen presents an approach to statistical data science based on quantile functions, density-quantile functions, and information and entropy measures. He outlines a new approach to density estimation based on using exponential probability densities as exact and approximate models.

Peters discusses, for a probability model of a finite mixture of multivariate distributions, the asymptotic consistency, normality, and efficiency of the maximum likelihood estimators of the parameters of this

model.

An important technique of estimating a smooth function $g(t)$ given data values x_i , $i = 1, \dots, n$ which are noisy measurements of $A g(t_i)$, for a known linear operator A , is to choose g to minimize

$$\frac{1}{n} \sum_{i=1}^n \{x_i - A g(t_i)\}^2 + \lambda \int_0^1 |g''(t)|^2 dt$$

Rice and Rosenblatt examine this procedure in the cases of numerical differentiation and deconvolution.

Schuster summarizes results reported in several papers by Schuster and Gregory on their experience in applying non-parametric maximum likelihood techniques of density estimation to judge the comparative quality of various estimators.

Scott summarizes his experience in comparing the effects of smoothing parameters on probability density estimators for univariate and bivariate data.

Silverman introduces, and discusses the asymptotic behavior of, a test statistic for hypotheses concerning the number of modes in a probability density.

Thompson introduces a method for generating random vectors from the distribution of a random vector x which is based on a random sample of x without estimating the underlying density.

Redner and Walker review the theory of estimation of parameters of mixture density models, and discuss in detail iterative procedures for numerical approximation of maximum likelihood estimates based on the EM algorithm.

Yakowitz and St'edrovsky provide a comprehensive review of "kriging" methods for fitting functions to spatial data.

Wendelberger discusses multidimensional smoothing splines, the method of generalized cross-validation, and applications to meteorology.

NASA WORKSHOP ON DENSITY ESTIMATION
AND FUNCTION SMOOTHING

Texas A&M University
March 11-13, 1982
Room 5i0, Rudder Tower

Thursday, March 11:

- 8:15 Coffee and donuts
- 8:30 Remote Sensing Fundamental Research Program: An Overview
R. B. MacDonald, NASA/Johnson Space Center
L. F. Guseman, Jr., Texas A&M University
- 9:20 Topics in Global Convergence of Density Estimates
Luc Devroye, McGill University
- 10:10 Coffee
- 10:30 Smoothing Splines: Regression, Derivatives, and Deconvolution
John Rice, University of California at San Diego
- 11:20 On Statistics and Density Estimation
Herbert Robbins, Columbia University
- 12:00 Lunch
- 1:40 A Bootstrap Approach to Bump Hunting
B. W. Silverman, University of Bath
- 2:30 Cross Validation for Densities and Regressions
Stu Geman, Brown University
- 3:20 Coffee
- 3:40 Estimation of Planar Sets from Poisson Projections
Donald McClure, Brown University
- 4:30 Quantiles, Parametric-Select Density Estimation, and Bi-Information
Density Estimators
Emanuel Parzen, Texas A&M University
- 6:15 Assemble in Aggieldand Inn Lobby for Cocktails & Dinner

Friday, March 12:

- 8:15 Coffee and donuts
- 8:30 Considerations in Cross Validation Type Density Smoothing with
a Look at Some Data
Eugene F. Schuster, University of Texas at El Paso

- 9:20 Nonparametric Regression and Kriging Methods for Spatial Data
Sidney Yakowitz, University of Arizona
- 10:10 Coffee
- 10:30 Multidimensional Smoothing Splines and Their Restriction to the
Sphere
Jim Wendelberger, University of Wisconsin--Madison
- 11:20 Remote Sensing of Temperature Profiles in the Atmosphere
Finbarr O'Sullivan, University of Wisconsin--Madison
- 12:00 Lunch
- 1:40 A Data Based Random Number Generator for a Multivariate Distribution
J. R. Thompson, Rice University
- 2:30 Review of Some Results in Bivariate Density Estimation
David Scott, Rice University
- 3:20 Coffee
- 3:40 Consistency and Other Large Sample Properties of Maximum Likelihood
Estimates of Mixture Parameters
B. Charles Peters, University of Houston
- 4:30 Mixture Densities, Maximum Likelihood, and the Em Algorithm
Homer Walker, University of Houston

Saturday, March 13:

- 9:00 Group discussion to define research areas critical to NASA
- 10:00 Coffee
- 10:30 Group discussion, cont'd.
- 12:00 Adjourn

NASA WORKSHOP ON DENSITY ESTIMATION AND FUNCTION SMOOTHING

Speakers:

Professor Luc Devroye
School of Computer Science
McGill University
805 Sherbrooke Street West
Montreal, Canada H3A 2K6

Professor Stu Geman
Division of Applied Mathematics
Brown University
Providence, Rhode Island 02913

Professor L. F. Guseman, Jr.
Department of Mathematics
Texas A&M University
College Station, Texas 77843

Mr. R. B. MacDonald
Chief, Earth Resources Research Division
NASA/Johnson Space Center
Mail Code SG
Houston, Texas 77058

Professor Donald McClure
Division of Applied Mathematics
Brown University
Providence, Rhode Island 02913

Finbarr O'Sullivan
Department of Statistics
University of Wisconsin--Madison
1210 West Dayton Street
Madison, Wisconsin 53706

Professor Emanuel Parzen
Institute of Statistics
Texas A&M University
College Station, Texas 77843

Professor B. Charles Peters
Department of Mathematics
University of Houston
Houston, Texas 77004

Professor John Rice
Mathematics Department
University of California, San Diego
La Jolla, California 92093

Professor Herbert Robbins
Department of Mathematical Statistics
Columbia University
New York, New York 10027

Professor Eugene F. Schuster
Mathematics Department
University of Texas, El Paso
El Paso, Texas 79968

Professor David Scott
Mathematical Sciences Department
Rice University
P. O. Box 1892
Houston, Texas 77001

Dr. B. W. Silverman
School of Mathematics
University of Bath
Claverton Down
Bath BA 2 7 AY, England

Professor J. R. Thompson
Mathematical Sciences Department
Rice University
P. O. Box 1892
Houston, Texas 77001

Professor Homer Walker
Mathematics Department
University of Houston
Houston, Texas 77004

Jim Wendelberger
Department of Statistics
University of Wisconsin--Madison
1210 West Dayton Street
Madison, Wisconsin

Professor Sidney Yakowitz
Systems Engineering Department
University of Arizona
Tucson, Arizona 85721

NASA WORKSHOP ON DENSITY ESTIMATION AND FUNCTION SMOOTHING

Other Participants:

Mr. Taskin Atilgan
Department of Statistics
University of Wisconsin--Madison
1210 West Dayton Street
Madison, Wisconsin 53706

Professor Jack Bryant
Department of Mathematics
Texas A&M University
College Station, Texas 77843

Professor Philip Cheng
Department of Mathematics
University of Houston
Houston, Texas 77004

Dr. Richard P. Heydorn
Earth Observation Division
NASA/Johnson Space Center
Mail Code SF3
Houston, Texas 77058

Mr. Kent Lenington
Lockheed
1830 NASA Road 1
Houston, Texas 77058

Professor Richard Redner
Division of Mathematical Sciences
University of Tulsa
Tulsa, Oklahoma 74104

Professor Larry L. Schumaker
Department of Mathematics
Texas A&M University
College Station, Texas 77843

Ms. Sylvia Shen
Lockheed
1830 NASA Road 1
Houston, Texas 77058

Mr. C. Sorensen
Lockheed
NASA Road 1
Houston, Texas 77058

N83

15775

UNCLAS

TOPICS IN GLOBAL CONVERGENCE OF DENSITY ESTIMATES

by

Luc Devroye
 McGill University

SUMMARY.

We take a critical look at the problem of estimating a density f on R^d from a sample X_1, \dots, X_n of independent identically distributed random vectors, and review some recent results in the field. Among other things, we will qualify the following statements :

- (i) For any sequence of density estimates f_n , any arbitrary slow rate of convergence to 0 is possible for $E(\int |f_n - f|)$.
- (ii) In theoretical comparisons of density estimates, one should use $\int |f_n - f|$ and not $\int |f_n - f|^p$, $p > 1$.
- (iii) For most reasonable nonparametric density estimates, either we have convergence of $\int |f_n - f|$ (and then the convergence is in the strongest possible sense for all f), or we have no convergence (and then we don't even have convergence in the weakest possible sense for a single f). There is no intermediate situation.

* Research of the author was supported by NSERC Grant A3455. The author is with the School of Computer Science, McGill University, 805 Sherbrooke Street West, Montreal, Canada H3A 2K6.

1. INTRODUCTION.

In this paper, we discuss various issues related to the problem of estimating a density on R^d from a sample X_1, \dots, X_n of independent identically distributed random vectors having density f , such as : how should one judge the goodness of an estimate; is there an optimal estimate; how good can estimates be for small n and large n ; and does it pay to use sophisticated estimates ? The discussion will be supplemented with a selected survey of recent results in the field.

A density estimate is a sequence $f_1, f_2, \dots, f_n, \dots$ where for each n ,

$$f_n(x) = f_n(x, X_1, \dots, X_n)$$

is a real-valued Borel measurable function of $x \in R^d$ and the data X_1, \dots, X_n . A density estimate can be parametric or nonparametric, but this distinction is not important in what follows. The prototype parametric estimate is defined as follows for $d=1$:

$$f_n \text{ is normal } (\hat{\mu}, \hat{\sigma}^2), \hat{\mu} = \frac{1}{n} \sum_{i=1}^n X_i, \hat{\sigma}^2 = \frac{1}{n} \sum_{i=1}^n (X_i - \hat{\mu})^2. \tag{1}$$

The most frequently used nonparametric estimate is the kernel estimate (Rosenblatt (1956) and Parzen (1962)) :

$$f_n(x) = \frac{1}{n} \sum_{i=1}^n h^{-d} K((X_i - x)/h), \tag{2}$$

$h > 0$ is a number depending upon n ,
 K is a given density (kernel).

For bibliographies on density estimation, see Wegman (1972), Wertz (1978), Wertz and Schneider (1979) and Bean and Tsokos (1980).

MEASURES OF GOODNESS
OF POOR QUALITY

2. MEASURES OF GOODNESS.

We would like to obtain a number that measures how close f_n is to f in order to carry out theoretical comparisons between estimates later on. For a variety of reasons, but mostly for the sake of convenience, researchers have proposed the criterion

$$\int (f_n - f)^2 . \quad (3)$$

All integrals in this paper are with respect to Lebesgue measure (dx) . Note that (3) is a random variable, and that it is necessary to take its expected value. In general, we can consider all integral measures of goodness :

$$E(\int |f_n - f|^p) , p \geq 1.$$

We will now argue that the only reasonable integral measure is the L_p measure with $p=1$. Our argument is based on a couple of observations.

1. Let g be an estimate of f . If X has density f on R^1 , then aX has density $\frac{1}{a}f(\frac{x}{a})$, and this density should be approximated by $\frac{1}{a}g(\frac{x}{a})$. But

$$\int |\frac{1}{a}f(\frac{x}{a}) - \frac{1}{a}g(\frac{x}{a})|^p = a^{1-p} \int |f-g|^p .$$

Thus, the only L_p measure that is independent of the scale is the L_1 measure.

2. By Minkowski's inequality we have

$$(\int |f-g|^p)^{1/p} \geq |(\int |f|^p)^{1/p} - (\int |g|^p)^{1/p}| ,$$

where the lower bound is infinite if one of the terms is infinite and the other one is finite. Thus, in any reasonable theory involving the L_p measure, we must assume first that $f \in L_p$. However, the only space to which all densities belong without discrimination is L_1 .

3. If f and g are both densities, then for any set $B \subseteq \mathbb{R}^d$, the probabilities of B defined by f and g respectively differ by at most

$$\Delta = \sup_B | \int_B f - \int_B g | .$$

For example, if Δ is known to be less than 10^{-4} , then two independent samples of size 10^4 , one from f and one from g , are all but statistically indistinguishable. Thus, keeping Δ small has a true practical impact in the area of simulation. But clearly,

$$\Delta = \frac{1}{2} \int |f-g| .$$

No other L_p measure has any connection with Δ in the sense that for any $p > 1$ and any f , there exist sequences of densities f_n and g_n such that

$$(i) \int |f_n - f| \rightarrow 0, \int |f_n - f|^p \rightarrow \infty ,$$

$$(ii) \int |f_n - f| = c > 0, \int |f_n - f|^p \rightarrow 0 .$$

4. If f and g are normal densities with zero mean and variances σ^2 and τ^2 , then $\int |f-g|$ depends only upon σ/τ , and tends to 0 if and only if $\sigma/\tau \rightarrow 1$. However, for $p > 1$, $\int |f-g|^p$ can tend to ∞ even if σ/τ tends to 1 (let $\sigma \rightarrow 0, \tau = \sigma + \sigma^{3p/(2p+1)}$), and it can tend to 0 even if σ/τ tends to ∞ (let $\tau \rightarrow \infty$ and $\sigma/\tau \rightarrow \infty$).

3. NEGATIVE RESULTS.

Many a density estimate (such as the kernel estimate) has been criticized for not performing well "for small sample sizes". Recent work in the area of density estimation has been in the direction of improved small sample performance and automatization of the estimate (automatization of the kernel estimate means that the parameter h is chosen as a function of the data). For research in this direction, see Deheuvels (1977a,b), Duin (1976), Scott et. al. (1977), Silverman (1978), Davis (1977), Scott et. al. (1981), Wahba (1977, 1978), de Montricher et. al. (1975), Good and Gaskins (1980), Breiman et.al. (1977), Nadaraya (1974), and Devroye and Wagner (1980). Most automatization schemes are so sophisticated that it is hard to prove that f_n converges to f in any sense at all. In fact, many schemes should be avoided altogether. For example, Schuster and Gregory (1981) have shown that the cross-validation method for determining "h" in the kernel estimate will not lead to a consistent estimate for most densities f with an infinite tail (such as the exponential density). Consistent cross-validated density estimation is also discussed by Chow, Geman and Wu (1981).

Even if an estimate is known to be consistent for all densities f , its small sample and large sample properties may be terrible. The search for always better estimates is doomed to be frustrating. In part, this frustration is captured in the following result.

Theorem 1. (Devroye, 1981a)

For every density estimate, and every $p \geq 1$, and every sequence of positive numbers tending to 0 (a_n), there exists a density f on R^d such that

$$E(|f_n - f|^p) \geq a_n \text{ infinitely often.}$$

We can always find such an f among the class of densities bounded by 2 and vanishing outside $[0,1]^d$. Moreover, for $p=1$, the density f in question can also be taken from the class of infinitely many times continuously differentiable functions.

Thus, any kind of continuity condition alone, however strong, is not sufficient for the study of the rate of convergence to 0 of $E(|f_n - f|)$, regardless of the type of estimate that is used ! For such studies, it seems that one needs combinations of continuity and tail conditions. Theorem 1 is in the spirit of a theorem proved by Boyd and Steele in 1979.

Theorem 2. (Boyd and Steele, 1979,

For every density estimate, there exists a normal density f with zero mean such that

$$E(|f_n - f|^2) \geq c(f)/n \text{ infinitely often,}$$

where $c(f) > 0$ is a constant depending upon f only.

In a sense, Theorem 2 gives us new information. Even if f is known to be normal with zero mean and unknown variance, it is impossible to find an estimate with an L_2 rate of convergence that is better than $1/n$. The theorem cannot be improved in the sense that the parametric estimate (1) satisfies $E(|f_n - f|^2) \leq c(f)/n$, all n (Maniya, 1969).

Let us finally point out that several results that have received widespread attention to date are practically vacuous. For example, Rosenblatt (1971) has shown that the kernel estimate (2) satisfies

$$E(|f_n - f|^2) \sim \frac{a}{nh} + \frac{b}{4} h^4$$

as $n \rightarrow \infty$, $h \rightarrow 0$, when K is bounded and symmetric, $d=1$, $\int x^2 K < \infty$, $a = \int K^2$, $b = (\int x^2 K)^2 \int f''^2$, and $f \in \mathfrak{J} = \{\text{all densities on } \mathbb{R}^1 \text{ that are twice continuously differentiable and for which } \int f^2 < \infty \text{ and } \int f''^2 < \infty \text{ and } f \text{ is bounded}\}$. Thus, if we take $h = (a/(bn))^{1/5}$, then

$$E(|f_n - f|^2) \sim \frac{5}{4} a^{4/5} b^{1/5} / n^{4/5}. \quad (4)$$

Thus, there are densities f in \mathfrak{J} for which (4) is valid and for which at the same time, $E(|f_n - f|) \geq 1/\log \log \log n$ infinitely often (theorem 1). But without guarantees for the performance of f_n in L_1 , Rosenblatt's result loses credibility. Thus, the choice $h = (a/(bn))^{1/5}$ for the kernel estimate, even if a and b were known, may not be "optimal" after all !

OF POSITIVE QUALITY

4. POSITIVE RESULTS.

A thorough study of global rates of convergence for density estimates in general and the kernel estimate in particular was carried out by Bretagnolle and Huber (1979). We cite one of their results that is closest to what we need in the present discussion.

Theorem 3. (Bretagnolle and Huber, 1979)

If $d=1$ and $f \in \mathcal{G}_s = \{ \text{all densities with compact support, that are } s \text{ times differentiable (} s \geq 1 \text{ is an integer) such that } \int |f^{(s)}| < \infty \}$, and if the kernel K in (2) satisfies: $\int K = 1$, $\int x^j K = 0$ ($0 < j < s$), $\int |x|^s |K| < \infty$, K has compact support, then a sequence $h=h(n)$ can be found such that for the kernel estimate

$$\limsup_n \frac{s}{2s+1} E(|f_n - f|) \leq c(s) \left(\int |f^{(s)}| \right)^{\frac{1}{2s+1}} \left(\int |f| \right)^{\frac{s}{2s+1}}, \text{ some } c(s) > 0.$$

This does not contradict theorem 1 because \mathcal{G}_s combines a continuity condition and a compactness condition. Unfortunately, \mathcal{G}_s does not include many common densities such as the normal and exponential densities.

A second positive development is related to the observation that for most reasonable nonparametric density estimates, $E(|f_n - f|) \rightarrow 0$ for all densities f on R^d . If we cannot say much about rates of convergence, at least we are guaranteed that the estimates are consistent. The first result of this type is due to Abou-Jaoude (1976a, 1976b, 1976c), who studies the histogram estimates. Here we consider a sequence of partitions P_n of R^d , where $P_n = \{A_{n1}, A_{n2}, \dots\}$, and we denote the set A_{ni} to which x belongs by $A_n(x)$. The histogram estimate is defined by

$$f_n(x) = (n\lambda(A_n(x)))^{-1} \sum_{i=1}^n I_{A_{ni}}(x) (X_i), \quad (5)$$

where I is the indicator function and λ is Lebesgue measure. Although Abou-Jaoude treats very general sorts of partitions, we will only state his results for the most common partitions: P_n consists of all sets

$$\prod_{i=1}^d [a_i b_n, (a_i+1)b_n) \quad (6)$$

where a_1, \dots, a_d can take all the integer values, and b_n is a sequence of positive numbers.

Theorem 4. (Abou-Jaoude, 1976a,c)

For the histogram estimate defined by (5-6), the following conditions are equivalent :

- A. $\int |f_n - f| \rightarrow 0$ in probability as $n \rightarrow \infty$, for all f .
- B. $\int |f_n - f| \rightarrow 0$ almost surely as $n \rightarrow \infty$, for all f .
- C. $\int |f_n - f| \rightarrow 0$ completely as $n \rightarrow \infty$, for all f .
- D. $\lim_{n \rightarrow \infty} b_n = 0$, $\lim_{n \rightarrow \infty} n b_n^d = \infty$.

(A sequence of random variables X_n converges completely to 0 if for all $\epsilon > 0$, $\sum_{n=1}^{\infty} P(|X_n| > \epsilon) < \infty$. Thus, complete convergence implies almost sure convergence.)

For histogram estimates, all types of L_1 convergence are equivalent. The L_1 convergence of the kernel estimate for all densities f was first observed by Devroye and Wagner (1979). Devroye (198b) showed a strong equivalence theorem for the kernel estimate :

Theorem 5. (Devroye, 198b)

For the kernel estimate (2) with a compact support kernel $K \geq 0$ which integrates to 1, the following statements are equivalent :

- A. $\int |f_n - f| \rightarrow 0$ in probability as $n \rightarrow \infty$, for some f .
- B. $\int |f_n - f| \rightarrow 0$ almost surely as $n \rightarrow \infty$, for some f .
- C. $\int |f_n - f| \rightarrow 0$ completely as $n \rightarrow \infty$, for some f .
- D. $\int |f_n - f| \rightarrow 0$ completely as $n \rightarrow \infty$, for all f .
- E. $\lim_{n \rightarrow \infty} h = 0$, $\lim_{n \rightarrow \infty} n h^d = \infty$.

Furthermore, E implies D whenever K is absolutely integrable and $\int K = 1$.

The difference with Theorem 4 is that weak convergence (A) for one f is enough to conclude E in Theorem 5, while weak convergence for all f is needed to conclude D in Theorem 4. Thus, either we have convergence in L_1 for kernel estimates (and then the convergence is in the strongest possible sense, and for all f), or we have no convergence in L_1 for kernel estimates (and then the estimate does not even converge in the weakest sense for a single f).

5. REFERENCES.

- ABOU-JAOUDE, S. (1976a). Sur une condition nécessaire et suffisante de L_1 -convergence presque complète de l'estimateur de la partition fixe pour une densité. Comptes Rendus de l'Académie des Sciences de Paris, Série A 283 1107-1110.
- ABOU-JAOUDE, S. (1976b). Sur la convergence L_1 et L_∞ de l'estimateur de la partition aléatoire pour une densité. Annales de l'Institut Henri Poincaré 12 299-317.
- ABOU-JAOUDE, S. (1976c). Conditions nécessaires et suffisantes de convergence L_1 en probabilité de l'histogramme pour une densité. Annales de l'Institut Henri Poincaré 12 213-231.
- BEAN, S.J. and TSOKOS, C.P. (1980). Developments in nonparametric density estimation. International Statistical Review 48 267-287.
- BOYD, D.W. and STEELE, J.M. (1978). Lower bounds for nonparametric density estimation rates. Annals of Statistics 6 932-934.
- BREIMAN, L., MEISEL, W. and PURCELL, E. (1977). Variable kernel estimates of multivariate densities. Technometrics 19 135-144.
- BRETAGNOLLE, C. and HUBER, C. (1979). Estimation des densités: risque minimax. Zeitschrift für Wahrscheinlichkeitstheorie und verwandte Gebiete 47 119-137.
- CHOW, Y.S., GEMAN, S. and WU, L. (1981). Consistent cross-validated density estimation. Manuscript, Brown University, Providence, Rhode Island.
- DAVIS, K.B. (1977). Mean integrated square error properties of density estimates. Annals of Statistics 5 530-535.
- DEHEUVELS, P. (1977a). Estimation non paramétrique de la densité par histogrammes généralisés. Revue de Statistique Appliquée 25 5-42.
- DEHEUVELS, P. (1977b). Estimation non paramétrique de la densité par histogrammes généralisés. Publications de l'Institut de Statistique de l'Université de Paris 22 1-23.
- de MONTRICHER, G.F., TAPIA, R.A. and THOMPSON, J.R. (1975). Nonparametric maximum likelihood estimation of probability densities by penalty function methods. Annals of Statistics 3 1329-1348.

- DEVROYE, L. and WAGNER, T.J. (1979). The L1 convergence of kernel density estimates. Annals of Statistics 7 1136-1139.
- DEVROYE, L. and WAGNER, T.J. (1980). The strong uniform consistency of kernel density estimates. Multivariate Analysis V, P.R.Krishnaiah Ed., North Holland 59-77.
- DEVROYE, L. (1981a). On arbitrarily slow rates of global convergence in density estimation. Manuscript, McGill University, Montreal.
- DEVROYE, L. (1981b). The equivalence of weak, strong and complete convergence in L1 for kernel density estimates. Manuscript, McGill University, Montreal.
- DUIN, R.P.W. (1976). On the choice of smoothing parameters for Parzen estimators of probability density functions. IEEE Transactions on Computers C-25 1175-1179.
- GOOD, I.J. and GASKINS, R.A. (1980). Density estimation and bump-hunting by the penalized method exemplified by scattering and meteorite data. Journal of the American Statistical Association 75 42-73.
- LOFTSGAARDEN, D.O. and QUESENBERRY, C.P. (1965). A nonparametric estimate of a multivariate density function. Annals of Mathematical Statistics 36 1049-1051.
- MANIYA, G.M. (1969). The square error of the density estimate of a multidimensional normal distribution for a given sample. Theory of Probability and its Applications 14 149-153.
- NADARAYA, E.A. (1974). On the integral mean square error of some nonparametric estimates for the density function. Theory of Probability and its Applications 19 133-141.
- PARZEN, E. (1962). On the estimation of a probability density function and the mode. Annals of Mathematical Statistics 33 1065-1076.
- ROSENBLATT, M. (1956). Remarks on some nonparametric estimates of a density function. Annals of Mathematical Statistics 27 832-837.
- ROSENBLATT, M. (1971). Curve estimates. Annals of Mathematical Statistics 42 1815-1842

- SCHUSTER, E.F. and GREGORY, G.G. (1981). On the nonconsistency of maximum likelihood nonparametric density estimators. Computer Science and Statistics: 13th Symposium on the Interface, Springer Verlag.
- SCOTT, D.W., TAPIA, R.A. and THOMPSON, J.R. (1977). Kernel density estimation revisited. Nonlinear Analysis 1 339-372.
- SCOTT, D.W. and FACTOR, L.E. (1981). Monte Carlo study of three data-based nonparametric probability density estimators. Journal of the American Statistical Association 76 9-15.
- SILVERMAN, B.W. (1978). Choosing the window width when estimating a density. Biometrika 65 1-11.
- WAHBA, G. (1977). Optimal smoothing of density estimates. Classification and Clustering, J.Van Ryzin Ed., Academic Press, New York, 423-458.
- WAHBA, G. (1978). Data-based optimal smoothing of orthogonal series estimates. Annals of Statistics, to appear.
- WEGMAN, E.J. (1972). Nonparametric probability density estimation: I. A summary of available methods. Technometrics 14 533-546.
- WERTZ, W. (1978). Statistical Density Estimation. A Survey. Vandenhoeck and Ruprecht, Göttingen, Applied Statistics and Econometrics Series, vol. 13.
- WERTZ, W. and SCHNEIDER, B. (1979). Statistical density estimation : a bibliography. International Statistical Review 47 155-175.

LUC DEVROYE
SCHOOL OF COMPUTER SCIENCE
MCGILL UNIVERSITY
805 SHERBROOKE STREET WEST
MONTREAL, CANADA H3A 2K6

N83

15776

UNCLAS

This Page Intentionally Left Blank

R 100

D2

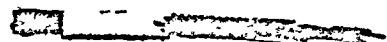
CROSS-VALIDATION FOR DENSITIES AND REGRESSIONS

Stuart Geman
Division of Applied Mathematics
Brown University
Providence, Rhode Island 02912

lecture delivered at NASA-sponsored workshop on
"DENSITY ESTIMATION AND FUNCTION SMOOTHING"
Texas A&M University
March 11,12,13, 1982

PRECEDING PAGE BLANK NOT FILMED

This work was partially supported by the Department of the
Army contract DAAG29-80-K-0006 and the U.S. Air Force grant
AFOSR 78-3514.



Contents

- I. Introduction
- II. Cross-validation for choosing smoothing parameters
 - A. Kernel and histogram
 - B. Ridge regression
- III. Analytic results
 - A. Why should cross-validation work?
 - B. Ridge regression
 - C. Density estimation

I. Introduction

Virtually all nonparametric (infinite dimensional) problems require the choice of a "smoothing parameter".

Example: x_1, x_2, \dots i.i.d. from a distribution with unknown density "f". Consider the Parzen-Rosenblatt kernel estimator with window width $1/\lambda$:

$$f_{n,\lambda}(x) = \frac{1}{n} \sum_{i=1}^n \lambda k(\lambda(x-x_i))$$

where k is a probability kernel, or the histogram with bin width $1/\lambda$:

$$f_{n,\lambda}(x) = \frac{\lambda}{n} \# \{x_i : \frac{k-1}{n} \leq x_i < \frac{k}{n}\} \quad x \in [\frac{k-1}{n}, \frac{k}{n}).$$

In each case λ serves as a smoothing parameter. It is well-known that if $\lambda_n \uparrow \infty$ sufficiently slowly then $f_{n,\lambda_n} \rightarrow f$ (e.g. almost surely in $L_1(\mathbb{R}, \mathcal{B}, dx)$). Depending on the assumptions made, optimal rates can be specified for λ_n , but these will always depend on the unknown density f . How should λ be chosen for a fixed, finite, sample? For moderate sample sizes, both estimators are sensitive to the choice of λ . This is the "smoothing problem". It has its analogue for virtually all (non-Bayesian) nonparametric density estimators. For example, the maximum penalized likelihood estimator requires the choice of a weight to be given the penalty term. Orthogonal series estimators (for densities or regressions) require that we specify the number of terms to be used in a truncated series

OF POOR QUALITY

expansion. Splines for nonparametric regression typically arise from solving a least squares problem with penalty, which may be, for example, the integral of the squared second derivative of the estimator. As with penalized maximum likelihood, the smoothing parameter here is the weight given the penalty term.

Some estimators of finite dimensional parameters also contain unspecified smoothing parameters. In fact, it is sometimes useful to introduce a smoothing parameter into an estimator that is otherwise completely specified. Consider, for example, the linear regression problem:

$$y_i = x_{i1}\beta_1 + \dots + x_{ip}\beta_p + \epsilon_i, \quad 1 \leq i \leq n, \quad \epsilon_i \text{ iid } N(0, \sigma^2).$$

Or, in vector-matrix notation:

$$Y = X\beta + \epsilon \quad \epsilon \sim N(0, \sigma^2 I).$$

The least squares (maximum likelihood) estimator for β is

$$\hat{\beta} = (X^T X)^{-1} X^T Y.$$

The ridge estimator for β is

$$\hat{\beta}_\lambda = (X^T X + n\lambda I)^{-1} X^T Y \quad \lambda \geq 0.$$

Observe that $\hat{\beta}_0$ is the least squares estimator. The introduction of λ into the least squares estimator may be motivated by any of the following considerations. (1) $\hat{\beta}_\lambda$ minimizes an equation of the form

$$\|Y - X\hat{\beta}\|^2 + \gamma \|\hat{\beta}\|^2.$$

OF POOR QUALITY

Hence $\hat{\beta}_\lambda$ may be viewed as a penalized least squares estimator, with penalty for large values of $\hat{\beta}$. (2) When $X^T X$ is "nearly singular" (poorly conditioned) $\hat{\beta}_0$ has large MSE due to the fact that the inverse of $X^T X$ is involved in its derivation. Adding $n\lambda I$ to $X^T X$ improves the conditioning and may be expected to reduce MSE. (3) Perhaps the best justification for ridge regression is the following easily demonstrated fact: for every n , β , and $\sigma^2 > 0$, there exists a $\lambda > 0$ such that

$$E\|\beta - \hat{\beta}_\lambda\|^2 < E\|\beta - \hat{\beta}_0\|^2.$$

Unfortunately, the optimal λ (in terms of MSE) depends on β and σ^2 , so that we are again faced with a version of the "smoothing problem".

It is natural to attempt to use the data to guide the choice of smoothing parameter. For each of the above examples many such "data-driven" estimators have been proposed. Perhaps the most widely applicable (certainly the most widely studied) data-driven technique is cross-validation. Simulations show that cross-validation can be a very effective means for choosing smoothing parameters. However, the technique can badly fail, and the conditions for success are not well-understood. In fact, almost nothing is known of the analytic properties of cross-validated estimators. In collaboration with Drs. Y.S. Chow and L.-D. Wu (previously visiting Brown University) and Aytul Irdul (currently a graduate student at Brown University) I have been attempting to establish some of the analytic properties of

cross-validated estimators. In the remainder of this talk I will introduce, by example, the method of cross-validation, and announce results which establish consistency for certain cross-validated density estimators and consistency as well as asymptotic normality for ridge regression.

II. Cross-validation for choosing smoothing parameters.

This method is best introduced by example.

A. Kernel and histogram.

Recall: x_1, x_2, \dots is an i.i.d. sample from a distribution with unknown density "f". $f_{n,\lambda}$ is either the kernel with window width $1/\lambda$, or the histogram with bin width $1/\lambda$. The problem is to choose λ when faced with a fixed and finite sample x_1, x_2, \dots, x_n . The first step in applying cross-validation is to form the estimator from the sample after first deleting one of the observations:

$$f_{n-1,\lambda}^i(x) = \frac{1}{n-1} \sum_{j \neq i} \lambda k(\lambda(x-x_j)).$$

$f_{n-1,\lambda}^i(x_1)$ is a measure of the appropriateness of λ for smoothing the estimator. If $f_{n-1,\lambda}^i(x_1)$ is large, we could say, loosely, that $f_{n-1,\lambda}^i(x)$ "anticipated the observation x_1 " (for fixed λ , $f_{n-1,\lambda}^i(x)$ is formed independent of x_1). If $f_{n-1,\lambda}^i(x_1)$ is small, then x_1 was measured as "unlikely", evidence that λ does not properly smooth the estimator. Through this procedure, applied n times, we arrive at a likelihood-like expression:

$$L_\lambda = \prod_{i=1}^n f_{n-1,\lambda}^i(x_1).$$

ORIGINAL PAGE IS
OF POOR QUALITY

We now choose $\lambda = \lambda_n$ to maximize L_λ . The cross-validated estimator (due to Habbema et al. (5) and, independently, Duin (3)) is f_{n, λ_n} . Simulations strongly support the use of this technique for certain combinations of the kernel and target density. However, the method can fail, and in surprisingly innocent looking situations. For example, Schuster and Gregory (6) have shown that the cross-validated kernel, using compact kernel, is not consistent for the exponential density. With this, as with all cross-validated estimators, very little is known analytically. In fact, with the exception of the results mentioned below for kernels and histograms, conditions for the consistency of cross-validated density estimators are unknown.

B. Ridge regression

Recall that the ridge estimator for β in the model

$$Y = X\beta + \epsilon \quad \epsilon \sim N(0, \sigma^2 I)$$

is

$$\hat{\beta}_\lambda = (X^T X + n\lambda I)^{-1} X^T Y \quad \lambda \geq 0.$$

Define $\hat{\beta}_\lambda^i$ to be the ridge estimator obtained by deleting (ignoring) the i 'th observation. The squared error in predicting the i 'th observation:

$$(y_i - \sum_{j=1}^n x_{ij} \hat{\beta}_{\lambda_j}^i)^2$$

measures the appropriateness of λ as a smoothing parameter. Define

$$L_\lambda = \frac{1}{n} \sum_{i=1}^n (y_i - \sum_{j=1}^n x_{ij} \hat{\beta}_{\lambda_j}^i)^2$$

and choose $\lambda = \lambda_n$ to minimize L_λ . The cross-validated ridge estimator (due to Allen (1)) is $\hat{\beta}_{\lambda_n}$. Our simulations, and those of others, indicate that $\hat{\beta}_{\lambda_n}$ is an extremely good estimator for β , especially when $X^T X$ is nearly singular or σ is large. Although they may exist, we have not found any situations in which the mean squared error of the cross-validated ridge regression estimator exceeds that of the ordinary least-squares estimator. Often, the ridge estimator reduces the MSE of least squares by 50 or more percent.

There is a closely related estimator, due to Golub, Heath, and Wahba (4), called the "generalized cross-validation" (GCV) ridge regressor. The GCV ridge regressor is computed by first rotating the coordinate system and then deriving the ordinary cross-validation estimator. Simulations demonstrate the GCV generally performs somewhat better than ordinary cross-validation, and GCV proves to be more mathematically tractable. Although the above-mentioned analytic results are for GCV, I will not formally define the GCV estimator since this would require that I introduce somewhat involved notation. Suffice it to say that GCV is ordinary cross-validation in a rotated coordinate system.

I should emphasize that cross-validation has its version for all of the estimators mentioned earlier, each of which requires the choice of a smoothing parameter to be fully defined. Quite generally simulations support its good potential, and quite generally there are no theoretical results available about the cross-validated estimator. Thus questions of distribution, efficiency, robustness, and even consistency are almost completely unanswered.

III. Analytic ResultsA. Why should cross-validation work?

Before stating some analytic results about cross-validated estimators, let me outline a heuristic argument in favor of the technique in the ridge regression context. This argument has its analogue for most cross-validated estimators, whether the target parameter is a density or a regression. In some cases it can be made into a proof of consistency (as it can for cross-validated ridge regression), but in nonparametric problems it appears that one must take a different approach. Nevertheless, the motivation is similar for nonparametric as well as parametric problems.

The cross-validated ridge regressor is

$$\hat{\beta}_{\lambda_n} = (X^T X + n\lambda_n I)^{-1} X^T Y;$$

where λ_n is chosen to minimize

$$(*) \quad \frac{1}{n} \sum_{i=1}^n (y_i - \sum_{j=1}^n x_{ij} \hat{\beta}_{\lambda_j}^i)^2.$$

Although $\hat{\beta}_{\lambda_j}^i$ depends implicitly on Y , it is reasonable to expect that a version of the law of large numbers will be in force uniformly in $\hat{\beta}_{\lambda_j}^i$. This leads us to expect that for large n (*) is close to

$$E_Y \frac{1}{n} \sum_{i=1}^n (y_i - \sum_{j=1}^n x_{ij} \hat{\beta}_{\lambda_j}^i)^2$$

where " E_Y " means integration with respect to explicit appearances of the components of Y , treating $\hat{\beta}_{\lambda_j}^i$ as constant. It is also

reasonable to expect that $\hat{\beta}_{\lambda_j}^i$ will differ very little from $\hat{\beta}_{\lambda_j}$, especially when n is large. Thus we choose λ_n to minimize an expression which we might expect, for large n , to be close to

$$\begin{aligned} E_Y \frac{1}{n} \sum_{i=1}^n (y_i - \sum_{j=1}^n x_{ij} \hat{\beta}_{\lambda_j})^2 \\ = \frac{1}{n} E_Y \|Y - X\hat{\beta}_{\lambda}\|^2 = \frac{1}{n} \|X\beta - X\hat{\beta}_{\lambda}\|^2 + \sigma^2. \end{aligned}$$

The conclusion is that the cross-validated estimator attempts to minimize the positive definite quadratic form

$$(**) \quad (\beta - \hat{\beta}_{\lambda})^T \frac{X^T X}{n} (\beta - \hat{\beta}_{\lambda}).$$

Since

$$(\beta - \hat{\beta}_0)^T \frac{X^T X}{n} (\beta - \hat{\beta}_0) \rightarrow 0$$

(recall that $\hat{\beta}_0$ is the least squares estimator), we expect that (***) will also converge to 0, and at least as fast.

B. Ridge regression

Here, loosely stated, is what we know about the analytic properties of the cross-validated ridge estimator:

THEOREM (with Aytul Erdal). If $\hat{\beta}_{\lambda_n}$ is the GCV ridge regressor then

$$\|\hat{\beta}_{\lambda_n} - \beta\| \rightarrow 0 \quad \text{a.s.}$$

and $(X^T X)^{1/2} (\hat{\beta}_{\lambda_n} - \beta) \sim N(0, \sigma^2 I)$.

Observe that for least squares the distribution of

$$(X^T X)^{1/2} (\hat{\beta}_0 - \beta)$$

is exactly $N(0, \sigma^2 I)$. Thus the GCV estimator asymptotically assumes all of the distributional properties of the least squares estimator.

C. Density estimation

Results are much more difficult for infinite dimensional target parameters. So far, for the cross-validated kernel and histogram we have only a consistency result (stated here without all of the technical details - see (2) for the precise formulation):

THEOREM (with Y.S. Chow and L.-D. Wu). If f (the target density) has compact support, then the cross-validated histogram and compact-kernel density estimators are consistent:

$$\int |f_{n, \lambda_n}(x) - f(x)| dx \rightarrow 0 \quad \text{a.s. .}$$

References

1. Allen, D.M. (1974). The relationship between variable selection and data augmentation and a method for prediction. *Technometrics*, 16, 125-127.
2. Chow, Y.S., Geman, S., and Wu, L.-D. (1982). Consistent cross-validated density estimation. *Annals of Statistics* (to appear).
3. Duin, R.P.W. (1976). On the choice of smoothing parameters for Parzen estimators of probability density functions. *IEEE Trans. on Computers*, C-25, 1175-1179.
4. Golub, G.H., Heath, M., and Wahba, G. (1979). Generalized cross-validation as a method for choosing a good ridge parameter. *Technometrics*, 21, 215-223.
5. Habbema, J.D.F., Hermans, J., and van den Brock, K. (1977). Selection of variables in discriminant analysis by F-statistic and error rate. *Technometrics*, 19, 487-493.
6. Schuster, E.F. and Gregory, G.G. (1981). On the nonconsistency of maximum likelihood nonparametric density estimators. In: *Computer Science and Statistics: Proceedings of the 13th Symposium on the Interface*. Ed. W.F. Eddy, Springer-Verlag, New York, 295-298.

Summary of talk presented at the
NASA Workshop on Density Estimation and Function Smoothing
at Texas A&M University
March 11-13, 1982

[This version contains only the Summary and References of the complete technical report. The full report will be made available on request to the author.]

Estimation of Planar Sets
from Poisson Projections

by

Donald E. McClure

Reports in Pattern Analysis No. 115

Division of Applied Mathematics
Brown University
Providence, Rhode Island 02912

April 1982

Research supported in part by the Army Research Office under contract DAAG29-80-K-0006 and the Air Force Office of Scientific Research through grant No. 78-3514 to Brown University.

OF POOR QUALITY,

1. Summary

This report summarizes ongoing work concerned with the reconstruction of planar sets that can be only partially observed. Details of the problem formulations and of the results reported here are being incorporated in a report describing a broader class of problems, specifically the estimation of an intensity function of a planar Poisson process based on observations of stochastically independent fixed-angle projections of the process.

First, the set-estimation problem is formulated and connected to reconstruction methods of emission computed tomography. Then the inference problem per se will be isolated and approached by traditional estimation methods.

I shall focus on the special case of estimating an unknown planar convex body K_2 that is a subset of a known convex body K_1 . Poisson events occur with an intensity $\lambda(x,y)$ that is spatially inhomogeneous (and temporally homogeneous) within the larger set K_1 ; we assume for our prototypal problem that $\lambda(x,y) \equiv \lambda_2$ within K_2 and $\lambda(x,y) \equiv \lambda_1 < \lambda_2$ within $K_1 - K_2$. The Poisson events are projected on a line \mathcal{L}_θ with fixed arbitrary orientation θ relative to the horizontal axis, and only the projected points are observable.

The underlying model for generation of the projected point process implies that its univariate intensity function μ_θ is a superposition of the "shadows" of K_1 and K_2 . In particular,

$$\mu_\theta(\xi) = \lambda_1 w_1(\xi) + (\lambda_2 - \lambda_1) w_2(\xi) \quad (1)$$

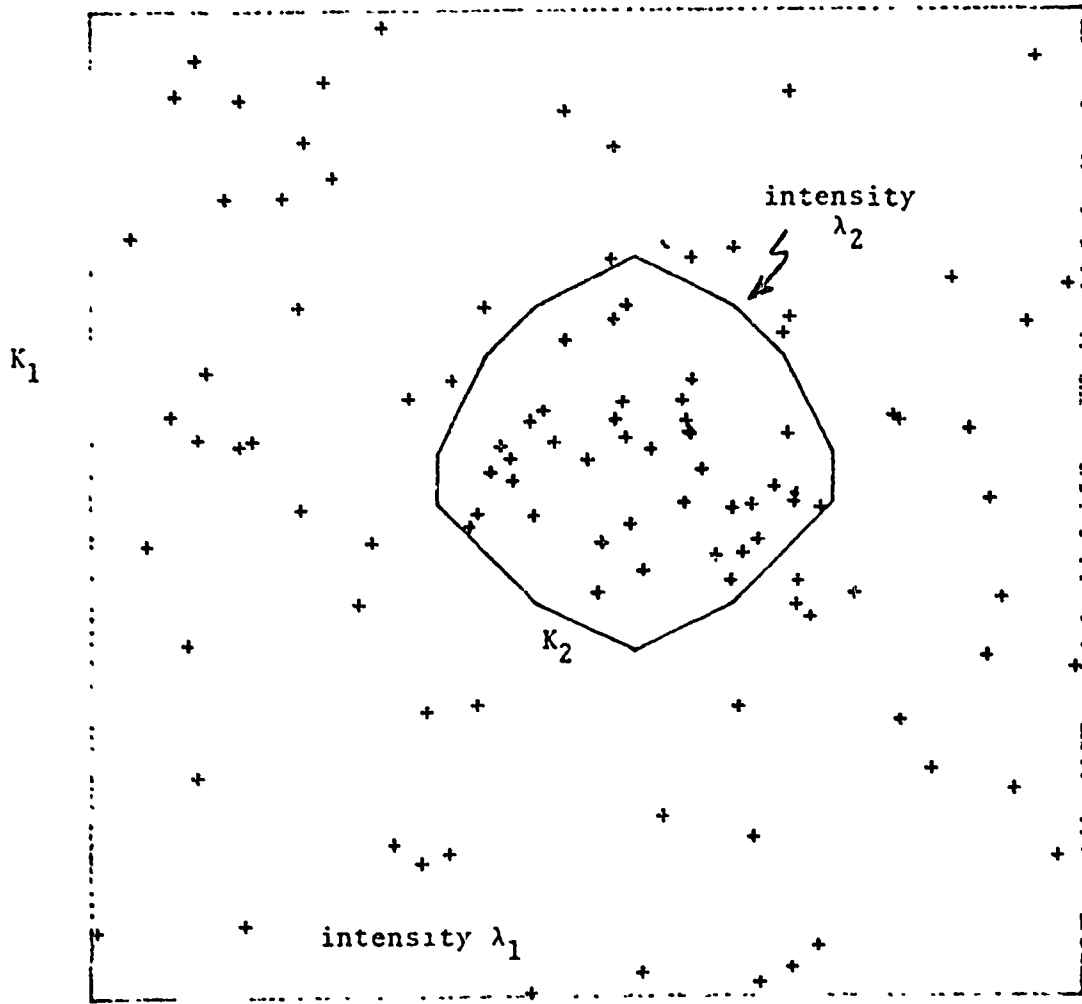
ORIGINAL PAGE IS
OF POOR QUALITY

where (i) $w_1(\xi)$ is the known width of K_1 , in direction $\theta + \pi/2$ and at location ξ along the line \mathcal{L}_θ , (ii) $w_2(\xi)$ is the unknown width function of K_2 , and (iii) λ_1 and λ_2 are the unknown values of $\lambda(x,y)$ within $K_1 - K_2$ and K_2 , respectively. When K_1 and K_2 are convex then w_1 and w_2 are unimodal and analogies with familiar nonparametric inference problems can be drawn.

The problem that is solved in this report is the characterization of the maximum likelihood estimates of λ_1 and of $u = (\lambda_2 - \lambda_1)w_2$, under the constraints on the structure of u that follow from convexity of K_2 . The characterization is patterned after ones that are familiar in the context of isotonic estimation and regression. Specifically, the m.l.e. u^* of u attains a maximum value on a nondegenerate interval $[\xi_0, \xi_1]$. To the left of ξ_0 (and to the right of ξ_1), u^* is the slope of the greatest convex minorant of a modified counting function for the univariate point process.

The characterization of u^* is finite-dimensional and its computation is feasible. The intrinsic complexity of the computation of u^* is discussed and an implemented algorithm is described. Finally, a simulation example illustrates the performance of the estimator u^* and the use of u^* to reconstruct the boundary of K_2 .

OF POOR QUALITY



Model Generating Two-dimensional Poisson Process

Figure 2

References

- [1] R.E. Barlow, D.J. Bartholomew, J.M. Fremner and H.D. Brunk, Statistical Inference under Order Restrictions, John Wiley & Sons, New York (1972).
- [2] T.F. Budinger, Computed tomography: three-dimensional imaging with photons and nuclear magnetic resonance, in Biomedical Pattern Recognition and Image Processing, K.S. Fu and T. Pavlidis, editors, Verlag Cnemie, Weinheim, West Germany (1979).
- [3] T.F. Budinger, G.T. Gullberg and R.H. Huesman, Emission computed tomography, in Image Reconstruction from Projections: Implementation and Applications, G.T. Herman, editor, Volume 32 of Topics in Applied Physics, Springer-Verlag, New York (1979), 147-246.
- [4] H. Cramér and H. Wold, Some theorems on distribution functions, J. London Math. Soc. 11 (1936), 290-294.
- [5] U. Grenander, On the theory of mortality measurement, Skand. Aktuarietidskr 39 (1956), 71-153.
- [6] G.T. Gullberg, The attenuated Radon transform: theory and application in medicine and biology, Ph.D. thesis, University of California at Berkeley (1979).
- [7] G.T. Herman, Image Reconstruction from Projections, Academic Press, New York (1980).
- [8] D.E. McClure, Image Models in Pattern Theory, Computer Graphics and Image Processing 12 (1980), 309-325.
- [9] B.L.S. Prakasa Rao, Estimation of a unimodal density, Sankhyā A 31 (1969), 23-36.
- [10] L.A. Shepp and J.B. Kruskal, Computerized tomography: the new medical X-ray technology, Amer. Math. Monthly 85 (1978), 420-439.
- [11] L.A. Shepp and B.F. Logan, The Fourier reconstruction of a head section, IEEE Trans. Nucl. Sci. NS-21 (1974), 21-43.
- [12] K.T. Smith, D.C. Solomon and S.L. Wagner, Practical and mathematical aspects of the problem of reconstructing objects from radiographs, Bull. Amer. Math. Soc. 83 (1977), 1227-1270.
- [13] E.J. Wegman, Maximum likelihood estimation of a unimodal density, II, Ann. Math. Statist. 41 (1970), 2169-2174.

N83

15777

UNCLAS

Characterization of a maximum-likelihood
nonparametric density estimator of
kernel type

by

Stuart Geman and Donald E. McClure
Reports in Pattern Analysis No. 114
Division of Applied Mathematics
Brown University
Providence, Rhode Island 02912

March 1982

Research supported in part by the Department of the Army under
contract DAAG-80-K-0006 and by the Air Force Office of
Scientific Research through grant no. 78-3514 to Brown University.

1. Introduction

As an instance of Grenander's method of sieves [2] for adapting the maximum-likelihood approach to settings where the target parameter is infinite dimensional, we have considered density functions of the form

$$f(x) = \int_{-\infty}^{\infty} \frac{1}{\sigma} \phi((x-y)/\sigma) G(dy) = (\phi_{\sigma} * G)(x). \quad (1)$$

Here G is an arbitrary cdf and ϕ is the standard normal density function. In this note, we shall derive a characterization of the cdf G^* that solve the maximum-likelihood equation:

$$\mathcal{L}(G^*) = \max_G \mathcal{L}(G) \quad (2)$$

where $\mathcal{L}(G)$ is the likelihood function

$$\mathcal{L}(G) = \prod_{i=1}^n f(x_i) \quad (3)$$

determined by a random sample x_1, x_2, \dots, x_n from an unknown population density f_0 .

Geman and Hwang [1] have described the connection between this optimization problem and nonparametric maximum-likelihood estimation. In brief, if we specify a sequence $\{\sigma_m\}_{m=1}^{\infty}$ of positive values with $\sigma_m \downarrow 0$ as $m \rightarrow \infty$, then the sequence of sets

$$S_m = \{f : f = \phi_{\sigma_m} * G, \quad G \text{ an arbitrary cdf}\}$$

defines a sieve of subsets of L_1 , the so-called convolution sieve. The method-of-sieves (i) fixes an index m , depending

on sample size n and on the sequence $\{\sigma_m\}$, (ii) seeks the solution G_m^* of (2) determined by the sample $\{x_i\}_{i=1}^n$ and σ_m , and (iii) forms the estimator $f_m^* = \phi_{\sigma_m} * G_m^*$.

The familiar Parzen-Rosenblatt kernel estimator fits within this framework. The kernel estimator prescribes G to be the empirical cdf. One motivation for introducing the convolution sieve is to study the relationship between the kernel estimator and ones derived through the principle of maximum likelihood.

Our characterization theorem for G^* exhibits a rather close relationship between f_m^* and the kernel estimator based on the Gaussian kernel. We shall show that the solution G^* of (2) is a discrete cdf and that it contains no more than n points in its support. Thus, the estimator f_m^* obtained from the method-of-sieves admits a representation of the form

$$f_m^*(x) = \sum_{j=1}^q p_j \phi_{\sigma_m}(x-y_j),$$

analogous to a familiar form of the kernel estimator. In contrast to the kernel estimator, the support $\{y_j\}$ of G^* does not coincide with the sample $\{x_i\}_{i=1}^n$ and, in general, the weights $\{p_j\}$ will not be identically equal to n^{-1} . Computational experiments with closely related sieves strongly indicate that the number q of points in the support of G^* will typically be much smaller than sample size n .

2. Characterization Theorem

Theorem. Let x_1, x_2, \dots, x_n be a random sample from a population with density f_0 . Let $\sigma > 0$ and consider estimators f of f_0 defined by (1).

(i) There exists a solution G^* of the maximum-likelihood problem (2)-(3).

(ii) If G^* satisfies (2), then G^* is a discrete cdf with finite support. Denote $\text{supp}(G) = \{s_j\}_{j=1}^q$. Then $q \leq n$.

(iii) If $x_{(1)} = \min(\{x_i\}_{i=1}^n) < \max(\{x_i\}_{i=1}^n) = x_{(n)}$,

then $x_{(1)} < \min(\{s_j\}_{j=1}^q)$ and $\max(\{s_j\}_{j=1}^q) < x_{(n)}$.

Proof: We may assume, for convenience and without loss of generality, that $\sigma=1$. The sample values can be rescaled, setting $\hat{x}_i = x_i/\sigma$, if $\sigma \neq 1$.

The maximum of $\mathcal{L}(G)$, if it exists, will be attained by a cdf with support in $[x_{(1)}, x_{(n)}]$. To see this, consider an arbitrary right-continuous cdf G and defined G_0 in terms of G by

$$G_0((-\infty, x]) = \begin{cases} 0 & , \text{ for } x < x_{(1)} \\ G((-\infty, x]), & \text{ for } x_{(1)} \leq x < x_{(n)} \\ 1 & , \text{ for } x_{(n)} \leq x. \end{cases}$$

G_0 is designed so that $G_0(\{x_{(1)}\}) = G((-\infty, x_{(1)}])$ and $G_0(\{x_{(n)}\}) = G([x_{(n)}, \infty))$. Since ϕ is monotone on the separate intervals $(-\infty, 0]$ and $[0, \infty)$, we have

$$\begin{aligned} \phi(x_i - x_{(n)})G_0(\{x_{(n)}\}) &\geq \int_{x_{(n)}^-}^{\infty} \phi(x-y)G(dy) \quad \text{and} \\ \phi(x_i - x_{(1)})G_0(\{x_{(1)}\}) &\geq \int_{-\infty}^{x_{(1)}^+} \phi(x-y)G(dy). \end{aligned}$$

Consequently $(\phi * G_0)(x) \geq (\phi * G)(x)$ for all x in $[x_{(1)}, x_{(n)}]$ and hence $\mathcal{L}(G_0) \geq \mathcal{L}(G)$.

The existence of a solution G^* of (2) follows from (i) the compactness of the (tight) family of cdfs having support in $[x_{(1)}, x_{(n)}]$, and (ii) the observation that $\mathcal{L}(G)$ is a bounded and continuous functional on this set of cdfs, i.e. continuous with respect to the topology of weak convergence.

Let G^* be a solution of (2) and set $f^* = (\phi * G^*)$. A variational argument characterizes the points in the support of G^* as roots of a transcendental equation. Let s be an arbitrary point in the support of G^* . For any $\epsilon > 0$ and for any z , define a measure $H_{s,\epsilon,z}$ by

$$H_{s,\epsilon,z}(B) = G^*((s-\epsilon, s+\epsilon] \cap (B-z))$$

$H_{s,\epsilon,z}$ is a rigid shift through distance z of G^* restricted to $(s-\epsilon, s+\epsilon]$. Define $G_{s,\epsilon}^* = G^* - H_{s,\epsilon,0}$. Then $G_{s,\epsilon}^* + H_{s,\epsilon,z}$ is a cdf for any z , and it may be regarded as a local perturbation near s of G^* .

Set $f_{s,\epsilon,z} = \phi * [G_{s,\epsilon}^* + H_{s,\epsilon,z}]$ and observe that $f^* = f_{s,\epsilon,0}$. Since $\Pi f^*(x_i)$ is maximal, we have

$$0 = \frac{d}{dz} \sum_{i=1}^n \log f_{s,\epsilon,z}(x_i) \Big|_{z=0}.$$

Evaluation of the derivative gives

$$\begin{aligned}
 0 &= \sum_{i=1}^n \frac{1}{f^*(x_i)} \frac{d}{dz} (\phi * H_{s,\epsilon,z})(x_i) \Big|_{z=0} \\
 &= \sum_{i=1}^n \frac{1}{f^*(x_i)} \frac{d}{dz} \int_{s-\epsilon}^{s+\epsilon} \phi(x_i - y - z) G^*(dy) \Big|_{z=0} \\
 &= \sum_{i=1}^n \frac{1}{f^*(x_i)} \int_{s-\epsilon}^{s+\epsilon} (x_i - y) \phi(x_i - y) G^*(dy).
 \end{aligned}$$

Dividing this expression by $G^*((s-\epsilon, s+\epsilon])$ and letting $\epsilon \rightarrow 0$ yields

$$\sum_{i=1}^n \frac{(x_i - s)}{f^*(x_i)} \phi(x_i - s) = 0,$$

for any s in the support of G^* .

Now consider the function

$$T(y) = \sum_{i=1}^n \frac{(x_i - y)}{f^*(x_i)} \phi(x_i - y).$$

The support of G^* is a subset of the set of roots of T . Properties of this set follow from the connection of T with an extended Tchebycheff system. We can re-express T as

$$\begin{aligned}
 T(y) &= \frac{e^{-y^2/2}}{\sqrt{2\pi}} \sum_{i=1}^n [x_i e^{-x_i^2/2} e^{x_i y} - e^{-x_i^2/2} y e^{x_i y}] \\
 &= e^{-y^2/2} \left[\sum_{i=1}^n (a_i e^{x_i y} + b_i y e^{x_i y}) \right].
 \end{aligned}$$

The expression in braces is a simple linear combination of the $2n$ functions $\left[e^{x_i y}, y e^{x_i y} \right]_{i=1}^n$. When the x_i 's are distinct, this

set is an extended Tchebycheff system of order $2n$. (And of course if $\{x_i\}_{i=1}^n$ is a random sample from population density f_0 , then the x_i 's are distinct w.p.1. If the x_i 's were not distinct, we could reduce the order of the system accordingly to express $T(y)$ in terms of an extended Tchebycheff system with fewer than $2n$ elements.) The Tchebycheff property implies:

- (i) $Z^0 = \{y : T(y)=0\}$ has at most $2n-1$ elements, and
- (ii) $Z^{+-} = \{y : T(y)=0, T'(y) \leq 0\}$ has at most n elements (Karlin and Studden [3]).

Since the support of G^* is contained in Z^0 , G^* is discrete with at most $2n-1$ jumps.

In order to show that G^* has at most n jumps, it suffices to show that the support of G^* is actually contained in Z^{+-} , i.e. that $T'(s) \leq 0$ for any s in the support of G^* . For f^* , we can now write

$$f^*(x) = \sum_{j=1}^q p_j \phi(x-s_j)$$

where $\{s_j\}_{j=1}^q$ is the support of G^* , $q \leq 2n-1$, $p_j > 0$, and $\sum_{j=1}^q p_j = 1$. Set $s = s_\ell$, for fixed ℓ between 1 and q . Let $\epsilon > 0$ and define a perturbation f_ϵ of f^* by

$$f_\epsilon(x) = \sum_{j \neq \ell} p_j \phi(x-s_j) + \frac{p_\ell}{2} \phi(x-s+\epsilon) + \frac{p_\ell}{2} \phi(x-s-\epsilon).$$

The density f_ϵ admits a representation of the form (1) and $f^* = f_0$. Since $\Pi f^*(x_1)$ is maximal,

$$\frac{d^2}{d\epsilon^2} \sum_{i=1}^n \log f_{\epsilon}(x_i) \Big|_{\epsilon=0} \leq 0.$$

Straightforward calculation yields

$$\frac{d^2}{d\epsilon^2} \sum_{i=1}^n \log f_{\epsilon}(x_i) \Big|_{\epsilon=0} = p_{\ell} T'(s),$$

and hence, as claimed, $T'(s) \leq 0$.

Finally, to confirm the last statement in the theorem, observe that if $s \leq x_{(1)}$ for some s in the support of G^* , then $\phi(x_i - s)$ is strictly increasing for sufficiently small increases in s and for all x_i , except perhaps $x_{(1)}$. Further, $\frac{d}{ds} \phi(x_{(1)} - s) \geq 0$ as long as $s \leq x_{(1)}$; hence $\Pi f^*(x_i)$ is a strictly increasing function of s , contradicting the maximum-likelihood property of G^* and f^* . The same reasoning precludes $s > x_{(n)}$. \square

3. Concluding Remarks

The characterization theorem was announced in the paper by Geman and Hwang [1], where consistency questions for f^* are analyzed. The consistency results guarantee that $f^* \rightarrow f_0$ in L_1 -norm, with probability one, provided that $\sigma \rightarrow 0$ sufficiently slowly as sample size $n \rightarrow \infty$.

H. Robbins recently restimulated interest in the maximum-likelihood problem per se during his lecture at the NASA Workshop on Density Estimation and Function Smoothing at Texas A&M University, March 11-13, 1982. Professor Robbins recalled his 1950 formulation of the maximum-likelihood problem (1)-(3) in [4] wherein connections are made with statistical decision problems.

References

- [1] S. Geman and C-R. Hwang, Nonparametric maximum-likelihood estimation by the method of sieves, to appear in Ann. Statist.
- [2] U. Grenander, Abstract Inference, John Wiley & Sons, New York (1981).
- [3] S. Karlin and W.J. Studden, Tchebycheff systems: with applications in analysis and statistics, Interscience, John Wiley & Sons, New York (1966).
- [4] H. Robbins, A generalization of the method of maximum likelihood: estimating a mixing distribution (abstract), Ann. Math. Statist. 21 (1950), 314-15.

N83

15778

UNCLAS

1. Introduction

Remote sensing of the atmosphere is a rapidly developing science. Today's meteorological satellites such as those in the TIROS-N series have high resolution instruments on board which measure the intensity of upwelling radiation in selected channel frequencies. A description of the data retrieved by the radiometers on the TIROS-N type satellites can be found in [7]. From these data it is possible to obtain information on the atmosphere's temperature, moisture and wind structure. One of the goals of the current Satellite Meteorology program is to improve the quality of atmospheric information obtained from satellite soundings to a point where it can be used for weather forecasting purposes. A major challenge in this direction is to develop refined numerical and statistical methods for inverting the equations of radiative transfer given a finite number of noisy measurements.

For a non-scattering atmosphere in local thermodynamic equilibrium the radiative transfer equations (RTE's) describe how the satellite upwelling radiance measurements relate to the underlying temperature distribution T:-

$$I_{\nu}(T) = B_{\nu}[T(p_0)]\tau_{\nu}(p_0) - \int_0^{p_0} B_{\nu}[T(p)]\frac{d}{dp}\tau_{\nu}(p)dp \quad (1.1)$$

where p_0 is the surface pressure, $\tau_{\nu}(p)$ is the transmittance of the atmosphere above pressure p at frequency ν , and B_{ν} is Planck's function given by:-

ORIGINAL PAGE IS
OF POOR QUALITY

$$\begin{aligned} B_v[T(p)] &= c_1 v^3 / \{\exp(c_2 v/T(p)) - 1\} \\ c_1 &= 1.19061 \times 10^{-5} \text{ erg-cm}^2\text{-sec}^{-1} \\ c_2 &= 1.43868 \text{ cm-deg(K)} \end{aligned} \quad (1.2)$$

The R.T.E's are of course an idealization. They describe the intensities the satellite radiometer would record in the absence of such things as atmospheric attenuation due to clouds or instrument noise. However, by using high resolution radiometers like the HIRS or AVHRR, sets of intensity measurements from many FOV's (fields of vision) can be combined to obtain data of the form

$$z_i = \int_{v_i} (T) + e_i \quad i = 1, \dots, n \quad (1.3)$$

where e_i 's are errors. These data relate to an area of about 119 by 140 km on the earth's surface. See [6] for more details.

We are interested in refining the methods used to obtain temperature distribution estimates from the above data. The procedure currently used to process TIROS-N temperature sounding data is a linear regression technique see [6]. Here we begin to discuss how the method of regularization (M.O.R.) might be used to improve the quality of temperature profiles obtainable by this procedure.

OF POOR QUALITY

Let T be the true temperature profile in the atmosphere. Then T can be written as

$$T = T_0 + \delta \quad (1.4)$$

where T_0 is the current best guess of T and δ is the update or correction to T_0 to be estimated from the data $\{z_i\}$ in hand. Using M.O.R. to estimate δ involves consideration of a functional I_λ given by

$$I_\lambda(\delta) = \sum_{i=1}^n [z_i - \mathcal{Q}_v(T_0 + \delta)]^2 + \lambda \int_0^{p_0} [\delta^{(m)}(p)]^2 dp \quad (1.5)$$

and picking the estimated update δ_λ to minimize this functional¹ over some class of physically plausible candidates, for instance the set of functions δ in $W_2^m[0, p_0]$ for which $T_0 + \delta$ is positive or perhaps, if the location of the temperature inversion were reliably known, one would look for minimizers of I_λ subject to an additional constraint involving temperature inversion.

The statistical reasoning for considering regularized estimates of this type is well documented in the literature, see for example [3] and [1]. Intuitively δ_λ has been designed to match the observed data and possess certain smoothness qualities. The parameter λ controls a tradeoff between

the smoothness of a solution (measured by $\int_0^{p_0} [\delta_\lambda^{(m)}(p)]^2 dp$) and how well it

[1] This corresponds to the case when the measurement errors are iid $N(0, \sigma^2)$. A more "robust" method would be to consider functionals of the form

$$I_\lambda(\delta) = \sum_{i=1}^n \rho[z_i - \mathcal{Q}_v(T_0 + \delta)] + \lambda \int_0^{p_0} [\delta^{(m)}(p)]^2 dp$$

where ρ reflected the possible non-Gaussian nature of the noise.

matches the data (the $\sum_{i=1}^n [z_i - \int_{\nu_i} (T_0 + \delta_\lambda)]^2$ term).

Inverting the R.T.E.'s with noisy data can be viewed as a special case of a more general situation in which the scientist wishes to estimate a function x given data

$$z_i = N_i(x) + e_i \quad i = 1, \dots, n \quad (1.6)$$

where x is in some Hilbert space H , the N_i 's are non-linear functionals and e_i 's are noise. Here, assuming the e_i 's are iid $N(0, \sigma^2)$, an appropriate regularization function I_λ is

$$I_\lambda(x) = \sum_{i=1}^n [z_i - N_i(x)]^2 + \lambda J(x) \quad (1.7)$$

where J is a roughness penalty functional on H . To estimate x one proceeds to minimize I_λ over some subset of physical interest in H . This report summarizes recent results we have obtained on the existence and numerical approximability of minimizers of such I_λ 's in certain subsets of H . We indicate how these results apply to the radiative transfer equations case.

There are three sections: section 2 talks about the existence theory; a Gauss-Newton algorithm for minimizing the regularization functionals is outlined in section 3, while the final section briefly describes how to estimate the smoothing parameter using a first order approximation to the generalized cross validation function given in [8]. We assume the reader is familiar with the basic mathematical tools for discussing minimization problems in Hilbert spaces. Part 1 of Ekeland and Temam's book [2] is an inspiring introduction to this subject.

OF POOR QUALITY

2. Existence Theory

Preliminaries

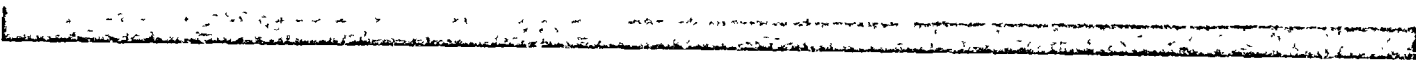
Before describing our main results, let's pause a moment to get our notation straight. H is a real Hilbert space with inner product $\langle \cdot, \cdot \rangle$ and norm $\|\cdot\|$ (so $\langle x, x \rangle = \|x\|^2$). P is a projection operator in H with finite dimensional null space; the complementary projection $I-P$ is denoted by P_0 . H^* is the dual space of H , i.e. the space of all continuous linear maps from H into R . $L(H, H^*)$ is the space of linear operators from H into H^* . We will discuss functionals, I say, acting on H (so $I: H \rightarrow R$). The first and second Frechet derivatives of I at some point $x \in H$ will be denoted by $I'(x)$ and $I''(x)$ respectively. Think of $I'(x)$ as an element of H^* and $I''(x)$ as an element of $L(H, H^*)$. Our concern here is with regularization functionals I_λ on H given by

$$I_\lambda(x) = \sum_{i=1}^n [z_i - N_i(x)]^2 + \lambda \|Px\|^2 \tag{2.1}$$

where N_i 's are functionals on H , z_i 's are in R , $x \in H$ and $\lambda > 0$. Whenever we write I_λ the form (2.1) will be what is meant. So we are considering regularization procedures in which the roughness penalty $J(x)$ is a semi-norm on H given by $J(x) = \|Px\|^2$.

Main Results

We now specify conditions on the non-linear functionals N_i which guarantee the existence of minimizers of I_λ in closed convex subsets K of H . In the R.T.E. case a reasonable choice for K is the set of all functions in $W_2^m[0, p_0]$ for which $T_0 + \delta$ is positive. It is very easy to check that this K is a closed convex subset of $W_2^m[0, p_0]$ for any m . Our existence results are summarized in the following three theorems.



UNIFORM BOUNDS
OF POOR QUALITY

Theorem 1 (proof in [2] pp. 34-35).

Let K be a closed convex subset of a Hilbert space H . Suppose $I_\lambda: K \rightarrow \mathbb{R}$ is coercive on K (i.e. $I_\lambda(x) \rightarrow \infty$ as $\|x\| \rightarrow \infty$ in K) and moreover that I_λ is weakly lower semi-continuous (w.l.s.c.) on K then I_λ attains its infimum on K .

Theorem 2 (proof in [4])

Let $\phi: \mathbb{R} \rightarrow \mathbb{R}$ be a monotonic increasing function in the modulus of its argument. Suppose

$$(i) \quad \sum_{i=1}^n \phi(N_i(x)) \text{ is convex on } K$$

$$(ii) \quad \sum_{i=1}^n \phi[N_i(x)] = \phi \Leftrightarrow P_0 x = P_0 \theta \text{ for some } \theta \text{ in } K$$

then I_λ is coercive on K .

Remark: The above theorem can be generalized somewhat but we refrain from doing so because the form given has more intuitive appeal.

Theorem 3

If N_i is weakly continuous (w.c.) on K for each i then I_λ is w.l.s.c. on K .

Proof: If the N_i are w.c., then it surely follows that $\sum_{i=1}^n [z_i - N_i(x)]^2$ is

w.c. But $\|Px\|^2$ is well known to be w.l.s.c. Therefore I_λ is w.l.s.c. QED

Application to the R.T.E.'s (see [4] for details)

The \int_{v_i} arising here can be shown to satisfy the hypotheses of Theorem 2 with ϕ taken to be

$$\phi(x) = |x|, \quad x \in \mathbb{R}$$

Also, each \int_{v_i} is w.c. We therefore have that for each $\lambda > 0, \exists \delta_\lambda \in K = \{\delta \in W_2^m[0, p_0] \mid T_0 + \delta > 0\}$, s.t.

$$I_\lambda(\delta_\lambda) = \min_{\delta \in K} \left\{ \sum_{i=1}^n [z_i - \int_{v_i}(T_0 + \delta)]^2 + \lambda \int_0^{p_0} [\delta^{(m)}(p)]^2 dp \right\}$$

There exist regularized solutions to the R.T.E.'s.

3. A numerical procedure for minimizing I_λ in K

Let x^k be the k^{th} approximation to the minimizer in K of I_λ . Define the functional I_λ^k on K as follows

$$I_\lambda^k(x) = \sum_{i=1}^n [z_i - N_i(x^k) - N_i'(x^k)[x - x^k]]^2 + \lambda \|Px\|^2 \quad (3.1)$$

each N_i is simply linearized about x^k . Define x^{k+1} to be the minimizer in K of I_λ^k .

Under suitable regularity conditions the iterates x^k are well defined and can be shown to satisfy

$$\begin{aligned} x^{k+1} &= x^k - \left\{ \sum_{i=1}^n N_i'(x^k) N_i'(x^k) + \lambda \langle P, \dots \rangle \right\}^{-1} I_\lambda'(x^k) \\ &\equiv x^k - A^{-1}(x^k) I'(x^k) \end{aligned} \quad (3.2)$$

ORIGINAL SOURCE
OF POOR QUALITY.

That this equation makes good sense is evident once one realizes that $A(x^k)$ belongs to $L(H, H^*)$ and $I_\lambda'(x^k)$ is in H^* .

Those in the know will have recognized that the above procedure is nothing more than an infinite dimensional version of the Gauss-Newton algorithm. The finite dimensional case is discussed in [5]. The major advantage of using a Gauss-Newton procedure to minimize our regularization functionals is the ease with which successive iterates can be obtained. At each stage we have a regularization problem involving linear functionals, the $N_i'(x^k)$'s, consequently we can take advantage of available software tools.

With the appropriate assumptions it is possible to show that the procedure is a decent method and the sequence x^k converges at least R-linearly to a critical point of I_λ in K .

Theorem 4 (proof in [4]).

Suppose that the $N_i(\cdot)$'s are twice continuously differentiable and $N_i'(\cdot)$'s are w.c. on $\text{int } K$. Let $x^0 \in \text{int } K$ be such that

$$L^0 = \{x \mid I_\lambda(x) < I_\lambda(x^0)\}$$

is weakly compact and I_λ has only finitely many critical points in L^0 . Moreover, suppose that μ_0, μ_1, γ_1 all positive with $\mu_0 - 1/2\gamma_1 > 0$ satisfying

$$\mu_0 \|h\|^2 < \langle h, A(x)h \rangle < \mu_1 \|h\|^2, \quad I_\lambda''(x)hh < \gamma_1 \|h\|^2 \quad \forall x \in L^0, h \in H$$

then the sequence of iterates $\{x^k\} \subseteq L^0$, $\lim_k x^k = x^*$ where $I_\lambda'(x^*) \equiv 0$ and if $I_\lambda''(x^*)$ is non-singular, then the convergence is at least R-linear.

The proof follows an argument similar to that used in 14.4.6 of [5].

4. The choice of λ

The generalized cross validation method for choosing λ works as follows.

Let $x_\lambda^{[k]}$ be the minimizer² in K of

$$\sum_{\substack{i=1 \\ i \neq k}}^n [z_i - N_i(x)]^2 + \lambda \|Px\|^2 \quad (4.1)$$

Then λ is chosen to minimize

$$V(\lambda) = \frac{\frac{1}{n} \sum_{k=1}^n [z_k - N_k(x_\lambda^{[k]})]^2}{\left[1 - \frac{1}{n} \sum_{k=1}^n a_{kk}^*(\lambda)\right]^2} \quad (4.2)$$

where $N_k(x_\lambda^{[k]})$ is the prediction of z_k given the data $z_1, z_2, \dots, z_{k-1}, z_{k+1}, \dots, z_n$ and $a_{kk}^*(\lambda)$ is the "differential influence" of the z_k 'th data point on the estimate x_λ (x_λ is the minimizer in K of I_λ).

$$a_{kk}^*(\lambda) = \frac{N_k(x_\lambda^{[k]}) - N_k(x_\lambda)}{N_k(x_\lambda^{[k]}) - z_k} \quad (4.3)$$

From a computational viewpoint $V(\lambda)$ is prohibitively expensive so one needs to find some convenient approximation. Following Wahba [8], $V(\lambda)$ can be approximated by

$$v_{\text{approx}}(\lambda) = \frac{\frac{1}{n} \sum_{k=1}^n [z_k - N_k(x_\lambda)]^2}{[1 - \mu_1(\lambda)]^2} \quad (4.4)$$

[2] Assumed to be uniquely defined.

ORIGINAL PAGE IS
OF POOR QUALITY

where μ_1 given by

$$\mu_1(\lambda) = \frac{1}{n} \sum_{k=1}^n \frac{\partial N_k(x_\lambda)}{\partial z_k}$$

is an easily computed functional of x_λ . We hope to study this procedure more closely in the near future.

Acknowledgements

This work was done with the help of my thesis advisor, Professor Grace Wahba.

REFERENCES

- [1] D.D. Cox, "Asymptotics for M-Type Smoothing Splines," Statistics Department, University of Wisconsin, Madison, Technical Report No. 654, November 1981.
- [2] J. Ekeland and R. Teman, Analyse Convexe et Problems Variationelles, Herman, Paris (1973).
- [3] G.S. Kimeldorf and G. Wahba, "A correspondence between Bayesian Estimation in Stochastic Processes and Smoothing by Splines," Annals of Mathematical Statistics A1, pp. 495-562 (1970).
- [4] F. O'Sullivan, Thesis to appear. (1982)
- [5] J.M. Ortega and W.C. Rheinbold, Iterative Solutions of Non-linear Equations to Several Variables, Academic Press (1970).
- [6] W.L. Smith and H. M. Woolf, "The Use of Eigenvectors of Statistical Covariance Matrices for Interpreting Sastellite Sounding Radiometer Observations," Journal of the Atmospheric Sciences 33, 7, pp. 1127-1140, July 1976.
- [7] W.L. Smith, H.M. Woolf, C.M. Hayden, D.O. Wark, and L.M. McMillin, "The TIROS-N Operational Vertical Sounder," Bulletin of the American Meteorological Society 50, 10, pp. 1177-1187, October 1979.
- [8] G. Wahba, "Constrained Regularization for Ill-Posed Linear Operator Equations, with Applications in Meteorology and Medicine," Statistics Department, University of Wisconsin, Madison, Technical Report No. 646, August 1981.

N83

15779

UNCLAS

QUANTILES, PARAMETRIC-SELECT DENSITY ESTIMATION,
AND BI-INFORMATION PARAMETER ESTIMATORS

by

EMANUEL PARZEN
Institute of Statistics
Texas A&M University

Abstract

This paper outlines a quantile-based approach to statistical analysis and probability modeling of data which formulates statistical inference problems as functional inference problems in which the parameters to be estimated are density functions. Density estimators can be non-parametric (computed independently of model identified) or parametric-select (approximated by finite parametric models that can provide standard models whose fit can be tested). Exponential models and autoregressive models are approximating densities which can be justified as maximum entropy for respectively the entropy of a probability density and the entropy of a quantile density. Applications of these ideas are outlined to the problems of modeling: (1) univariate data; (2) bivariate data and tests for independence; and (3) two samples and likelihood ratios. It is proposed that bi-information estimation of a density function can be developed by analogy to the problem of identification of regression models.

Research supported by the Army Research Office Grant
DAAG29-80-C-0070.

CONTENTS

1. Statistical Science, Data Analysis, and Buffalo Snowfall
2. Functions that describe probability distributions
3. Raw functions that describe samples
4. Smooth functions that describe samples and estimate probability distributions
5. Parameter estimation and information divergence
6. Information and bi-information parameter estimation, and comparison distribution functions
7. Statistical inference reduced to density estimation
8. Parametric-select density estimation and maximum entropy densities
9. Exact-parametric and parametric-select estimation of probability density functions using exponential models
10. Case studies of bi-information density estimation

1. Statistical Science, data analysis, and Buffalo snowfall

Statisticians complain about the failure of universities to adequately educate students on how to analyze statistical data. At the same time some statisticians state that data analysis is an art, and thus cannot be taught. When these statisticians speak of statistical science it is difficult to imagine to what they are alluding since they seem to sneeringly reject all attempts to reason, and reach consensus, about the evaluation of methods to be used as part of the process of statistical data analysis.

I would like to propose a data set which I believe provides a useful test case for various approaches to data analysis, namely the annual time series of snowfall in Buffalo, N.Y. The segment of that series which I will discuss is 1910-1972, although it has many interesting features when extended to 1981. The data analysis question to be considered is What probability distributions can be used to describe Buffalo snowfall. An ever-present hypothesis to be considered is whether Buffalo snowfall is normal.

2. Functions that describe probability distributions

The probability law of a continuous random variable X can be described by one or more of the following functions:

(1) Distribution Function $F(x) = \Pr [X \leq x]$

(2) Probability Density Function $f(x) = F'(x)$

- (3) Quantile Function $Q(u) = F^{-1}(u)$
 $= \inf \{x: F(x) \geq u\}$
 $= \inf \{x: F(x) = u\}$ if F is continuous
 $= x$ such that $F(x) = u$ if F increasing at x

(4) Quantile-Density Function $q(u) = Q'(u)$

(5) Density-Quantile Function $fQ(u) = f(Q(u))$

Theorem: For F continuous

$$FQ(u) = u, \quad fQ(u) q(u) = 1$$

3. Raw functions that describe samples

Data X_1, \dots, X_n is called a random sample of X when X_1, \dots, X_n are independent random variables identically distributed as X . An important role in the analysis of a sample is played by the order statistics $X_{(1)} < X_{(2)} < \dots < X_{(n)}$

- (1) Sample Distribution $\tilde{F}(x) = \text{fraction } X_1, \dots, X_n \leq x$
 $= \frac{j}{n}, X_{(j)} \leq x < X_{(j+1)}$

(2) Sample Probability Density, or Histogram, estimates $f(x)$ by a numerical derivative

$$\tilde{f}(x) = \frac{\tilde{F}(x+h) - \tilde{F}(x-h)}{2h}$$

- (3) Sample Quantile $\tilde{Q}(u) = \tilde{F}^{-1}(u)$
 $= X_{(j)}, \frac{j-1}{n} < u \leq \frac{j}{n}$

A universal display of any data set is provided by the quantile box plot introduced in Parzen (1979).

(4) Sample Quantile-Density is a numerical derivative

$$\tilde{q}(u) = \frac{\tilde{Q}(u+h) - \tilde{Q}(u-h)}{2h}$$

(5) Sample Density-Quantile = $\tilde{f}\tilde{Q}(u) = 1/\tilde{q}(u)$.

An important formula is

$$\tilde{f}(X_{(j)}) = \tilde{f}\tilde{Q}\left(\frac{j}{n+1}\right) = 2 \{(n+1)(X_{(j+1)} - X_{(j-1)})\}^{-1}$$

4. Smooth functions that describe samples and estimate probability distributions

The functions F, f, Q, q, fQ that represent the true probability distribution of a random variable X are estimated by smooth functions $\hat{F}, \hat{f}, \hat{Q}, \hat{q}, \hat{f}\hat{Q}$ which are derived from the raw descriptive functions $\tilde{F}, \tilde{f}, \tilde{Q}, \tilde{q}, \tilde{f}\tilde{Q}$. One distinguishes between parametric and non-parametric methods of estimating smooth functions.

A parametric estimation method : (1) assumes a family $F_\theta, f_\theta, Q_\theta, q_\theta, f_\theta Q_\theta$ of functions, called parametric models, which are indexed by a parameter $\theta = (\theta_1, \dots, \theta_k)$; (2) forms estimators $\hat{\theta} = (\hat{\theta}_1, \dots, \hat{\theta}_k)$ of θ ; (3) forms smooth functions by

$$\begin{aligned} \hat{F}(x) &= F_{\hat{\theta}}(x), \hat{f}(x) = f_{\hat{\theta}}(x), \\ \hat{Q}(u) &= Q_{\hat{\theta}}(u), \hat{q}(u) = q_{\hat{\theta}}(u), \\ \hat{f}\hat{Q}(u) &= f_{\hat{\theta}}Q_{\hat{\theta}}(u). \end{aligned}$$

A non-parametric estimation method forms estimators which are not based on parametric models. Important examples of non-parametric estimators of a probability density f(x) and a

quantile-density $q(u)$ are respectively

$$\hat{f}(x) = \frac{1}{\delta} \int_{-\infty}^{\infty} K\left(\frac{x-y}{\delta}\right) d\tilde{F}(x)$$

$$\hat{q}(u) = \frac{1}{\delta} \int_0^1 K\left(\frac{u-t}{\delta}\right) d\tilde{Q}(u)$$

for suitable kernels $K(\cdot)$ and bandwidth δ .

5. Parameter estimation and information divergence

When a parametric model f_{θ} is assumed, parameter estimators $\hat{\theta}$ are often determined by minimizing a "distance" between $\tilde{f}(x)$ and $f_{\theta}(x)$. A "distance" between two probability densities $f(x)$ and $g(x)$ is denoted $I(f;g)$ and is called an information divergence between $f(x)$ and $g(x)$. It is usually not symmetric in f and g . It does not satisfy the triangle inequality for a metric. But it does satisfy $I(f;g) \geq 0$ and $I(f;g) = 0$ if and only if $f = g$.

The most famous, and most important, definition of information divergence is

$$I_1(f;g) = \int_{-\infty}^{\infty} -\log\left\{\frac{g(x)}{f(x)}\right\} f(x) dx$$

called the information divergence of order 1, or Kullback-Liebler information divergence. Information divergence of order α is defined for $\alpha > 0$ (but $\alpha \neq 1$) by

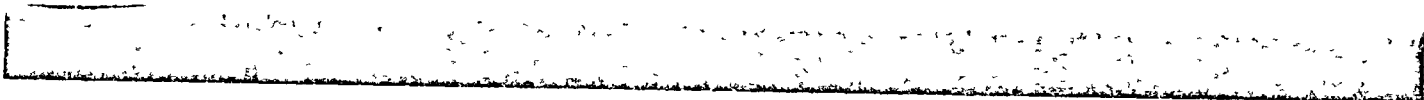
$$I_{\alpha}(f;g) = \frac{-1}{1-\alpha} \log \int_{-\infty}^{\infty} \left\{\frac{g(x)}{f(x)}\right\}^{1-\alpha} f(x) dx.$$

The most important values of α are $0.5 \leq \alpha \leq 2$.

Bi-information divergence is defined by

$$II(f;g) = \int_{-\infty}^{\infty} \left| \log \left\{ \frac{g(x)}{f(x)} \right\} \right|^2 f(x) dx;$$

it may be regarded as related to $I_2(g;f)$.



INFORMATION DIVERGENCE OF ORDER α

$$I_1(f;g) = H(f;g) - H(f)$$

defining

$$H(f;g) = \int_{-\infty}^{\infty} \{-\log g(x)\} f(x) dx,$$

$$H(f) = H(f;f) = \int_{-\infty}^{\infty} \{-\log f(x)\} f(x) dx.$$

We call $H(f;g)$ the cross-entropy of f and g , and call $H(f)$ the entropy of f .

Maximum likelihood parameter estimation can be shown to be equivalent to minimum cross-entropy estimation. The likelihood function of a parametric model f_θ is defined by

$$\begin{aligned} L(f_\theta) &= \log f_\theta(X_1, \dots, X_n) \\ &= \sum_{t=1}^n \log f_\theta(X_t) \end{aligned}$$

One may verify that

$$\begin{aligned} L(f_\theta) &= n \int_{-\infty}^{\infty} \log f_\theta(x) d\tilde{F}(x) \\ &= -n H(\tilde{f}; f_\theta). \end{aligned}$$

The maximum likelihood parameter estimator $\hat{\theta}$, defined by

$$L(f_{\hat{\theta}}) = \max_{\theta} L(f_\theta),$$

clearly satisfies

$$H(\tilde{f}; f_{\hat{\theta}}) = \min_{\theta} H(\tilde{f}; f_\theta).$$

It also satisfies

$$I_1(\tilde{f}; f_{\hat{\theta}}) = \min_{\theta} I_1(\tilde{f}; f_\theta).$$

In general parameter estimators $\hat{\theta}$ are found by minimizing $I_\alpha(\tilde{f}; f_\theta)$ or $I_\alpha(f_\theta, \tilde{f})$. Chi-squared estimators minimize $I_2(f_\theta; \tilde{f})$ while modified chi-squared estimators minimize $I_2(\tilde{f}; f_\theta)$.

To compute $I_1(\tilde{f}; f_\theta)$ one needs to compute $H(\tilde{f})$. A useful formula for accomplishing this is

$$\begin{aligned} H(f) &= \int_{-\infty}^{\infty} \{-\log f(x)\} dF(x) \\ &= \int_0^1 \{-\log fQ(u)\} du \\ &= \int_0^1 \log q(u) du. \end{aligned}$$

The value of $I_1(\tilde{f}; f_{\hat{\theta}})$ can be used to test the goodness of fit of the parametric model f_θ .

6. Information and bi-information parameter estimation, and comparison distribution functions

Given a sample with sample probability density function \tilde{f} and parametric model f_θ , one can form diverse parameter estimators, denoted $\hat{\theta}$ and $\check{\theta}$, corresponding to two choices of information divergence which we take to be: (1) $I_1(\tilde{f}; f_\theta)$, and (2) $I_2(f_\theta; \tilde{f})$ or $II(\tilde{f}, f_\theta)$. We call $\hat{\theta}$ and $\check{\theta}$ diverse parameter estimators. For greater precision we call $\hat{\theta}$ the (order 1) information estimator, and $\check{\theta}$ the bi-information estimator.

When the parametric model f_θ is exact, the diverse parameter estimators have equivalent statistical properties; they are both asymptotically efficient estimators, and are not significantly different from each other.

When the values of $\hat{\theta}$ and $\check{\theta}$ computed from a sample are significantly different one should suspect that the parametric model f_θ does not fit the data. The Shapiro-Wilk statistics

for testing normality and exponentiality can be regarded as comparing diverse estimators which minimize information of order 1 and 2 respectively.

One can interpret $\hat{\theta}$ and $\check{\theta}$ as parameter values of "best approximating" models.

One wishes to evaluate $F_{\hat{\theta}}(x)$ and $F_{\check{\theta}}(x)$ as smooth estimators of $F(x)$. For any parameter value θ , define

$$\tilde{D}_{\theta}(u) = F_{\theta}(\tilde{Q}(u))$$

which is the sample quantile function of the transformed random variables

$$U_1 = F_{\theta}(X_1), \dots, U_n = F_{\theta}(X_n).$$

The true parameter value θ has the property that U_1, \dots, U_n are distributed with a uniform $[0,1]$ distribution. Then parameter estimators $\hat{\theta}$ and $\check{\theta}$ are compared by the character of the closeness to the identity function $D(u) = u$ of $\tilde{D}_{\hat{\theta}}(u)$ and $\tilde{D}_{\check{\theta}}(u)$.

We call $\tilde{D}_{\theta}(u)$ a comparison distribution function. Its derivative

$$\tilde{d}_{\theta}(u) = \{\tilde{D}_{\theta}(u)\}'$$

plays a basic role and is called a comparison density; formulas for the comparison density are

$$\begin{aligned} \tilde{d}_{\theta}(u) &= f_{\theta}(\tilde{Q}(u)) \tilde{q}(u) \\ &= \frac{f_{\theta}(\tilde{Q}(u))}{\tilde{f} \tilde{Q}(u)} \end{aligned}$$

An alternative comparison density introduced in Parzen (1979), is

$$\tilde{d}(u) = f_0 Q_0(u) \tilde{q}(u) \div \tilde{\sigma}_0,$$

$$\tilde{\sigma}_0 = \int_0^1 f_0 Q_0(u) \tilde{q}(u) du,$$

$$\tilde{D}(u) = \int_0^u \tilde{d}(t) dt$$

where $f_0 Q_0(u)$ is a specified density-quantile function.

Parameter estimators can be justified as minimizing information divergence

$$I_1(\tilde{d}_\theta) = \int_0^1 -\log \tilde{d}_\theta(u) du = I_1(\tilde{f}; f_\theta)$$

$$II(\tilde{d}_\theta) = \int_0^1 |\log \tilde{d}_\theta(u)|^2 du = II(\tilde{f}; f_\theta)$$

$$I_\alpha(\tilde{d}_\theta) = \frac{-1}{1-\alpha} \log \int_0^1 \{\tilde{d}_\theta(u)\}^{1-\alpha} du$$

$$\int_0^1 |\tilde{d}_\theta(u) - 1|^2 du = \int_0^1 |\tilde{d}_\theta(u)|^2 du - 1$$

These measure the closeness to 1 of $\tilde{d}_\theta(u)$, or the closeness to $D(u) = u$ of $\tilde{D}_\theta(u)$. However the final decision about parameter estimators should be based on visual inspection of the graph of $\tilde{D}_\theta(u)$.

10

ORIGINAL PAPERS
OF POOR QUALITY

Another consequence of considering information of order α is that we can unify the estimation criterion used to form maximum likelihood estimators with the estimation criterion used to form Gaussian time series parameter estimators:

$$I_{sp}(\tilde{f}; f_{\theta}) = \log \int_0^1 \frac{\tilde{f}(w)}{f_{\theta}(w)} dw .$$

where \tilde{f} and f_{θ} are spectral densities. It is comparable to

$$I_2(\tilde{d}_{\theta}) = \log \int_0^1 \frac{\tilde{f}Q(u)}{f_{\theta}Q(u)} du$$

7. Statistical inference reduced to density estimation

The quantile approach to statistical data analysis being developed by Parzen [since Parzen (1979)] is based on the proposition that conventional problems of statistical inference concerning (1) a random sample X_1, \dots, X_n , (2) a bivariate sample $(X_1, Y_1), \dots, (X_n, Y_n)$, or (3) two samples X_1, \dots, X_m and Y_1, \dots, Y_n should be transformed to problems of functional inference, estimating and testing hypotheses about density functions $d(u)$, $d(u_1, u_2), \dots, d(u_1, \dots, u_k)$, on the unit interval $0 \leq u \leq 1$, unit square $0 \leq u_1, u_2 \leq 1$, unit hypercube $0 \leq u_1, \dots, u_k \leq 1$. To illustrate how this is done consider the following problems.

Modeling Bivariate Data and Tests for Independence. Let X and Y be continuous random variables with joint density function $f_{X,Y}(x,y)$. The hypothesis, H_0 : X and Y are independent can be expressed

$$H_0: f_{X,Y}(x,y) = f_X(x) f_Y(y)$$

or in terms of information divergence

$$I(f_{X,Y}; f_X f_Y) = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} \left\{ -\log \frac{f_X(x) f_Y(y)}{f_{X,Y}(x,y)} \right\} f_{X,Y}(x,y) \, dx \, dy$$

by

$$H_0: I(f_{X,Y}; f_X f_Y) = 0 \quad .$$

Define

$$D(u_1, u_2) = F_{X,Y}(Q_X(u_1), Q_Y(u_2))$$

$$d(u_1, u_2) = \frac{\partial^2}{\partial u_1 \partial u_2} D(u_1, u_2)$$

$$= \frac{f_{X,Y}(Q_X(u_1), Q_Y(u_2))}{f_X(Q_X(u_1)) f_Y(Q_Y(u_2))}$$

We call $d(u_1, u_2)$ the quantile dependence density.

The hypothesis H_0 can be expressed

$$H_0: D(u_1, u_2) = u_1 u_2, \quad d(u_1, u_2) = 1.$$

One can verify that

$$I_1(f_{X,Y}; f_X f_Y) = \int_0^1 \int_0^1 \{\log d(u_1, u_2)\} d(u_1, u_2) du_1 du_2$$

$$= - H_1(d(u_1, u_2))$$

Thus estimating the information divergence between $f_{X,Y}$ and $f_X f_Y$ is equivalent to estimating the negative of the entropy of $d(u_1, u_2)$.

Estimators $\hat{d}_m(u)$ dependent on a finite number of parameters can be formed from the raw estimator

$$\tilde{D}(u_1, u_2) = \tilde{F}_{X,Y}(\tilde{Q}_X(u_1), \tilde{Q}_Y(u_2)).$$

Modeling likelihood ratios and testing equality of distributions. Let X and Y be continuous random variables.

The hypothesis

$$H_0: F_X(x) = F_Y(x), \text{ or } f_X(x) = f_Y(x)$$

can be expressed in terms of information divergence

$$\begin{aligned}
 I(f_Y; f_X) &= \int_{-\infty}^{\infty} -\log \frac{f_X(x)}{f_Y(x)} dF_Y(y) \\
 &= \int_0^1 -\log d(u) du \\
 &= -H_q d(d(u))
 \end{aligned}$$

defining the comparison distribution function and comparison density function

$$D(u) = F_X Q_Y(u), \quad d(u) = \frac{d}{du} D(u) = \frac{f_X(Q_Y(u))}{f_Y(Q_Y(u))}$$

Estimating the information divergence between f_Y and f_X is equivalent to estimating the negative of the entropy in the quantile-density sense of the comparison density $d(u)$.

8. Parametric-select density estimation and Maximum Entropy Densities

A density $d(u) = D'(u)$ can be approximated in many ways by sequences $d_m(u), m=1,2,\dots$ of functions which converge to $d(u)$. For $m=1,2,\dots$, let $\hat{d}_m(u)$ be an estimator of $d_m(u)$; the sequence $\hat{d}_m(u)$ then estimates $d(u)$.

If $d_m(u)$ corresponds to a standard finite parametric model $d(u)$ for which one could consider testing the hypothesis that $d_m(u)$ provides an exact model, we call $d_m(u)$ a parametric-select representation, and $\hat{d}_m(u)$ a parametric-select estimator,

OF POOR QUALITY

to indicate that we are free to select the number of parameters in $d_m(u)$ to provide an adequate approximation or representation of $d(u)$.

We call $d_m(u)$ a non-parametric representation, and $\hat{d}_m(u)$ a non-parametric estimator, if $d_m(u)$ does not correspond to a standard finite parameter model which could be interpreted as an exact model.

An important criterion for developing the functional form of exact models for densities is the maximum entropy principles.

A density $f(x)$, $-\infty < x < \infty$, which maximizes entropy $H(f) = \int_{-\infty}^{\infty} \{-\log f(x)\} f(x) dx$ subject to constraints

$$\int_{-\infty}^{\infty} T_j(x) f(x) dx = \tau_j, \quad j=1, \dots, k,$$

where $T_j(x)$ are specified functions (called sufficient statistics) and τ_j are specified moments can be shown to have the representation, called an exponential model,

$$\log f(x) = \sum_{j=1}^k \theta_j T_j(x) - \psi(\theta_1, \dots, \theta_k)$$

where

$$\psi(\theta_1, \dots, \theta_k) = \log \int_{-\infty}^{\infty} \exp \left\{ \sum_{j=1}^k \theta_j T_j(x) \right\} dx$$

guarantees that $f(x)$ integrates to 1.

A quantile function $q(u)$, $0 < u < 1$, which maximizes entropy $H_q(q) = \int_0^1 \log q(u) du$ subject to the constraints

$$\frac{\int_0^1 \exp(2\pi iuv) f_0 Q_0(u) q(u) du}{\int_0^1 f_0 Q_0(u) q(u) du} = \rho(v), \quad v=0, \pm 1, \dots, \pm m$$

where $f_0 Q_0(u)$ is a specified density quantile function must have the representation, called an autoregressive model,

$$q(u) = q_0(u) \sigma_m^2 |1 + \alpha_m(1)e^{2\pi iu} + \dots + \alpha_m(m)e^{2\pi ium}|^{-2}$$

9. Exact-Parametric and Parameter-select Estimation of Probability density Functions using Exponential Models

Two important exponential models for a density $f(x)$, $-\infty < x < \infty$ are the normal density and the gamma density.

The normal density, denoted Normal (μ, σ)

$$f_{\mu, \sigma}(x) = \frac{1}{\sigma} \phi\left(\frac{x-\mu}{\sigma}\right),$$

$$\phi(x) = \frac{1}{\sqrt{2\pi}} \exp - \frac{1}{2} x^2$$

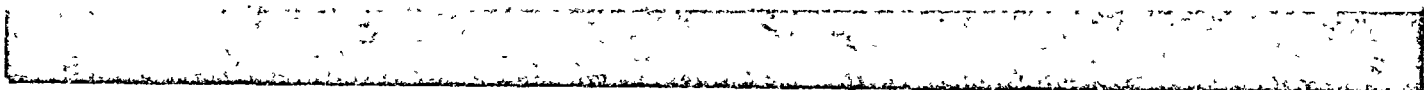
is exponential with sufficient statistics $T_1(x) = x$ and $T_2(x) = x^2$.

The Gamma density, denoted Gamma (r, λ) where $\lambda = 1/\sigma$,

$$f_{r, \sigma}(x) = \frac{1}{\sigma} f_r\left(\frac{x}{\sigma}\right),$$

$$f_r(x) = \frac{1}{\Gamma(r)} x^{r-1} e^{-x}, \quad x > 0,$$

$$= 0, \quad x < 0,$$



ORIGINAL PAGE IS
OF POOR QUALITY

as exponential with sufficient statistics $T_1(x) = x$ and $T_2(x) = \log x$.

A location scale parameter Gamma density

$$f_{r, \mu, \sigma}(x) = \frac{1}{\sigma} f_r\left(\frac{x-\mu}{\sigma}\right)$$

is not an exponential model. We can treat it as one by estimating μ (say, by the minimum $X_{(1)}$ of the random sample X_1, \dots, X_n), and treating $X_j - \hat{\mu}$ as a sample from $f_{r, \sigma}(x)$.

The hypothesis that the data is fit by a normal distribution versus the hypothesis that the data is fit by a Gamma distribution can be tested by forming an over-parametrized exponential model with sufficient statistics

$$T_1(x) = x, \quad T_2(x) = x^2, \quad T_3(x) = x^3, \quad T_4(x) = \log x.$$

The (order 1) information divergence, or maximum likelihood, estimators $\hat{\theta}_1, \hat{\theta}_2, \hat{\theta}_3, \hat{\theta}_4$, which minimize information divergence of order 1 $\int_0^1 -\log \tilde{d}_\theta(u) du$, may be found for an exponential model by solving

$$\hat{\tau}_j = E_{\hat{\theta}}[T_j]$$

where $\tau_j = E_\theta[T_j]$ is estimated by

$$\hat{\tau}_j = \bar{T}_j = \frac{1}{n} \sum_{j=1}^n T_j(X_{(j)})$$

The bi-information divergence estimators $\check{\theta}_1, \check{\theta}_2, \check{\theta}_3, \check{\theta}_4$, which minimize information divergence $\int_0^1 |\log \tilde{d}_\theta(u)|^2 du$, may be found using least squares regression analysis techniques by minimizing with respect to $\theta_1, \dots, \theta_k$ the sum of squares

$$\sum_{j=2}^{n-1} |\log \tilde{f}(X_{(j)}) - \{\log \tilde{f}(X_j)\}^- - \theta_1 (T_1(X_{(j)}) - \bar{T}_1) - \dots - \theta_k (T_k(X_{(j)}) - \bar{T}_k)|^2$$

Stepwise regression is used to suggest parsimonious parametrizations.

Graphical procedures to determine which parameter values fit best are as follows: estimate $\tilde{D}_\theta(\frac{j}{n+1})$, $j=2, \dots, n-1$, by adding

$$\tilde{d}_\theta(\frac{j}{n+1}) = f_\theta(X_{(j)}) \div \tilde{f}(X_{(j)})$$

and normalizing the sum to go from 0 to 1. One inspects its graph to see how it deviates from $D(u) = u$.

10. Case studies of bi-information density estimation

The density estimators corresponding to the bi-information parameter estimates of the normal, gamma, and four-parameter exponential models are presented for four simulated random samples:

- 1) Exponential or Gamma ($r = 1, \sigma = 1$)
- 2) Gamma ($r=10, \sigma = 1$)

- 3) Normal ($\mu = 0, \sigma = 1$),
- 4) Contaminated normal: $100N(0,1), 5N(10,1)$

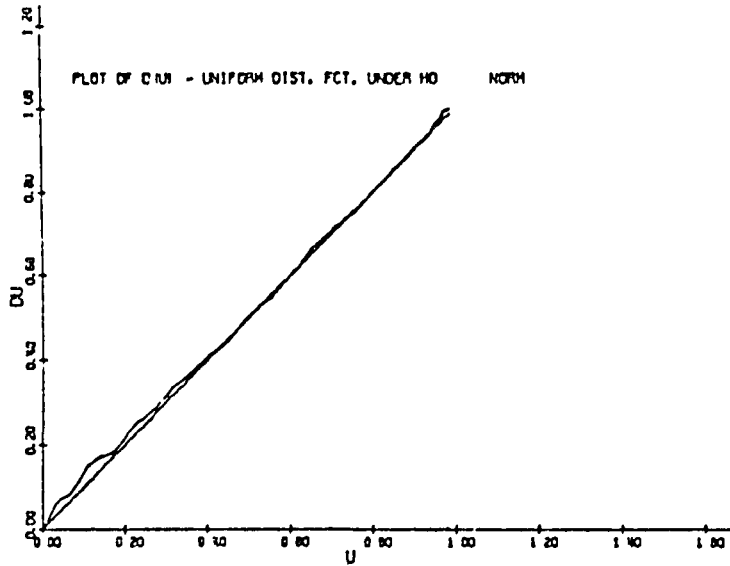
In addition density estimators, using bi-information parameters, are presented for the data set of Buffalo snowfall. Bi-information select regression estimation of the parameters of a 4-parametrial exponential model with sufficient statistics x, x^2, x^3 , and $\log x$ leads to the conclusion that Buffalo snowfall obeys a Gamma distribution. It is equally well fit by a normal distribution whose parameters are estimated by minimizing bi-information rather than order 1 information. The hypothesis that Buffalo snowfall is normal seems to be acceptable, but one can question whether the maximum likelihood estimators (sample mean and variance) provide the best-fitting normal distribution for Buffalo snowfall.

As in Parzen (1979), we reject a trimodal shape probability density estimate for Buffalo snowfall, which has been found by several non-parametric density estimation techniques; including Tapia and Thompson (1978).

REFERENCES

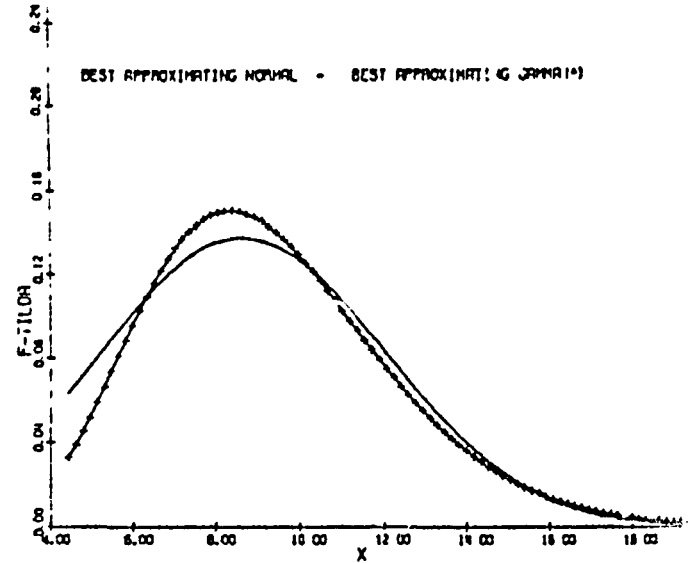
- Parzen, E. (1979) Nonparametric statistical data modeling. Journal of the American Statistical Association, 74, 105-131.
- _____. (1982) Maximum entropy interpretation of autoregressive spectral densities. Submitted for publication.
- Tapia, R. A. and Thompson, J. R. (1978) Nonparametric Probability Density Estimation, Baltimore: Johns Hopkins University Press.

D(u) FOR BI-INF-DIV NORMAL (8.64, 11.53)

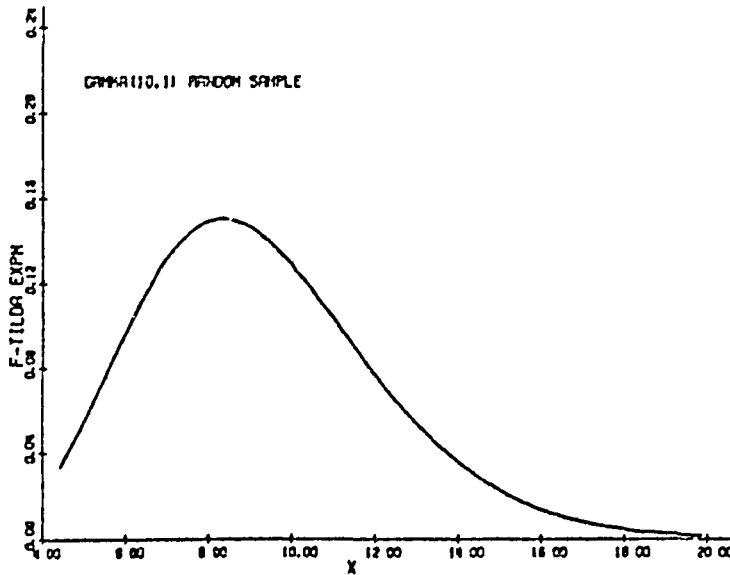


GAMMA (10,1) SIMULATED SAMPLE

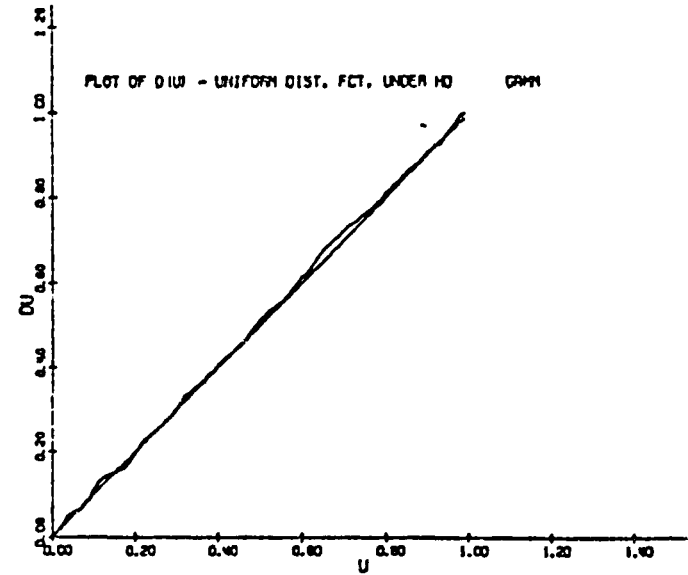
DENSITY ESTIMATES EVALUATED ON SAMPLE RANGE



PARSIMONIOUS 4-PARAMETER EXPONENTIAL DENSITY IS
GAMMA (10.08, 1.09)



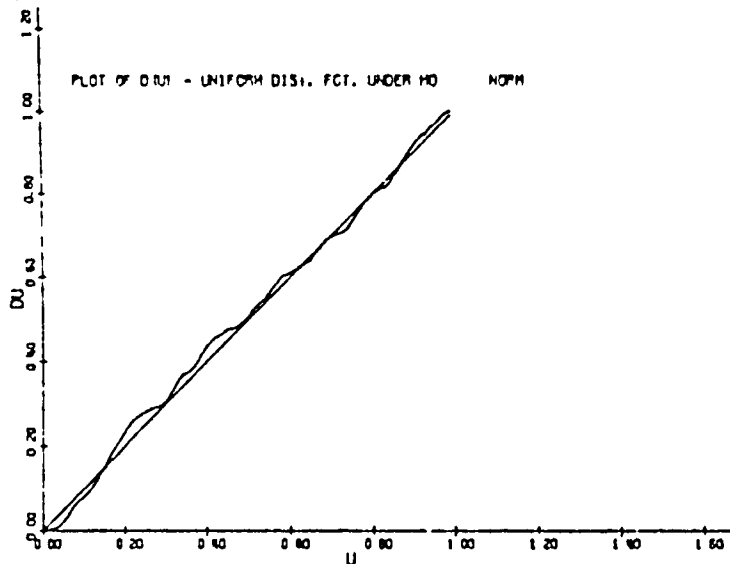
D(u) FOR BI-INF-DIV GAMMA (10.08, 1.09)



OF POOR QUALITY

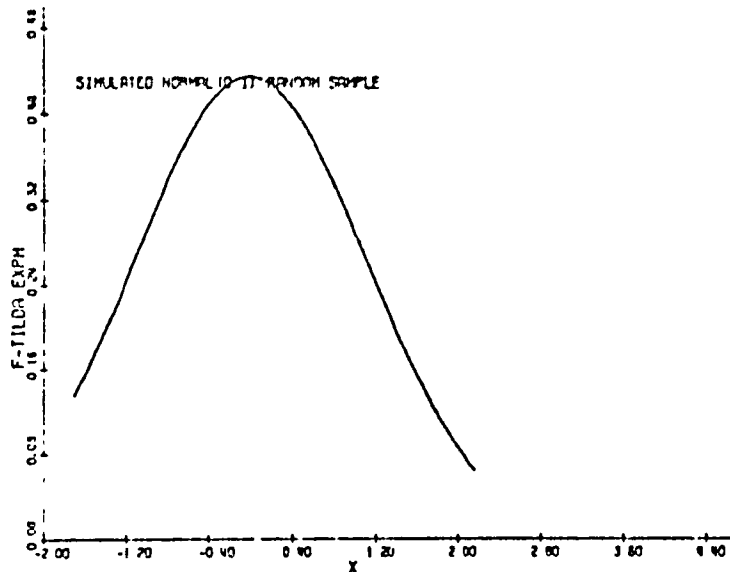
RE

$D(u)$ FOR BI-INF-DIV NORMAL (.03, 1.22)

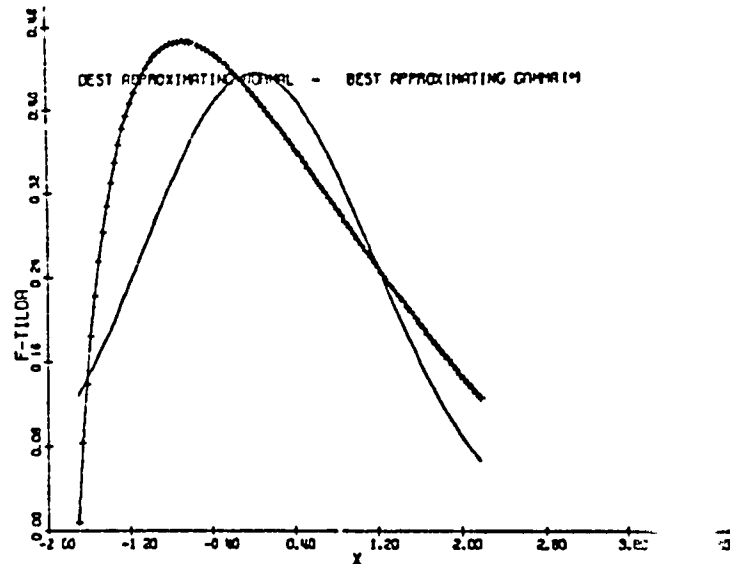


NORMAL (0, 1) SIMULATED SAMPLE
 Sample Mean .11, Variance .82

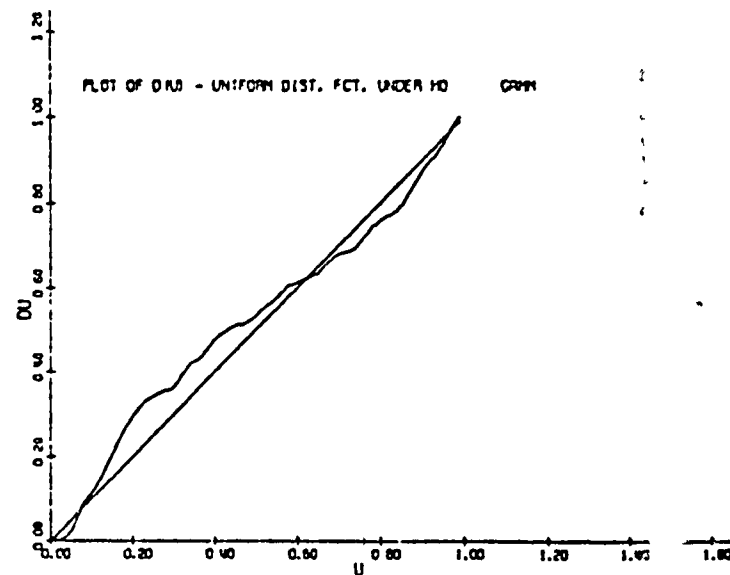
PARSIMONIOUS 4-PARAMETER EXPONENTIAL DENSITY IS
 NORMAL (.03, 1.22)



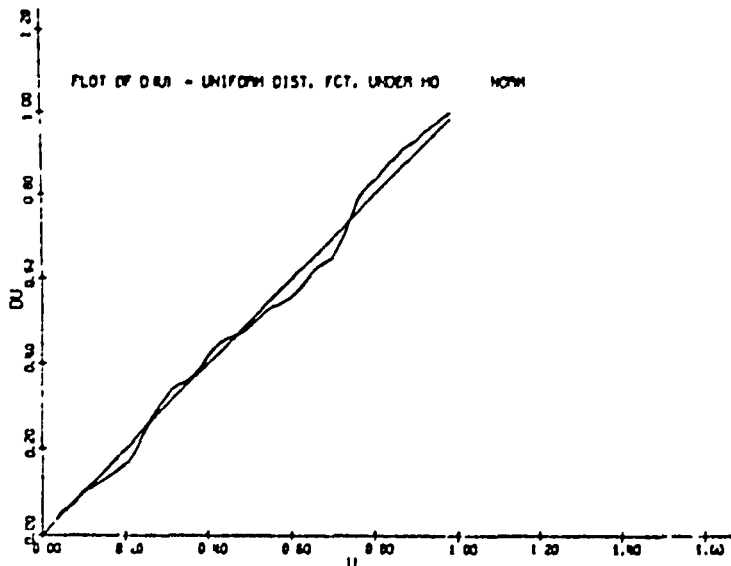
DENSITY ESTIMATES EVALUATED ON SAMPLE RANGE



$\tilde{D}(u)$ FOR BI-INF-DIV GAMMA (1.56, .53)

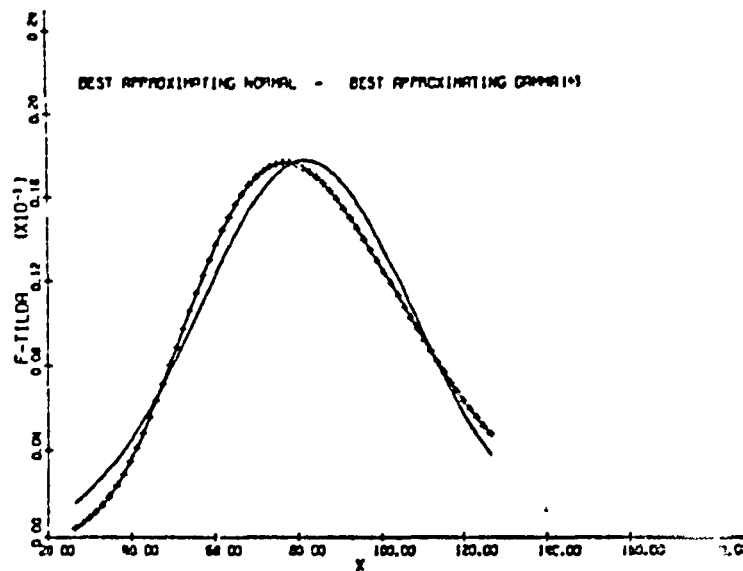


$D(u)$ FOR BI-INF-DIV NORMAL (81.9, 644)

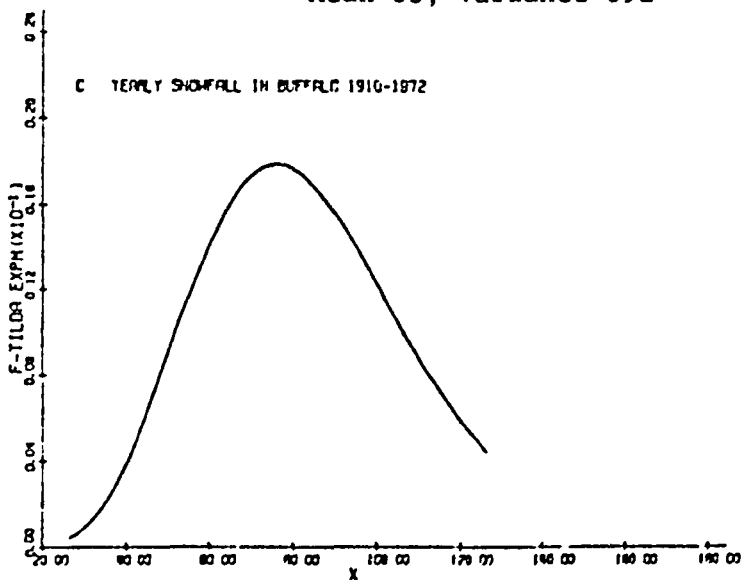


YEARLY SNOWFALL IN BUFFALO 1910-1972
Mean 80.5, Variance 487

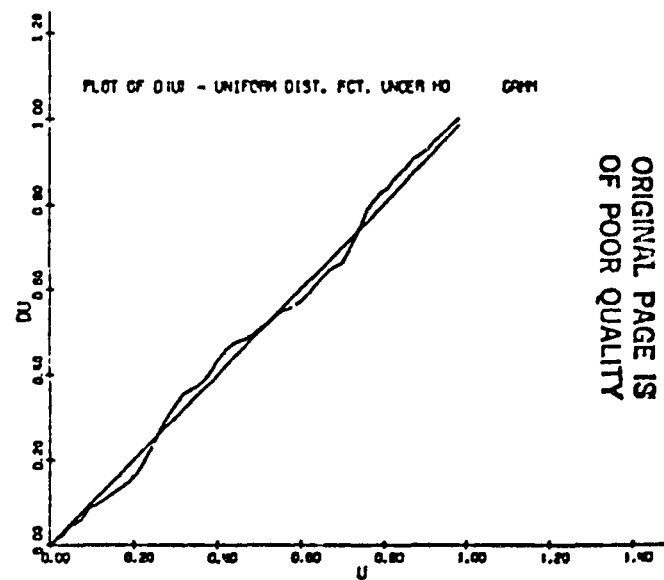
DENSITY ESTIMATES EVALUATED ON SAMPLE RANGE



PARSIMONIOUS 4-PARAMETER EXPONENTIAL DENSITY IS
GAMMA (9.96, .12) Mean 83, Variance 692

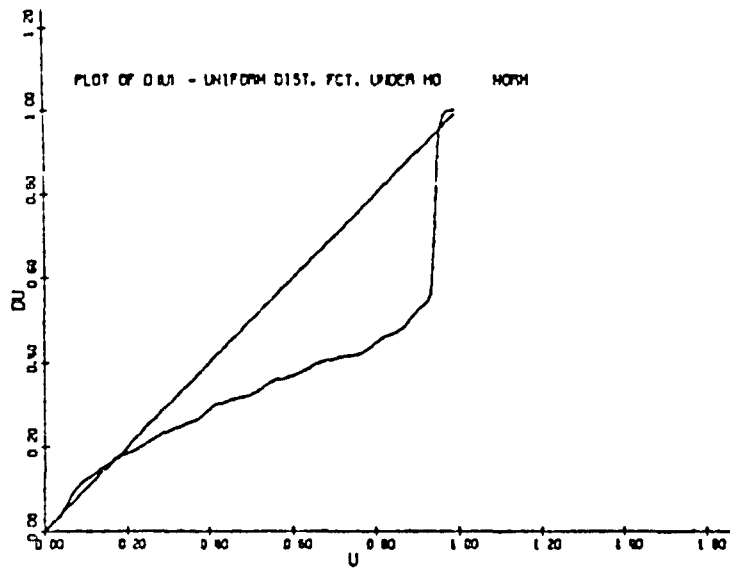


$\tilde{D}(u)$ FOR BI-INF-DIV GAMMA (9.96, .12)



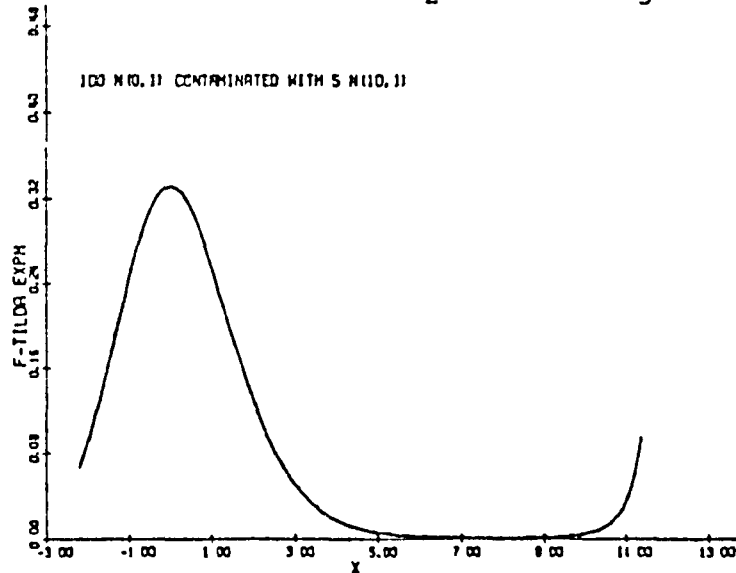
ORIGINAL PAGE IS
OF POOR QUALITY

D(u) FOR BI-INF-DIV NORMAL (-.63, 16.51)

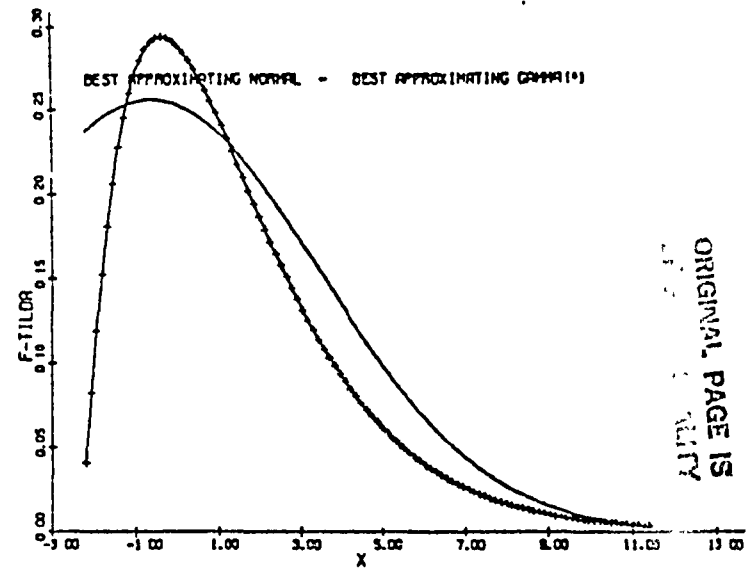


CONTAMINATED NORMAL 100N(0,1), 5N(10,1)
Mean .4, Variance 4.6

PARSIMONIOUS 4-PARAMETER EXPONENTIAL DENSITY HAS
SUFFICIENT STATISTICS x^2 ($\theta_2 = -.28$) x^3 ($\theta_3 = .02$)

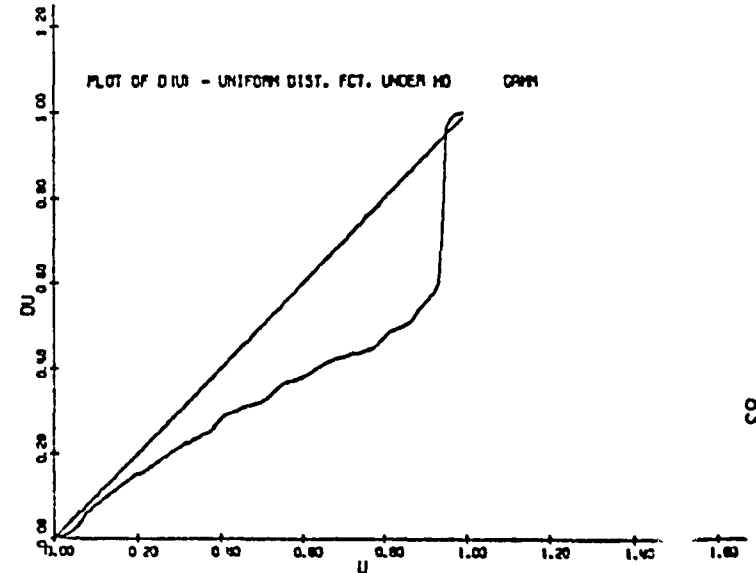


DENSITY ESTIMATES EVALUATED ON SAMPLE RANGE

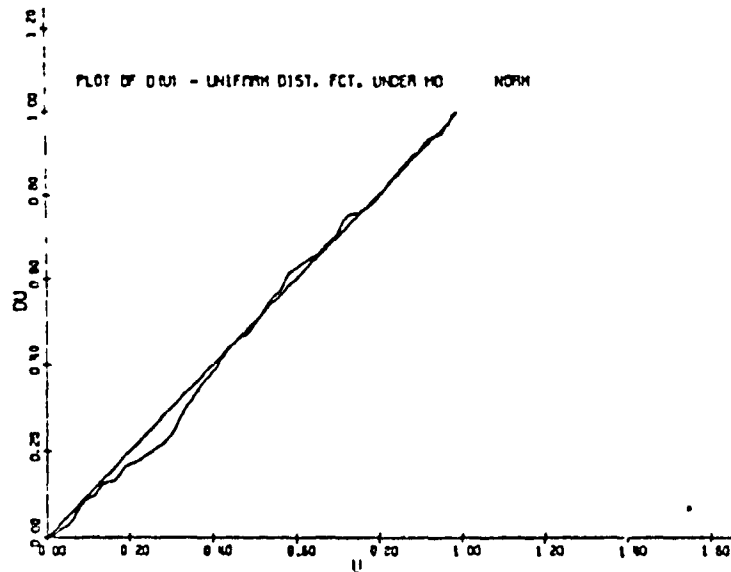


ORIGINAL PAGE IS
OF
QUALITY

D(u) FOR BI-INF-DIV GAMMA (2.11, .57)

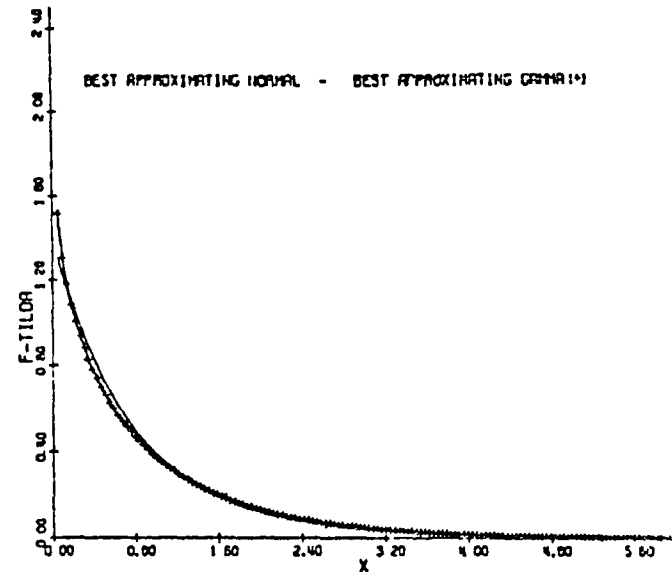


D(u) FOR BI-INF-DIV NORMAL (11.4, -8.6)

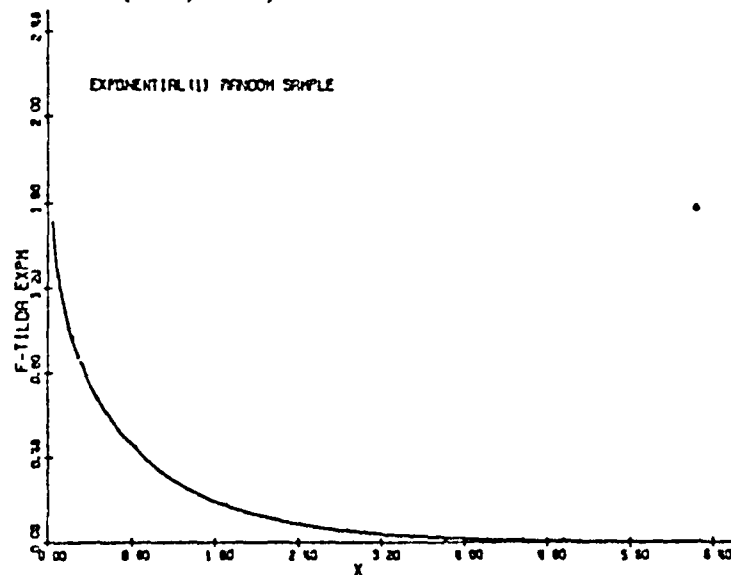


EXPONENTIAL GAMMA ($r=1, \lambda=1$) SIMULATED SAMPLE

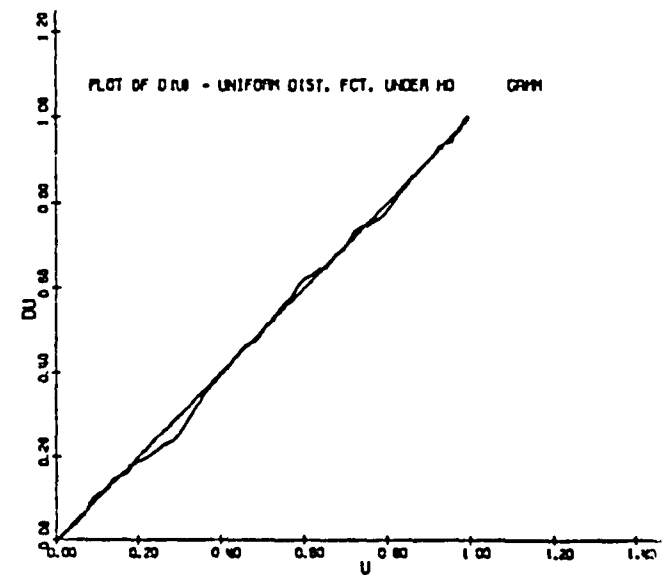
DENSITY ESTIMATES EVALUATED ON SAMPLE RANGE



PARSIMONIOUS 4-PARAMETER EXPONENTIAL DENSITY IS
GAMMA (.83, .93)



D(u) FOR BI-INF-DIV GAMMA (.83, .93)



ORIGINAL PAGE IS
UNCLASSIFIED

N83

15780

UNCLAS

Consistency and other large sample
properties of maximum likelihood
estimates of mixture parameters

by

Charles Peters
Department of Mathematics
University of Houston
Houston, Texas

Abstract

This paper discusses the strong consistency, asymptotic normality, and asymptotic efficiency of maximum likelihood estimates of the parameters in a finite mixture of multivariate distributions, as well as the asymptotic theory of some hypothesis tests for such mixtures.

Research supported in part by NASA/Johnson Space Center under contract
NAS 9 - 14689

1. Introduction

The use of multivariate mixture analysis techniques for unsupervised classification of large amounts of data has been feasible at least since it was proposed and implemented by J.H. Wolfe in 1970, [34]. Prior to that time estimation of parameters in a finite mixture of unknown component distributions had largely been confined to mixtures of a small number of univariate distributions, primarily because of the numerical difficulties and computational requirements of parameter estimation in larger mixture models. A variety of estimators for mixture parameters has been suggested, including moment estimators (Pearson [22] and Rao [26]), graphical methods (Blishke [4] Bhattacharyya [3], Cassie [6] and Harding [15]) and least squares and minimum chi squares estimators. However, recent attention has primarily been focused on maximum likelihood estimation (Day [9], Hasselblad [17], Dick [10], Peters and Coberly [24], and Peters and Walker [25].) and on nonparametric methods (Murray and Titterton [21], and Hall [14]).

As shown in the next section the likelihood equations for mixture parameters are not explicitly solvable and require the use of iterative methods of solution. Because there may be multiple roots of the likelihood equations, one is concerned that the iterative method chosen converge to the "right" solution, i.e., a consistent solution if one exists. This issue is discussed by Kale [20], and also by Peters and Walker [25]. For mixtures with a known number of components, the asymptotic theory is established rather easily using appropriate generalizations of the combined

results of Cramer and Huzurbazar in the single parameter case [8], [19]. For mixtures with an unknown number of components the problem is more difficult and, in particular, the large sample theory of tests for hypotheses about the number of components has not been worked out satisfactorily. These issues will be discussed in more detail in Section 4

The use of multivariate mixture methods in the analysis of remotely sensed data is, for the most part, as an alternative to clustering. Used in this fashion, the method is superior to most clustering methods in ease of implementation (with certain reservations) and in economy of output - it tells the investigator only the most important facts about the data distribution. Thus, the usefulness of mixture density estimation in this mode depends solely on the reality of some prior classification of the data into subpopulations accurately described by the given parametric family of component distributions. However, there is a growing tendency to use large sample considerations, with samples drawn from a multivariate mixture density, as a standard for judging clustering methods [11]. By this standard then, provided the expected consistency, normality, and efficiency properties hold, the maximum likelihood estimate of mixture parameters is the ideal alternative to clustering.

2. The Basic Likelihood Functions.

Let X be a random n -vector which is distributed according to a finite mixture density of the form

$$(2.1) \quad f(x | \alpha, \theta) = \sum_{i=1}^m \alpha_i f_i(x | \theta_i)$$

where the mixing proportions $\alpha_i > 0$ are unknown parameters satisfying

$$(2.2) \quad \sum_{i=1}^m \alpha_i = 1$$

and the $f_i(x | \theta_i)$ are distinct members of parametric families $\{f_i(x | \theta_i) | \theta_i \in \Theta_i\}$ of density functions. For the remainder of this section, we assume that m , the number of components in the mixture, is known, and that the densities $f_i(x | \theta_i)$ come from exponential families

$$(2.3) \quad f_i(x | \theta_i) = c_i(\theta_i) h_i(x) \exp[q_i(\theta_i) \cdot T_i(x)]$$

where $\theta_i = E_{\theta_i} [T_i(X)]$ is the mean value parametrization and $\theta_i \in \Theta_i$ an open subset of \mathbb{R}^{n_i} , [2]. Our aim is to investigate the consistency of roots of likelihood equations for the parameters (α, θ) , where $\alpha = (\alpha_1, \dots, \alpha_m)$ and $\theta = (\theta_1, \dots, \theta_m) \in \Theta_1 \times \dots \times \Theta_m$, for various types of samples. Mixture densities arise most naturally when it is known that X comes from one of m populations P_1, \dots, P_m and that the density of X given that it comes from P_i is of the form $f_i(x | \theta_i)$. If $\Pi \in \{1, \dots, m\}$ is the associated random variable which designates the population of origin, then $\alpha_i = \text{Prob}[\Pi = i]$. The r.v. Π is usually unobserved.

Independent unlabelled samples: $(X_1, \Pi_1), \dots, (X_n, \Pi_n)$ are independent and identically distributed according to (2.1) and the π_i are unknown. The corresponding log likelihood functions is

$$(2.3) \quad L_1(\alpha, \theta) = \sum_{j=1}^n \log f(X_j | \alpha, \theta)$$

Partially labelled samples: Here we consider two sample types, both introduced by Hosmer [18], and studied in detail by him and Walker [33]; see also Redner [29].

Type 1 - Fixed numbers M_1, \dots, M_m of samples are taken independently of one another from each of the component populations P_1, \dots, P_m . Let $\{X_{ij}\}_{i=1}^{M_j}$ be the sample from P_j . In addition a random sample X_1, \dots, X_n is taken from the mixture (2.1). The log likelihood is

$$(2.4) \quad L_2(\alpha, \theta) = \sum_{i=1}^m \sum_{j=1}^{M_j} \log f_i(X_{ij} | \theta_i) + L_1(\alpha, \theta)$$

Type 2 - After a random sample X_1, \dots, X_{N+M} of size $N + M$ is taken from (2.1), the originating populations of X_{N+1}, \dots, X_{N+M} are determined (with no error) and it is found that M_i of X_{N+1}, \dots, X_{N+M} come from P_i , $i = 1, \dots, m$. The log likelihood is

$$(2.5) \quad L_3(\alpha, \theta) = \log \frac{M!}{M_1! \dots M_m!} \alpha_1^{M_1} \dots \alpha_m^{M_m} + L_2(\alpha, \theta)$$

In this expression $L_2(\alpha, \theta)$ has the same form as in (2.4), although M_1, \dots, M_m are random.

Samples in blocks: For making inferences about the agricultural makeup of ground areas from satellite data, certain procedures have been designed which automatically delineate sets of geographically contiguous measurements which come from the same population (Bryant, [5]). Thus, the data is obtained in blocks $X_j = X_{j1}, \dots, X_{jN_j}$, $j = 1, \dots, p$, where the corresponding Π_{jk} have a common value Π_j . Various kinds of dependence can be assumed within each block leading to different likelihood functions of the form

$$(2.6) \quad L_4(\alpha, \theta) = \sum_{j=1}^p \log \sum_{i=1}^m \alpha_i f_{ij}(X_j | \theta_i),$$

where $f_{ij}(X_j | \theta_i)$ is the joint density of X_{j1}, \dots, X_{jN_j} given that $\Pi_j = i$. In deriving (2.6) it is assumed that the size of the block N_j is independent of Π_j , which may require careful stratification of blocks by size. Finally, we remark that, in applications, samples of each type are frequently degraded by missing components in the data vectors. In this case, a likelihood function like (2.6) is appropriate provided the pattern of missing components is independent of both the population of origin and the full data vector. The X_j in (2.6) become the vectors of observed components. Note that not all of the scalar components of θ_i are necessarily identifiable in the density $f_{ij}(X_j | \theta_i)$.

The simplest model (2.3) well illustrates the complications of maximum likelihood estimation. After introducing the appropriate Lagrange multipliers and setting the derivatives of $L_2(\alpha, \theta)$ equal to zero, the following likelihood equations are obtained (see Hasselblad [17] and Redner [29]).

$$(2.7) \quad \alpha_i = \frac{1}{N} \sum_{j=1}^N \frac{\alpha_i f_{ij}(X_j | \theta_i)}{f(X_j | \alpha, \theta)}$$

$$(2.8) \quad \theta_i = \frac{\sum_{j=1}^N \frac{f_{ij}(X_j | \theta_i)}{f(X_j | \alpha, \theta)} T_{ij}(X_j)}{\sum_{j=1}^N \frac{f_{ij}(X_j | \theta_i)}{f(X_j | \alpha, \theta)}}.$$

In addition to the implicitness of the likelihood equation a further difficulty is that the likelihood function may actually be unbounded. For example, if the $f(x | \theta_i)$ in (2.3) are multivariate normal, one of the means is set equal to a sample value, and the corresponding covariance

matrix tends toward singularity, then L_1 tends to infinity (Duda and Hart [12]). Redner [29] shows that L_1 has a global maximum if a penalty term $-\lambda \sum_{i=1}^m \|\Sigma_i^{-1}\|^\ell$ is added, where $\lambda, \ell > 0$ and $\|\Sigma_i^{-1}\|$ is a norm of the inverse of the i^{th} covariance matrix. For partially labelled samples the likelihood function is bounded provided that each multivariate normal population is adequately represented in the labelled portion of the sample.

3. Asymptotic properties of the mle when m is known.

Let X_1, \dots, X_N be independent random variables with densities $f_j(x_j | \theta^0)$, $\theta^0 \in \Theta$, an open subset of \mathbb{R}^n . When we say that there is a strongly consistent maximum likelihood estimator we mean that given a small enough neighborhood U of the true parameter θ^0 the probability is one that there is an interger N_1 such that for $N \geq N_1$ there is a unique solution $\tilde{\theta}_N$ of $\frac{\partial L_N(\theta)}{\partial \theta}$ in U and that $\tilde{\theta}_N$ locally maximize $L_N(\theta)$, where

$$(3.1) \quad L_N(\theta) = \sum_{j=1}^N \log f_j(x_j | \theta) .$$

The estimator $\tilde{\theta}_N$ is asymptotically normal and efficient if $C_N^{-1/2}(\tilde{\theta}_N - \theta^0)$ converges in distribution to $N(0, I)$, where C_N is the Cramer-Rao lower bound

$$(3.2) \quad C_N^{-1} = -E_{\theta^0} \left[\frac{\partial^2 L_N(\theta)}{\partial \theta^2} \right]_{\theta=\theta^0}$$

Under the regularity conditions to be assumed this is

$$(3.3) \quad C_N^{-1} = \sum_{j=1}^N E_{\theta^0} \left[\frac{\partial \log f_j(x_j | \theta)}{\partial \theta} \frac{\partial \log f_j(x_j | \theta)}{\partial \theta} \right]_{\theta=\theta^0} .$$

We observe that for all the sample types considered in the previous section, there are only a finite number of distinct densities $f_j(x_j | \theta)$ to be considered, whether the data has missing components or not (in sampling by blocks, the block sizes are bounded.) In each instance, a straightforward modification of the following theorem and its proof suffices to establish the required asymptotic properties of the model.

Theorem 1. Let $\{g_j(y_j | \theta) | 1 \leq j \leq p; \theta \in \Theta\}$ be a finite set of parametric families of density functions with the same parameter set Θ , an open subset of \mathbb{R}^n . Let X_1, X_2, \dots be independent random variables with densities $f_1(x_1 | \theta^0), f_2(x_2 | \theta^0)$ where each $f_j(x_j | \theta^0)$ is one of $\{g_j(y_j | \theta^0)\}_{j=1}^p$. Suppose each $g(y | \theta) \in \{g_j(y | \theta)\}_{j=1}^p$ satisfies the condition.

a. there is a neighborhood U of θ^0 such that for all

$$\theta \in U \text{ and almost all } y \quad \left\| \frac{\partial g(y | \theta)}{\partial \theta} \right\| \leq h_1(y),$$

$$\left\| \frac{\partial^2 g(y | \theta)}{\partial \theta^2} \right\| \leq h_2(y) \text{ and } \left\| \frac{\partial^3 \log g(y | \theta)}{\partial \theta^3} \right\| \leq h_3(y),$$

where h_1 and h_2 are integrable and $\int h_3(y)g(y | \theta^0) dy < \infty$.

Suppose that there is a positive number ϵ such that

$$b. \quad \frac{1}{n} C_N^{-1} = \frac{1}{N} \sum_{j=1}^N E_{\theta^0} \left[\frac{\partial \log f_j(X_j | \theta)}{\partial \theta} \frac{\partial \log f_j(X_j | \theta)^T}{\partial \theta} \right]_{\theta=\theta^0}$$

$\geq \epsilon I$ for sufficiently large N .

Then there is a strongly consistent solution $\tilde{\theta}_N$ of the likelihood equation

$$0 = \frac{\partial}{\partial \theta} \sum_{j=1}^N \log f_j(X_j | \theta).$$

Furthermore, $\tilde{\theta}_N$ is asymptotically normal, $\tilde{\theta}_N \sim N(\theta^0, C_N)$, and efficient.

Condition (a) and (b) are very similar to those of Chanda [7], who generalized to the multiparameter case the theorems of Cramer [8] and Huzurbazar [19]. For a proof of theorem 1 see Foutz [13], Peters and Walker [25], and Peters [23].

Returning to the mixture density likelihood functions of the previous section with component densities $f_i(x | \theta_i) = c(\theta_i)h_i(x) \exp[q_i(\theta_i) \cdot T_i(x)]$ assume that each pattern of missing components in X manifests itself in a certain pattern of missing components in $T_i(X)$ (as in the multivariate normal distribution). For a sample of size k , let $\phi(i,j,k)$ denote the relative frequency with which the j^{th} component of $T_i(X)$ is observed.

The next theorem is stated for fully observed data vectors; however, it remains valid for data with missing components provided that for each i and j $\lim_{k \rightarrow \infty} \phi(i,j,k) > 0$ for any sample of size k tending to infinity (see Peters [23] and Redner [29]).

Theorem 2. Suppose the functions $\{\exp[q_i(\theta_i) \cdot T_i(x)]\}_{i=1}^m$ together with the component functions of $\{T_i(x) \exp[q_i(\theta_i) \cdot T_i(x)]\}_{i=1}^m$ are all linearly independent. Then there is a consistent, asymptotically normal and efficient mle of (α^0, θ^0) for $L_1(\alpha, \theta)$ as $N \rightarrow \infty$, for $L_2(\alpha, \theta)$ as $N \rightarrow \infty$ and each $\frac{M_i}{N}$ remains bounded, and for $L_3(\alpha, \theta)$ as $M + N \rightarrow \infty$.

4. Mixtures with an unknown number of classes.

If the number m of classes in the mixture density is among the parameters to be estimated, then the results of the preceding section no longer

OF POOR QUALITY

apply. It is easy to see that the likelihood function $L_1(\alpha, \theta)$ can be made arbitrarily large if m is taken to be the sample size. For partially identified samples leading to $L_3(\alpha, \theta)$ the number of classes is eventually determined as M (the number of identified samples) tends to ∞ ; however, because of the expense of labelling samples, one would like to be able to include m as a parameter in $L_1(\alpha, \theta)$. An approach which has had some success in applications is the quasi-Bayesian approach used by Rassbach [27], which will not be discussed here, although it has some similarities to the use of the Akaike information criterion proposed by Redner and Coberly [30].

Suppose

$$(4.1) \quad f(x | \alpha, m, \psi) = \sum_{i=1}^m \alpha_i f_i(x | \theta_i)$$

is a mixture density family with parameters α , m and $\psi = (\theta_1, \dots, \theta_m)$ satisfying

$$(4.2) \quad \begin{aligned} 1 \leq m \leq \bar{m} \\ \sum_{i=1}^m \alpha_i = 1; \quad \alpha_i \geq 0 \\ \theta_i \in \Theta, \end{aligned}$$

a compact subset of \mathbb{R}^n . Since the parameter space is compact we could consider global maxima of the likelihood, except that unfortunately the parameters are no longer identifiable, even locally. This is a consequence of the particular compact parametrization chosen and not of any inherent non-identifiability of finite mixtures (Teicher [31] and Yakowitz [35]). Redner adapted Wald's consistency theorem (Wald [32]) to show that if F

is the set of all finite mixtures (4.1) satisfying (4.2), then under certain conditions, if X_1, X_2, \dots, X_N are iid from $f^0 \in F_m$, then there is a unique mle $f^N \in F_m$ satisfying

$$\sum_{j=1}^N \log f^N(X_j) = \max_{f \in F} \sum_{j=1}^N \log f(X_j)$$

and with probability 1, $f^N(x) \rightarrow f^0(x)$ for each x as $N \rightarrow \infty$, except perhaps for a null set depending only on f^0 (Redner [28]).

For estimating m , which is frequently of independent interest, it is necessary to further restrict the parameters as follows: Replace conditions (4.2) by

$$(4.3) \quad m = \bar{m}$$

$$\sum_{i=1}^m \alpha_i = 1, \quad \alpha_i \geq \epsilon_1 > 0 \quad \text{or} \quad \alpha_i = 0$$

$$\theta_i \in \Theta, \quad \text{a compact subset of } \mathbb{R}^n$$

$$\beta(\theta_i, \theta_j) \geq \epsilon_2 > 0 \quad \text{for } i \neq j$$

where $\beta: \Theta \times \Theta \rightarrow [0, \infty)$ is a continuous function such that $f(\cdot | \theta) = f(\cdot | \theta')$ if and only if $\beta(\theta, \theta') = 0$. A good example of β is the Bhattacharyya coefficient $\beta(\theta, \theta') = 1 - \int [f(x | \theta)f(x | \theta')]^2 dx$. Assume that θ is identifiable in $f(x | \theta)$.

Theorem 3 (Redner [28]). Let X_1, X_2, \dots be independent samples from a mixture density $f(x | \alpha^0, m, \psi^0)$ of type (4.1) subject to conditions (4.3). Let $N_r(\theta)$ be the closed ball of radius r at θ . Suppose the family $\{f(x | \theta)\}_{\theta \in \Theta}$ satisfies the conditions:

$$(i) \int \log f^*(x, \theta, r) f(x | \theta') dx < \infty \text{ for sufficiently small}$$

$$r = r(\theta, \theta'), \text{ where } f^*(x, \theta, r) = \max\{1, \sup_{\bar{\theta} \in N_r(\theta)} f(x | \bar{\theta})\}$$

$$(ii) \text{ for each } \theta \text{ there is a null set } S_\theta \text{ such that for all } x \notin S_\theta, \lim_{\theta' \rightarrow \theta} f(x | \theta') = f(x | \theta).$$

and

$$(iii) \int |\log f(x | \theta')| f(x | \theta) dx < \infty \text{ for all } \theta, \theta'.$$

Then the global mle $(\tilde{\alpha}_n, \tilde{\psi}_n)$ is a strongly consistent estimator of (α^0, ψ^0) . In particular, with probability one the number of nonzero components of $\hat{\alpha}_n$ is eventually the correct number and for the exponential families $\{f(x | \theta)\}$ discussed in the Section 3, $(\tilde{\alpha}_n, \tilde{\psi}_n)$ is asymptotically normal and efficient.

Wolfe [34] suggested a hypothesis testing approach to determining m , where the null hypothesis is that the mixture has m components against the alternative of $m + 1$ components. Specifically, let X_1, \dots, X_n be a sample from $f(x)$ where

$$H_0: f(x) = \sum_{i=1}^m \alpha_i f_i(x | \theta_i)$$

$$\alpha_i > 0, \sum_{i=1}^m \alpha_i = 1$$

$\theta_1, \dots, \theta_m$ are distinct elements of Θ

and

$$H_1: f(x) = \sum_{i=1}^m \alpha_i f_i(x | \theta_i)$$

$$\alpha_i \geq 0, \sum_{i=1}^{m+1} \alpha_i = 1$$

$\theta_1, \dots, \theta_{m+1}$ are distinct elements of Θ

are the null and alternative hypothesis and Θ is an open subset of \mathbb{R}^n . Let f_m^N and f_{m+1}^N be consistent mle's of f under H_0 and H_1 respectively. Wolfe bases his test on the assumption that under H_0 the likelihood ratio statistic

$$\lambda_N = 2 \sum_{j=1}^N \log f_{m+1}^N(X_j) - 2 \sum_{j=1}^N \log f_m^N(X_j)$$

has an asymptotic χ^2 distribution with d.f. = $n + 1$ as $N \rightarrow \infty$. Unfortunately, this does not always seem to hold (Hartigan, [16]). Hartigan suggests that λ_N is stochastically smaller than χ_{n+1}^2 , which would be a true, since then an upper bound at least for the size of the χ^2 test would be known. Apparently, the m -class model cannot be embedded in the $m + 1$ class model regularly enough so that the classical asymptotic theory is valid.

Finally, Redner and Coberly [30] have suggested using the Akaike information criterion to estimate the number of components in the model, whereby $m, \alpha_1, \dots, \alpha_m \geq 0$ and $\theta_1, \dots, \theta_m \in \Theta$ are chosen to maximize

$$(4.4) \quad \text{AIC}_m = \hat{L}_m - k_m$$

where \hat{L}_m is the maximum log likelihood for the m -class model

$$(4.5) \quad \hat{L}_m = \max_{f \in F_m} \sum_{j=1}^N \log f(X_j) = \sum_{j=1}^N \log f_m^N(X_j)$$

and k_m is the number of free parameters, namely

$$(4.6) \quad k_m = mn + m - 1.$$

If m is the true number of components and f_m^N, f_{m+1}^N are consistent mle's, and if the likelihood ratio statistic has a χ_{n+1}^2 distribution, then $E[AIC_{m+1} - AIC_m] = -\frac{n+1}{2}$ asymptotically. The use of the Akaike criterion then is subject to the same reservations as the use of the χ^2 -test, although there is no question of its utility in providing an adequate and economical description of a given data distribution.

Bibliography

1. H. Akaike, Information theory and an extension of the maximum likelihood principle. Second International Symposium on Information Theory. Budapest, 1973.
2. O. Barndorff-Nielsen, Information and Exponential Families in Statistical Theory, Wiley, New York, 1978.
3. C. T. Bhattacharye, A simple method of resolution of a distribution into Gaussian components, Bionstrics, V. 23 (1967), pp. 115-137.
4. W. R. Blischke, Mixtures of distributions, Classical and Contagious Discrete Distributions, edited by G. P. Patil, (1963), Statistical Publishing Society, Calcutta, pp. 351-373.
5. J. Bryant, On the clustering of multidimensional pictorial data, Pattern Recognition, v. 11 (1979), No. 2, pp. 115-125.
6. R. M. Cassie, Some uses of probability paper in the analysis of size frequency distributions, Austral. J. Marine and Freshwater Res., V. 5 (1954), pp. 513-522.
7. K. C. Chanda, A note on the consistency and maxima of the roots of likelihood equations, Biometrika, V. 41 (1954), pp. 56-61.
8. H. Cramer, Mathematical Methods of Statistics, Princeton University Press, Princeton, 1946.
9. N. E. Day, Estimating the components in a mixture of normal distributions, Biometrika, V. 56 (1969), pp. 463-474.
10. N. P. Dick, Maximum likelihood estimations for mixtures of two normal distributions, Biometrics, V. 29 (1973), pp. 781-790.
11. R. Dubes and A. K. Jain, Validity studies in clustering methodologies, Pattern Recognition, V. 11, (1979), No. 2, pp. 235-254.
12. R. O. Duda and P. E. Hart, Pattern Classification and Scene Analysis, Wiley, New York, 1973.

13. V. Fontz, On the unique consistent solution to the likelihood equations, J. Amer. Statist. Assoc. V. 72 (1977), pp. 147-148.
14. P. Hall, On the nonparametric estimation of mixture proportions, J. Roy. Statist. Soc. B, V. 43 (1981), pp. 147-156.
15. J. P. Harding, The use of probability paper for the graphical analysis of polymodal frequency distributions, J. Marine Biological Assoc., V. 28 (1949), pp. 141-153.
16. J. A. Hartigan, Distribution problems in clustering, Classification and Clustering, edited by J. Van Ryzin, publication No. 37 Mathematics Research Center, University of Wisconsin, Academic Press, New York, 1977.
17. V. A. Hasselblad, Estimation of finite mixtures from the exponential family. J. Amer. Statist. Assoc., 1969, pp. 1459-1471.
18. D. W. Hosmer, A comparison of iterative maximum likelihood estimates of the parameters of a mixture of two normal distributions under three different types of samples, Biometrics, V. 29 (1973), pp. 761-770.
19. V. S. Huzurbazar, The likelihood equation, consistency and the maxima of the likelihood function, Ann. Eugen. Lond V. 14 (1948), pp. 185-200.
20. B. K. Kale, On the solution of the likelihood equations by iteration processes, the multiparametric case, Biometrika, V. 49 (1962), pp. 479-486.
21. G. D. Murray and D. M. Tillerington, Estimation problems with data from a mixture, Appl. Statist., V. 27 (1978), No 3, pp. 325-334.
22. K. Pearson, Contributions to the mathematical theory of evolution, Phil. Trans. Roy. Soc., 185 (1894), pp. 71-110.
23. B. C. Peters, On the existence uniqueness and asymptotic normality of the likelihood equations for nonidentically distributed observations. Report 76, Department of Mathematics, University of Houston, 1980.
24. B. C. Peters and W. A. Coberly, The numerical evaluation of the mle of mixture proportions, Commun. Stat. Part A, Theory and Methods, V. A5 (1976), pp. 1127-1135.

25. D. G. Peters and M. F. Walker, Maximum likelihood estimates of the parameters for a mixture of normal distributions, SIAM J. Appl. Math. B, V. 35 (1978), pp. 362-378.
26. C. R. Rao, Advanced Statistical Methods in Biometric Research (1952), Wiley, New York, pp. 300-304.
27. M. E. Rassbach and R. K. Levington, CLASSY-An adaptive maximum likelihood clustering algorithm, LEC-12145, Ninth Annual Meeting of the Classification Society Clemson University, Clemson, S. C., May 21-23, 1978.
28. R. A. Redner, Note on the consistency of the maximum likelihood estimate for nonidentifiable distributions, Ann. Stat. (to appear).
29. R. Redner, Maximum likelihood estimation for mixture models (to appear as a NASA technical memorandum).
30. R. A. Redner and J. A. Coberly, Estimation of the number of components in a mixture model using AIC, Report, Department of Mathematical Sciences, University of Tulsa, Oct. 1981.
31. H. Teicher, Identifiability of finite mixtures, Ann. Math. Statist., V. 34 (1963), pp. 1265-1269.
32. A. Wald, Note on the consistency of the maximum likelihood estimate, Ann. Math. Statist. V. 20 (1949), pp. 595-601.
33. H. F. Walker, The numerical evaluation of the mle for the parameters for a mixture of normal distributions from a partially identified sample. Report 54, Department of Mathematics, University of Houston, 1976.
34. J. H. Wolfe, Pattern clustering by multivariate mixture analysis, Multiv. Beh. Res., V. 5 (1970), pp. 329-350.
35. S. Yakowitz and J. Spragins, On the identifiability of finite mixtures, Ann. Math. Statist., V. 39 (1968), pp. 209-214.

N83

15781

UNCLAS

Smoothing Splines: Regression, Derivatives
and Deconvolution

John Rice and Murray Rosenblatt

University of California, San Diego
La Jolla, California 92093

Research supported by NSF Grant MCS-79-01800 and ONR Contract N00014-81-K-0003.

AMS 1980 subject classification: 62G99, 62J99, 41A15.

Key words and phrases spline, smoothing spline, regularization, deconvolution.

Summary

The statistical properties of a cubic smoothing spline and its derivative are analyzed. It is shown that unless unnatural boundary conditions hold, the integrated squared bias is dominated by local effects near the boundary. Similar effects are shown to occur in the regularized solution of a translation-kernel integral equation. These results are derived by developing a Fourier representation for a smoothing spline.

1. Introduction and Summary

We consider statistical properties of smoothing splines and related procedures. Given $x_i = f(t_i) + \varepsilon_i$, $i = 1, \dots, n$ where f is an unknown smooth function and the ε_i are random errors, a cubic smoothing spline $g(t; \lambda)$ is the function which minimizes

$$\frac{1}{n} \sum_{i=1}^n [x_i - g(t_i)]^2 + \lambda \int (g''(t))^2 dt. \quad (1.1)$$

Smoothing splines were proposed by Whittaker (1923), Schoenberg (1964), and Reinsch (1967). Some analysis of their statistical properties in the case that f and g are periodic appears in Wahba (1975) and Rice and Rosenblatt (1981). The method of cross validation for choosing the smoothing parameter λ from the data has been discussed in Craven and Wahba (1979).

Smoothing splines may be viewed in a larger context. Given $x_i = (Af)(t_i) + \varepsilon_i$ where A is a linear operator, a "regularized" estimate of f is the function g which minimizes

$$\frac{1}{n} \sum_{i=1}^n [x_i - (Ag)(t_i)]^2 + \lambda \int (g''(t))^2 dt. \quad (1.2)$$

Frequently Af is of the form

$$(Af)(t) = \int k(t, s) f(s) ds. \quad (1.3)$$

Many examples of this type may be found in Tikhonov and Arsenin (1977).

The method of regularization is used to control the instability that would arise if one tried to invert A or A^*A . The regularized solutions have a formal resemblance to ridge-regression estimates; in both cases the variance

of the estimate is reduced at the cost of increasing bias. Although there is a large literature on this topic, there has been relatively little analysis of the statistical properties of the solutions.

In this paper we examine two cases of (1.3), numerical differentiation

$$(Af)(t) = \int_0^t f(u) du \quad (1.4)$$

and deconvolution,

$$(Af)(t) = \int_0^1 w(t-s)f(s) ds . \quad (1.5)$$

We next summarize and discuss our main results. Derivations and some further results are contained in later sections. We first deal with a cubic smoothing spline.

Consider observations

$$x_k = f(k/n) + \epsilon_k, \quad k = 0, 1, \dots, n$$

with f continuously differentiable, $f \in L^2$ and the ϵ_k random variables with

$$E \epsilon_k \equiv 0$$

$$E \epsilon_k \epsilon_j = \delta_{k,j} \sigma^2, \quad \sigma^2 > 0 .$$

We wish to determine a continuously differentiable function $g = g(t; \lambda, n)$ with $g \in L^2$ that minimizes

$$\frac{1}{n} \left\{ \frac{1}{4} (x_0 + x_n - g(0) - g(1))^2 + \sum_{k=1}^{n-1} (x_k - g(k/n))^2 \right\} \quad (1.6)$$

$$+ \lambda \int_0^1 (g^{\sim}(t))^2 dt .$$

Here $\lambda = \lambda(n) > 0$ and the object is to determine $\lambda(n)$ as a function of n so that

$$\int_0^1 E[g(t) - f(t)]^2 dt$$

tends to zero as $n \rightarrow \infty$ at a rapid rate. The term

$$\frac{1}{2} (x_0 + x_n) - \frac{1}{2} (g(0) + g(1))$$

appears in (1.6) because one wishes to allow for the possibility that $f(0) \neq f(1)$ and in that case the Fourier series of $f(t)$ will converge to $\frac{1}{2} (f(0) + f(1))$ at $t = 0, 1$.

Theorem 1. Let $f \in C^2$. If $\lambda^3(n)n^8 \rightarrow \infty$, $\lambda(n) \rightarrow 0$ as $n \rightarrow \infty$ then

$$\int \sigma^2[g(t)] dt \cong \frac{\sigma^2 \lambda^{-1/4}}{n} 3 \cdot 2^{-7/2} .$$

Theorem 2. Let $f \in C^4$. Assume that $\lambda^3(n)n^8 \rightarrow \infty$, $\lambda(n) \rightarrow 0$ as $n \rightarrow \infty$. Then if $f^{(2)}(0)$ or $f^{(2)}(1) \neq 0$

$$\int [E g(t) - f(t)]^2 dt \cong \{(f^{(2)}(0))^2 + (f^{(2)}(1))^2\} \lambda^{5/4} 2^{-3/2}$$

while if $f^{(2)}(0) = f^{(2)}(1) = 0$ but $f^{(3)}(0) \neq 0$ or $f^{(3)}(1) \neq 0$ we have

$$\int [E g(t) - f(t)]^2 dt \cong \{ (f^{(3)}(0))^2 + (f^{(3)}(1))^2 \} \lambda^{7/4} 3 \cdot 2^{-3/2} .$$

A common reason for nonparametric data smoothing is to calculate an estimate of the derivative of a function. Schemes for numerically differentiating noisy data that are closely related to the derivative of a smoothing spline have been proposed in Cullum (1971) and Anderssen and Bloomfield (1974). The properties of the derivative of a smoothing spline follow fairly directly from the properties of the smoothing spline itself.

Theorem 3. If $f \in C^2$ and if $\lambda n^5 \rightarrow \infty$ as $n \rightarrow \infty$ and $\lambda \rightarrow 0$, then

$$\int_0^1 2(g'(t)) dt = \frac{\sigma^2}{n} \lambda^{-3/4} \cdot 2^{-7/2} + o(n^{-1} \lambda^{-3/2}) .$$

Theorem 4. Assume that $f \in C^4$, and that $\lambda n^5 \rightarrow \infty$. Then if $f^{(2)}(0) \neq 0$ or $f^{(2)}(1) \neq 0$

$$\int_0^1 [E g'(t) - f'(t)]^2 dt \cong \{ (f^{(2)}(0))^2 + (f^{(2)}(1))^2 \} \cdot \lambda^{3/4} \cdot 3 \cdot 2^{-3/2} ;$$

If $f^{(2)}(0) = f^{(2)}(1) = 0$, but $f^{(3)}(0)$ or $f^{(3)}(1) \neq 0$ then

$$\int_0^1 [E g'(t) - f'(t)]^2 dt \cong \{ (f^{(3)}(0))^2 + (f^{(3)}(1))^2 \} \lambda^{5/4} \cdot 3 \cdot 2^{-3/2} .$$

Comparing these results to Theorems 1 and 2 we see that the variance and integrated squared bias of the derivative are a factor of $\lambda^{-1/2}$ larger than the variance and integrate square bias of the function itself.

Theorem 2 shows that the integrated squared bias is dominated by contributions from the boundary unless g satisfies the condition $g^{(k)}(0) = g^{(k)}(1) = 0, k = 2, 3$. Lemma 6 of section 3 gives a local approximation to the bias in the case that these conditions are not met. Roughly, the bias decays like $\exp(-2^{-1/2} \lambda^{-1/4} t)$ trigonometrically modulated. In the interior of $[0, 1]$ the squared bias is proportional to λ^2 .

These results are not unexpected. The smoothing spline is a "natural" spline and satisfies the two arbitrary end conditions $f''(0) = f''(1) = 0$. In the context of pure interpolation the use of a natural spline is usually not recommended since the error near the ends is of order h^2 where h is the mesh size whereas other methods can produce an error uniformly of order h^4 , if $f \in C^4$, de Boor (1978), Powell (1981). Similarly, it can be shown that the boundary effect dominates the integrated squared error, Rosenblatt (1976). In the nonstochastic framework methods of estimating the boundary constraints have been proposed in these references and it would appear plausible that a similar approach might work in the stochastic case.

Natural splines in the nonstochastic setting and smoothing splines in the stochastic setting are the optimal solutions of certain minimax problems, Powell (1981) and Speckman (1981). It appears that flexibility is lost by guarding against worst cases.

Smoothing splines have also been proposed in the case of spectral density estimation (see Cogburn and Davis (1974) and Wahba (1980)). Boundary effects similar to those studied here occur in the case of periodic smoothing splines unless the function is smoothly periodic (see Rice and Rosenblatt (1980)).

The aliasing in the case of spectral analysis of discretely sampled data implies that boundary behavior will not be smooth in this context.

In the deconvolution problem we consider observations

$$x_k = F(k/n) + \epsilon_k, \quad k = 0, \dots, n$$

where $F(k/n) = \int_0^1 w(k/n-u)f(u)du$, with $f \in L^2$ and the ϵ_k uncorrelated random variables with mean 0 and variance σ^2 . The regularized approximation to f is the function g that minimizes

$$\begin{aligned} & \frac{1}{4n} (x_0 + x_n - G(0) - G(1))^2 + \frac{1}{n} \sum_{k=1}^{n-1} (x_k - G(k/n))^2 \\ & + \lambda \int_0^1 (g(t))^2 dt. \end{aligned} \tag{1.7}$$

Here $G(k/n) = \int_0^1 w(k/n-u)f(u)du$. The kernel of the integral equation, w , is the periodic extension of a function defined on $[0,1]$, and it is assumed that $w \in L^2$. We assume that the Fourier coefficients w_k of w are nonzero for all k .

The constants that occur in the asymptotic expressions for the components of the integrated mean square error depend on the exact form of w , but the rates of decrease depend only on the rate of decrease of the Fourier coefficients w_k of w . Paralleling Theorems 1 and 2 we have

Theorem 5. Let $f \in C^2$ and suppose that $|w_k|^2 \sim k^{-2B}$, $B > 0$. If $\lambda n^{2B+3} \rightarrow \infty$

OF POOR QUALITY

$$\int \sigma^2[g(t)]dt \sim n^{-1} \lambda^{-(2B+1)/(2B+4)} .$$

Theorem 6. Let $f \in C^4$ and suppose that $|w_k|^2 \sim k^{-2B}$, $B > 0$ and $\lambda n^{2B+3} \rightarrow \infty$ as $n \rightarrow \infty$. Then if $f''(0)$ or $f''(1) \neq 0$

$$\int [E g(t) - f(t)]^2 dt \sim \lambda^{5/(2B+4)} .$$

If $f''(0) = f''(1) = 0$ but $f^{(3)}(0)$ or $f^{(3)}(1) \neq 0$, then

$$\int [E g(t) - f(t)]^2 dt \sim \lambda^{7/(2B+4)} .$$

If $f^{(k)}(0) = f^{(k)}(1)$, $k = 2, 3$ then

$$\int [E g(t) - f(t)]^2 dt \sim \lambda^{8/(2B+4)} .$$

Analytic expressions for the approximate local bias are not available, but the qualitative behavior is similar to that of a smoothing spline.

Note that if w is very smooth, B is large, and the integrated mean square error will tend to zero relatively slowly.

2. Examples

The function $f(t) = \cos(2\pi t) + 4 \cos(\pi t)$ satisfies $f''(0) = -8\pi^2$, $f''(1) = 0$, $f'''(0) = f'''(1) = 0$. Figures 1 and 2 show the exact bias of the smoothing spline estimate of the function and its derivative for 50 equi-spaced sampling points and $\lambda = 10^{-6}$. The effect of $f''(0)$ is clearly evident. The asymptotic analysis (Lemma 6) predicts that the bias,

$$b(t) \cong f''(0)\lambda^{1/2} \exp(-t 2^{-1/2}\lambda^{-1/4}) \\ \cdot [\sin(t 2^{-1/2}\lambda^{-1/4}) - \cos(t 2^{-1/2}\lambda^{-1/4})].$$

From this expression we see that the first zero-crossing of the bias should occur at $t = \pi\lambda^{1/4} 2^{-3/2} = .035$ and that $b'(t)$ should be zero at $t = \pi\lambda^{1/4} 2^{-1/2} = .070$, which is borne out in Figure 1. Figure 2 shows that the bias of the derivative is larger by a factor of about $\lambda^{-1/4}$.

We next consider the deconvolution problem wherein f is convolved with a function w , the graph of which is an isosceles triangle centered at 0 with height 20 and base .4. This is intended to correspond to a situation in which averaged values of f are measured with error. Since the analysis of section 4 requires that w be periodically extended, the triangle is also centered over -1 and 1 . To calculate the bias, (1.7) was discretized assuming 25 equi-spaced observations and the solution was computed at 50 equi-spaced points. Other mesh sizes were tried to insure that the results did not merely reflect the discretization. The calculations were done on a VAX 11/80 in double precision. Figure 3 shows the bias for $\lambda = 10^{-8}$; there is a clear effect near 0 and also an effect near 1. The shapes are qualitatively similar to Figure 1.

Since the assumption that w is periodically extended is clearly somewhat artificial, we also computed the bias for w just corresponding to a triangle centered over 0. The resulting bias is shown in Figure 4. Here the only effect is near 0; the effect near 1 of Figure 3 is apparently due to the periodicity of w .

3. The Smoothing Spline and Its Derivative

In this section we derive Theorems 1-4 and some auxiliary results.

In order to do this we carry out a Fourier analysis of the smoothing spline.

Notice that

$$g_k = \int_0^1 e^{2\pi ikt} g(t) dt = \Delta g^0 a_k - \Delta g^1 b_k + h_k b_k \quad (3.1)$$

for $k \neq 0$ where

$$\Delta g^0 = g(1) - g(0)$$

$$\Delta g^1 = g'(1) - g'(0)$$

$$h_k = \int_0^1 e^{2\pi ikt} g''(t) dt .$$

Let

$$y_j = \begin{cases} \frac{1}{2} (x_0 + x_n) & \text{if } j = 0 \\ x_j & \text{if } j = 1, \dots, n-1 \end{cases}$$

and set

$$\hat{y}_j = \frac{1}{\sqrt{n}} \sum_{j=0}^{n-1} y_j \exp(2\pi ijk/n) .$$

Given a sequence of coefficients p_k we will let $p_k^{(n)}$ denote the corresponding set of aliased coefficients arising in a discrete Fourier analysis

$$p_k^{(n)} = \sum_{s=-\infty}^{\infty} p_{k+sn}, \quad k = 0, 1, \dots, n-1.$$

Also let

$$\tilde{p}_0^{(n)} = p_0^{(n)} - p_0 = \sum_{s \neq 0} p_{sn}$$

and

$$\hat{z}_k = \frac{\hat{y}_k}{\sqrt{n}} - \Delta g^0 a_k^{(n)} + \Delta g^1 b_k^{(n)}, \quad k = 1, \dots, n-1. \quad (3.2)$$

Lemma 1. Let f and $\Delta g^0, \Delta g^1$ be given. Assume that f, g are continuously differentiable with $f, g \in L^2$. Then the function g minimizing (1.1) is determined by the following specification on Fourier coefficients:

$$g_0 = \frac{\hat{y}_0}{\sqrt{n}} - \Delta g^0 a_0^{(n)} + \Delta g^1 b_0^{(n)}, \quad (3.3)$$

$$h_{sn} = 0 \quad \text{for } s \neq 0, \quad (3.4)$$

$$h_{k+sn} = \frac{1}{\lambda + r_k} b_{k+sn} \hat{z}_k \quad (3.5)$$

for $k = 1, \dots, n-1$ and integral s . Here it is understood that

$$r_k = \sum_s (2\pi(k+sn))^4.$$

The Parseval relation implies that (1.1) can be rewritten as

OF POOR QUALITY

$$\left| \frac{y_0}{\sqrt{n}} - \bar{y}_0 - \bar{y}_0^{(n)} \right| + \sum_{k=1}^n \left| \frac{y_k}{\sqrt{n}} - \Delta g^0 a_k^{(n)} + \Delta g^1 b_k^{(n)} - (h_k b_k)^{(n)} \right| \quad (3.6)$$

$$+ \lambda \left[(\Delta g^1)^2 + \sum_{k \neq 0} \sum_s |h_{k+sn}|^2 + \sum_{s \neq 0} |h_{sn}|^2 \right].$$

In minimizing this expression, one can separately minimize the sum of the terms with k fixed for each value of k . Minimizing for $k = 0$ leads one to (3.3) and (3.4). For $k \neq 0$ we have

$$\lambda h_{k+sn} = (\hat{z}_k - (h_k b_k)^{(n)}) b_{k+sn}. \quad (3.7)$$

Multiplying by b_{k+sn} and summing over s leads to

$$(h_k b_k)^{(n)} = \frac{\hat{z}_k r_k}{\lambda + r_k} \quad (3.8)$$

and this together with (3.7) leads to (3.5).

Lemma 2. Insert (3.3), (3.4) and (3.5) in (3.6). Minimizing the resulting expression with respect to $\Delta g^0, \Delta g^1$ leads to

$$\Delta g^0 = \left\{ \sum_{k=1}^{n-1} \frac{\hat{y}_k}{\sqrt{n}} \frac{1}{a_k^{(n)}} / (\lambda + r_k) \right\} \left\{ \sum_{k=1}^{n-1} |a_k^{(n)}|^2 / (\lambda + r_k) \right\}^{-1} \quad (3.9)$$

and

$$\Delta g^1 = - \left\{ \sum_{k=1}^{n-1} \frac{\hat{y}_k}{\sqrt{n}} b_k^{(n)} / (\lambda + r_k) \right\} \left\{ 1 + \sum_{k=1}^{n-1} |b_k^{(n)}|^2 / (\lambda + r_k) \right\}^{-1}. \quad (3.10)$$

If we insert (3.3), (3.4) and (3.5) in the expression (3.7), the result can be written as

$$\lambda |\Delta g^1|^2 + \lambda \sum_{k=1}^{n-1} \frac{|\hat{z}_k|^2}{\lambda + r_k} \quad (3.11)$$

Minimizing this expression with respect to Δg^0 and Δg^1 leads to the following equations

$$\sum_{k=1}^{n-1} \frac{\hat{z}_k a_k^{-(n)}}{\lambda + r_k} = 0 \quad (3.12)$$

$$\Delta g^1 + \sum_{k=1}^{n-1} \frac{\hat{z}_k b_k^{(n)}}{\lambda + r_k} = 0 \quad (3.13)$$

On solving for Δg^0 , Δg^1 the expressions (3.7) and (3.10) are obtained.

Lemma 3. *The function g minimizing (1.1) has Fourier coefficients*

$$g_0 = \frac{\hat{y}_0}{\sqrt{n}} + \Delta g^1 b_0^{(n)} \quad (3.14)$$

$$g_{sn} = \Delta g^0 a_{sn} - \Delta g^1 b_{sn} \quad \text{for } s \neq 0 \quad (3.15)$$

and for $k = 1, \dots, n-1$ and s integral

$$\begin{aligned} g_{k+sn} = & \Delta g^0 \left[a_{k+sn} - \frac{1}{\lambda + r_k} |b_{k+sn}|^2 a_k^{(n)} \right] \\ & - \Delta g^1 \left[b_{k+sn} - \frac{1}{\lambda + r_k} |b_{k+sn}|^2 b_k^{(n)} \right] \\ & + \frac{|b_{k+sn}|^2 \hat{y}_k}{\lambda + r_k \sqrt{n}} \end{aligned} \quad (3.16)$$

with $\Delta g^0, \Delta g^1$ given by (3.9) and (3.10).

The fact that $\bar{a}_0^{(n)} = 0$ and (3.3) holds lead to (3.14). Also (3.4) and (3.5) inserted in (3.1) yield (3.15) and (3.16).

The integrated mean square error of $g(t)$ as a function of $f(t)$ is

$$\begin{aligned} & \int_0^1 E[g(t)-f(t)]^2 dt \\ &= \int_0^1 \text{var}(g(t))dt + \int_0^1 [E g(t)-f(t)]^2 dt . \end{aligned} \quad (3.17)$$

Moreover

$$\int_0^1 \text{var}(g(t))dt = \text{var}(g_0) + 2 \sum_{k=1}^{\infty} \text{var}(g_k) . \quad (3.18)$$

It should be noted that the g_k 's are complex-valued random variables. The covariance of two complex-valued random variables U, V is understood to be

$$\text{cov}(U, V) = E\{(U-EU)(\overline{V-EV})\} .$$

We shall now derive Theorem 1. Notice that Δg^0 and Δg^1 are real even though they are written in complex form. It is clear that

$$\text{cov}(\hat{y}_j, \hat{y}_k) = \left(\delta_{j,k} - \frac{1}{2n} \right) \sigma^2 \quad (3.19)$$

for $j, k = 0, 1, \dots, n-1$. From (3.19) it follows that

$$\sigma^2(\Delta g^0) \cong \frac{c^2}{n} \left\{ \sum_{k=1}^{n-1} |a_k^{(n)}|^2 / (\lambda+r_k)^2 \right\} \left\{ \sum_{k=1}^{n-1} |a_k^{(n)}|^2 / (\lambda+r_k) \right\}^{-2} \quad (3.20)$$

$$\sigma^2(\Delta g^1) \cong \frac{\sigma^2}{n} \left\{ \sum_{k=1}^{n-1} |b_k^{(n)}|^2 / (\lambda+r_k)^2 \right\} \left\{ 1 + \sum_{k=1}^{n-1} |b_k^{(n)}|^2 / (\lambda+r_k) \right\}^{-2} \quad (3.21)$$

Since

$$\sum_{k=1}^{n-1} |a_k^{(n)}|^2 / (\lambda+r_k) \cong \sum_{|k| \leq \frac{n}{2}} |2\pi k|^2 / (\lambda |2\pi k|^4 + 1) \cong \lambda^{-3/4} c_1, \quad (3.22)$$

$$\sum_{k=1}^{n-1} |a_k^{(n)}|^2 / (\lambda+r_k)^2 \cong \lambda^{-7/4} c_2, \quad (3.23)$$

$$\sum_{k=1}^{n-1} |b_k^{(n)}|^2 / (\lambda+r_k) \cong \lambda^{-1/4} c_3, \quad (3.24)$$

$$\sum_{k=1}^{n-1} |b_k^{(n)}|^2 / (\lambda+r_k)^2 \cong \lambda^{-5/4} c_4, \quad (3.25)$$

where

$$c_1 = \int \frac{|2\pi x|^2}{|2\pi x|^4 + 1} dx, \quad c_2 = \int \frac{|2\pi x|^6}{(|2\pi x|^4 + 1)^2} dx,$$

$$c_3 = \int \frac{dx}{|2\pi x|^4 + 1}, \quad c_4 = \int \frac{|2\pi x|^4}{(|2\pi x|^4 + 1)^2} dx$$

if $\lambda(n)n^4 \rightarrow \infty$ as $n \rightarrow \infty$, it can be seen that

$$\sigma^2(\Delta g^0) \cong \frac{\sigma^2}{n} C_2 C_1^{-2} \lambda^{-1/4}, \quad (3.26)$$

$$\sigma^2(\Delta g^1) \cong \frac{\sigma^2}{n} C_4 C_3^{-2} \lambda^{-3/4}, \quad (3.27)$$

if $\lambda(n)n^4 \rightarrow \infty$ as $n \rightarrow \infty$. The term

$$\sum_s \sum_{k=1}^{n-1} \left| a_{k+sn} - \frac{1}{\lambda+r_k} |b_{k+sn}|^2 a_k^{(n)} \right|^2 \quad (3.28)$$

occurs as a coefficient of $\sigma^2(\Delta g^0)$ in contributing to (3.18). However, (3.28) can be approximated by

$$\lambda^2 \sum_k \frac{|2\pi k|^6}{(\lambda|2\pi k|^4+1)^2} \cong \lambda^{1/4} C_2 \quad (3.29)$$

with an error $O\left(\frac{1}{n}\right)$ if $\lambda(n)n^4 \rightarrow \infty$ as $n \rightarrow \infty$. The term

$$\sum_s \sum_{k=1}^{n-1} \left| b_{k+sn} - \frac{1}{\lambda+r_k} |b_{k+sn}|^2 b_k^{(n)} \right|^2 \quad (3.30)$$

arises as a coefficient of $\sigma^2(\Delta g^1)$ in contributing to (3.18). An estimation procedure similar to that used in arriving at (3.29) shows that (3.30) can be approximated by

$$\lambda^2 \sum_k \frac{|2\pi k|^4}{(\lambda|2\pi k|^4+1)^2} \cong \lambda^{3/4} C_4 \quad (3.31)$$

with an error $O\left(\frac{1}{n}\right)$ if $\lambda^3(n)n^8 \rightarrow \infty$ as $n \rightarrow \infty$. The estimates obtained for (3.28) and (3.30) imply that the contribution to (3.18) from the terms involving Δg^0 and Δg^1 in (3.16) is

$$O\left(\frac{1}{n}\right)$$

if $\lambda^3(n)n^8 \rightarrow \infty$ as $n \rightarrow \infty$. Now consider the contribution from the last term on the right of (3.16). We shall see that it makes the major contribution to the integrated variance. The expression

$$\sum_s \sum_{k=1}^{n-1} \frac{|b_{k+sn}|^4}{|\lambda+r_k|^2} \frac{1}{n} \quad (3.32)$$

can be approximated by

$$\sum_{0 < |k| < \frac{n}{2}} \frac{1}{(|2\pi k|^4 + 1)^2} \frac{1}{n} \cong \frac{\lambda^{-1/4}}{n} C_5 \quad (3.33)$$

where

$$C_5 = \int \frac{dx}{(|2\pi x|^4 + 1)^2}$$

with an error $O\left(\frac{1}{n}\right)$ if $\lambda(n)n^3 \rightarrow \infty$ as $n \rightarrow \infty$. Theorem 1 follows from these estimates.

Our next object is to derive Theorem 2 for the integrated squared bias of g as an estimate of f . Notice that for $k \neq 0$ we have

of high quality

$$f_k = \int_0^1 e^{2\pi i k t} f(t) dt = \Delta f^0 a_k - \Delta f^1 b_k + m_k b_k \quad (3.34)$$

with

$$m_k = \int_0^1 e^{2\pi i k t} f^{\sim}(t) dt. \quad (3.35)$$

Using (3.34) it is clear that

$$\begin{aligned} \frac{1}{2} (f(0)+f(1)) &= \sum_{k=-\infty}^{\infty} f_k = \sum_{k=0}^{n-1} f_k^{(n)}, \\ f(j/n) &= \sum_{k=-\infty}^{\infty} f_k \exp(-2\pi i j k/n) \\ &= \sum_{k=0}^{n-1} f_k^{(n)} \exp(-2\pi i j k/n), \quad j = 1, \dots, n-1. \end{aligned} \quad (3.36)$$

This implies that

$$E \hat{y}_j / \sqrt{n} = f_j^{(n)}, \quad j = 0, 1, \dots, n-1. \quad (3.37)$$

From (3.34) it follows that

$$f_j^{(n)} = \Delta f^0 a_j^{(n)} - \Delta f^1 b_j^{(n)} + (m_j b_j)^{(n)}, \quad j = 1, \dots, n-1.$$

Relations (3.9), (3.10), and (3.37) imply that

$$E \Delta g^0 = \Delta f^0 + \left\{ \sum_{k=1}^{n-1} (m_k b_k)^{(n)} \bar{a}_k^{(n)} / (\lambda + r_k) \right\} \left\{ \sum_{k=1}^{n-1} |a_k^{(n)}|^2 / (\lambda + r_k) \right\}^{-1} \quad (3.38)$$

and

$$E \Delta g^1 = \Delta f^1 \left[1 - \left\{ 1 + \sum_{k=1}^{n-1} |b_k^{(n)}|^2 / (\lambda + r_k) \right\}^{-1} \right] - \left\{ \sum_{k=1}^{n-1} (m_k b_k)^{(n)} b_k^{(n)} / (\lambda + r_k) \right\} \left\{ 1 + \sum_{k=1}^{n-1} |b_k^{(n)}|^2 / (\lambda + r_k) \right\}^{-1}. \quad (3.39)$$

Since we are dealing with real-valued functions f it follows that

$$m_k = \bar{m}_{-k}$$

and

$$(m_k b_k)^{(n)} = \overline{(m_{-k} b_{-k})^{(n)}}.$$

These last two relations together with (3.38) and (3.39) imply that

$$E \Delta g^0 - \Delta f^0 \cong - \left\{ \sum_{k=-\infty}^{\infty} \frac{(2\pi k) \operatorname{Im} m_k}{1 + \lambda |2\pi k|^4} \right\} \left\{ \sum_{k=-\infty}^{\infty} \frac{(2\pi k)^2}{1 + \lambda |2\pi k|^4} \right\}^{-1} \quad (3.40)$$

and

$$E \Delta g^{1-\Delta} f^1 \cong - \left\{ \sum_{k=-\infty}^{\infty} \frac{\operatorname{Re} m_k}{1+\lambda|2\pi k|^4} \right\} \left\{ \sum_{k=-\infty}^{\infty} \frac{1}{1+\lambda|2\pi k|^4} \right\}^{-1}. \quad (3.41)$$

If $f \in C^3$ one can see that

$$\begin{aligned} \operatorname{Re} m_k &= \int_0^1 f''(x) \cos 2\pi kx \, dx \\ &= \int_0^1 \frac{1}{2} \{f''(x) + f''(-x)\} \cos 2\pi kx \, dx \end{aligned} \quad (3.42)$$

and

$$\begin{aligned} 2\pi k \operatorname{Im} m_k &= 2\pi k \int_0^1 f''(x) \sin 2\pi kx \, dx \\ &= -\Delta f^2 + \int_0^1 f^{(3)}(x) \cos 2\pi kx \, dx \\ &= -\Delta f^2 + \int_0^1 \frac{1}{2} \{f^{(3)}(x) + f^{(3)}(-x)\} \cos 2\pi kx \, dx \end{aligned} \quad (3.43)$$

with

$$\Delta f^2 = f^{(2)}(1) - f^{(2)}(0).$$

From (3.16) it follows that for $k = 1, \dots, n-1$

$$\begin{aligned}
 E g_{k+sn} - f_{k+sn} &= (E \Delta g^0 - \Delta f^0) \left\{ a_{k+sn} - \frac{|b_{k+sn}|^2}{\lambda + r_k} a_k^{(n)} \right\} \\
 &\quad - (E \Delta g^1 - \Delta f^1) \left\{ b_{k+sn} - \frac{|b_{k+sn}|^2}{\lambda + r_k} b_k^{(n)} \right\} \\
 &\quad + (m_k b_k)^{(n)} \frac{|b_{k+sn}|^2}{\lambda + r_k} - m_{k+sn} b_{k+sn} .
 \end{aligned} \tag{3.44}$$

Further, if $f \in C^4$ we have

$$m_k = \Delta f^2 a_k - \Delta f^3 b_k + f_k^{(4)} b_k \tag{3.45}$$

with

$$\Delta f^2 = f^{(2)}(1) - f^{(2)}(0) ,$$

$$\Delta f^3 = f^{(3)}(1) - f^{(3)}(0) ,$$

$$f_k^{(4)} = \int_0^1 \exp(2\pi ikt) f^{(4)}(t) dt .$$

The last term on the right of (3.44) can then be rewritten as

$$\begin{aligned}
 &\Delta f^2 \left\{ (a_k b_k)^{(n)} \frac{|b_{k+sn}|^2}{\lambda + r_k} - a_{k+sn} b_{k+sn} \right\} \\
 &\quad - \Delta f^3 \left\{ (b_k^2)^{(n)} \frac{|b_{k+sn}|^2}{\lambda + r_k} - b_{k+sn}^2 \right\} \\
 &\quad + \left\{ (f_k^{(4)} b_k^2)^{(n)} \frac{|b_{k+sn}|^2}{\lambda + r_k} - f_{k+2n}^{(4)} b_{k+sn}^2 \right\} .
 \end{aligned} \tag{3.46}$$

ORIGINAL PAGE IS
OF POOR QUALITY

Let

$$A_0(t) = - \sum_s \sum_{k=1}^{n-1} \left\{ (b_k^{(n)})^2 \frac{|b_{k+sn}|^2}{\lambda+r_k} - b_{k+sn}^2 \right\} \exp(-2\pi i(k+sn)t),$$

$$A_1(t) = - \sum_s \sum_{k=1}^{n-1} \left\{ (a_k b_k^{(n)}) \frac{|b_{k+sn}|^2}{\lambda+r_k} - a_{k+sn} b_{k+sn} \right\} \exp(-2\pi i(k+sn)t),$$

$$A_2(t) = \sum_s \sum_{k=1}^{n-1} \left\{ b_{k+sn} - \frac{|b_{k+sn}|^2}{\lambda+r_k} b_k^{(n)} \right\} \exp(-2\pi i(k+sn)t),$$

$$A_3(t) = \sum_s \sum_{k=1}^{n-1} \left\{ a_{k+sn} - \frac{|b_{k+sn}|^2}{\lambda+r_k} a_k^{(n)} \right\} \exp(-2\pi i(k+sn)t).$$

Set

$$B_j(t) = \lambda \sum_{k \neq 0} \frac{(2\pi i k)^j}{\lambda(2\pi k)^4 + 1} \exp(-2\pi i k t), \quad j = 0, 1, 2, 3.$$

Lemma 4. If $\lambda^3 n^8 \rightarrow \infty$, $\lambda \rightarrow 0$ as $n \rightarrow \infty$ then

$$\int_0^1 |A_j(t) - B_j(t)|^2 dt = o\left(\lambda^{\frac{7-2j}{4}}\right), \quad j = 0, 1, 2, 3. \quad (3.47)$$

Also $\int_0^1 |B_j(t)|^2 dt$ tends to zero at the rate of $\lambda^{\frac{7-2j}{4}}$, $j = 0, 1, 2, 3$.

The estimates required for this lemma parallel those used to obtain (3.29) and (3.31).

We wish to get more convenient representations or estimates of the $B_j(t)$'s. A contour integration shows that

$$C_0(t) = \frac{1}{2\pi} \int \frac{e^{itx}}{1+x^4} dx = \frac{1}{2\sqrt{2}} e^{-|t|2^{-\frac{1}{2}}} (\cos(t2^{-\frac{1}{2}}) + \sin(|t|2^{-\frac{1}{2}})) .$$

Successive differentiation then indicates that

$$C_1(t) = \frac{1}{2\pi} \int \frac{e^{itx} ix}{1+x^4} dx = -\frac{1}{2} e^{-|t|2^{-\frac{1}{2}}} \sin(t2^{-\frac{1}{2}}) ,$$

$$C_2(t) = \frac{1}{2\pi} \int \frac{e^{itx} (ix)^2}{1+x^4} dx = \frac{1}{2\sqrt{2}} e^{-|t|2^{-\frac{1}{2}}} (\sin(|t|2^{-\frac{1}{2}}) - \cos(t2^{-\frac{1}{2}})) ,$$

$$C_3(t) = \frac{1}{2\pi} \int \frac{e^{itx} (ix)^3}{1+x^4} dx = \frac{1}{2} \operatorname{sgn} t e^{-|t|2^{-\frac{1}{2}}} \cos(t2^{-\frac{1}{2}}) .$$

An application of the Poisson summation formula tells us that

$$B_j(t) = \lambda^{\frac{3-j}{4}} \sum_k C_j((k-t)\lambda^{-\frac{1}{2}}) . \quad (3.48)$$

Only the terms in the sum (3.48) corresponding to $k = 0$, $k = 1$ need to be considered since the sum of the remaining terms die off at the rate $e^{-\alpha\lambda^{-\frac{1}{2}}}$ with α a positive constant. Notice that the formulas for the $C_j(t)$ above imply that

$$C_1 = C_3 = \frac{1}{2\sqrt{2}} .$$

Lemma 5. Assume that $f \in C^4$. Then if $\Delta f^2 \neq 0$

$$E \Delta g^0_{-\Delta f^0} \cong -\lambda^{1/2} \Delta f^2 \quad (3.49)$$

while if $\Delta f^2 = 0$

ORIGINAL PAGE IS
OF POOR QUALITY

$$E \Delta g^0 - \Delta f^0 \cong 2\sqrt{2} \lambda^{3/4} \frac{1}{2} \{f^{(3)}(0) + f^{(3)}(1)\}. \quad (3.50)$$

If $f^{(2)}(0) + f^{(2)}(1) \neq 0$ we have

$$E \Delta g^1 - \Delta f^1 \cong -2\sqrt{2} \lambda^{1/4} \frac{1}{2} \{f^{(2)}(0) + f^{(2)}(1)\} \quad (3.51)$$

and if $f^{(2)}(0) + f^{(2)}(1) = 0$

$$E \Delta g^1 - \Delta f^1 \cong \Delta f^3 \lambda^{1/2} \quad (3.52)$$

as $\lambda = \lambda(n) \rightarrow 0$.

The asymptotic relations (3.49) and (3.50) follow from (3.40), (3.43) and (3.45). Formula (3.51) is a consequence of (3.41) and (3.42). If $f^{(2)}(0) + f^{(2)}(1) = 0$, since $\sum \operatorname{Re} m_k = \frac{1}{2} (f^{(2)}(0) + f^{(2)}(1))$ one can see that

$$\sum \frac{\operatorname{Re} m_k}{1 + \lambda(2\pi k)^4} = - \sum \frac{\lambda(2\pi k)^4 \operatorname{Re} m_k}{1 + \lambda(2\pi k)^4}. \quad (3.53)$$

However by (3.45)

$$\operatorname{Re} m_k = \frac{\Delta f^3}{(2\pi k)^2} - \frac{1}{(2\pi k)^2} \int \frac{1}{2} (f^{(4)}(x) + f^{(4)}(-x)) \cos 2\pi kx \, dx \quad (3.54)$$

This implies (3.52).

Lemma 6. Let $f \in C^4$. If $f^{(2)}(0) \neq 0$, $f^{(2)}(1) = 0$ then

$$E g(t)-f(t) = f^{(2)}(0) \lambda^{\frac{1}{2}} e^{-t 2^{-\frac{1}{2}} \lambda^{-\frac{1}{2}}} [\sin(t 2^{-\frac{1}{2}} \lambda^{-\frac{1}{2}}) - \cos(t 2^{-\frac{1}{2}} \lambda^{-\frac{1}{2}})] + e(t), \quad (3.55)$$

$0 < t < 1$, where the error term $e(t)$ is such that

$$\int e(t)^2 dt = o\left(\int [E g(t)-f(t)]^2 dt\right). \quad (3.56)$$

If $f^{(2)}(0) = f^{(2)}(1) = 0$, $f^{(3)}(0) \neq 0$, $f^{(3)}(1) = 0$, we have

$$E g(t)-f(t) = f^{(3)}(0) \lambda^{3/4} \sqrt{2} e^{-t 2^{-\frac{1}{2}} \lambda^{-\frac{1}{2}}} \cos(t 2^{-\frac{1}{2}} \lambda^{-\frac{1}{2}}) + e(t) \quad (3.57)$$

$0 < t < 1$, where the error term again satisfies (3.56). The approximations appropriate for the cases $f^{(2)}(0) = 0$, $f^{(2)}(1) \neq 0$ and $f^{(2)}(0) = f^{(2)}(1) = 0$, $f^{(3)}(0) = 0$, $f^{(3)}(1) \neq 0$ are obtained by replacing t by $1-t$ in the main expressions on the right of (3.55) and (3.57) respectively.

We next consider the variance and bias of the derivative g' of the smoothing spline. Theorems 2 and 3 follow from the previous analysis of g , after noting that the Fourier coefficients of g' are

$$g'_0 = \Delta g^0 \quad (3.58)$$

$$g'_k = a_k \Delta g^1 - a_k h_k, \quad k \neq 0. \quad (3.59)$$

We first consider the integrated squared variance

$$v = \sum \sigma^2(g'_k).$$

From (3.26),

$$\sigma^2(g_0) \cong \frac{\sigma^2}{n} c_2 c_1^{-2} \lambda^{-1/4}.$$

As in (3.16)

$$\begin{aligned} a_{k+sn} \Delta g' - a_{k+sn} h_{k+sn} &= a_{k+sn} \Delta g^1 \left(1 - \frac{1}{\lambda + r_k} b_{k+sn} b_k^{(n)} \right) \\ &+ a_{k+sn} a_k^{(n)} \Delta g^0 - \frac{a_{k+sn} b_{k+sn} \hat{y}_k}{\lambda + r_k \sqrt{n}} \end{aligned} \quad (3.60)$$

Estimates similar to those used in the analysis of the smoothing spline show that the contribution to the variance from the first term is of order $\lambda^{-1/2} n^{-1}$. The second term gives a contribution of order $\lambda^{-1/2} n^{-1}$; the third term dominates, giving a total contribution to V

$$\begin{aligned} &\cong \frac{\sigma^2}{n} \sum \frac{|2\pi i k|^2}{[\lambda(2\pi i k)^4 + 1]^2} \\ &\cong \frac{\sigma^2}{n} \lambda^{-3/4} \int_{-\infty}^{\infty} \frac{(2\pi x)^2}{((2\pi x)^4 + 1)^2} dx. \end{aligned} \quad (3.61)$$

Next, the bias:

$$\begin{aligned} E g'_{k+sn} - f'_{k+sn} &= a_{k+sn} (E \Delta g^1 - \Delta f^1) \\ &- a_{k+sn} (E h_{k+sn} - m_{k+sn}) \end{aligned} \quad (3.62)$$

which, as in (3.44)

$$\begin{aligned}
 &= (E \Delta g^0 - \Delta f^0) \frac{b_{k+sn} a_{k+sn} a_k^{(n)}}{\lambda + r_k} \\
 &+ (E \Delta g^1 - \Delta f^1) \left[a_{k+sn} - \frac{b_k^{(n)} a_k^{(n)} b_{k+sn}}{\lambda + r_k} \right] \\
 &- a_k \left[\frac{(m_k b_k)^{(n)} b_{k+sn}}{\lambda + r_k} - m_{k+sn} \right]. \tag{3.63}
 \end{aligned}$$

Making approximations as in the analysis of the spline function itself,

$$\begin{aligned}
 E g'(t) - f'(t) &\cong (E \Delta g^0 - \Delta f^0) + (E \Delta g^0 - \Delta f^0) \lambda^{-1} B_0(t) \\
 &+ (E \Delta g^1 - \Delta f^1) B_3(t) + \Delta f^2 B_2(t) - \Delta f^3 B_1(t).
 \end{aligned}$$

Using the Poisson-summation approximation and Lemma 5 if $f^{(2)}(0) \neq 0$, $f^{(2)}(1) = 0$,

$$E g'(t) - f'(t) \cong f^{(2)}(0) 2^{\frac{1}{2}} \lambda^{\frac{1}{2}} e^{-u} \cos u$$

where $u = 2^{-\frac{1}{2}} \lambda^{-\frac{1}{2}} t$. If $f^{(2)}(0) = f^{(2)}(1) = 0$, and $f^{(3)}(0) \neq 0$, $f^{(3)}(1) = 0$

$$E g'(t) - f'(t) \cong -f^{(3)}(0) \lambda^{\frac{1}{2}} e^{-u} (\sin u + \cos u).$$

Note that the approximate (in an L_2 sense) bias of the derivative is the derivative of the approximate bias (Lemma 6).

4. Deconvolution

We now sketch the development of the deconvolution results. Since this parallels closely the derivations of Section 3 the presentation will be somewhat sketchier. As before let g have Fourier coefficients

$$g_k = \Delta g^0 a_k - \Delta g^1 b_k + h_k b_k \quad k \neq 0 \quad (4.1)$$

and let

$$G_k = w_k g_k$$

$$A_k = w_k a_k$$

$$B_k = w_k b_k$$

$$H_k = w_k b_k h_k$$

and define y_j as in Section 3. Then (1.7) may be written as

$$\begin{aligned} & \left| \frac{\hat{y}_0}{\sqrt{n}} - G_0 - \tilde{G}_0^{(n)} \right|^2 + \sum_{k=1}^{n-1} \left| \frac{\hat{y}_j}{\sqrt{n}} - G_j^{(n)} \right|^2 \\ & + \lambda \left[(\Delta g^1)^2 + \sum_{j=1}^{n-1} \sum_s |h_{j+sn}|^2 \right]. \end{aligned} \quad (4.2)$$

Minimizing the 0th term gives $h_{sn} = 0$, $s \neq 0$, and $G_0 + \tilde{G}_0^{(n)} = \hat{y}_0/\sqrt{n}$.

As in the analysis of Section 3, we first fix Δg^0 and Δg^1 and minimize with respect to the h_j 's. If

$$\hat{z}_j = \frac{\hat{y}_j}{\sqrt{n}} - \Delta g^0 A_j^{(n)} + \Delta g^1 B_j^{(n)} .$$

Then (4.2) becomes

$$\sum | \hat{z}_j - H_j^{(n)} |^2 + \lambda \left[(\Delta g^1)^2 + \sum_k \sum_s | H_{k+sn} |^2 | B_{k+sn} |^{-2} \right] . \quad (4.3)$$

The minimizing coefficients can be calculated to be

$$h_{j+sn} = \bar{B}_{j+sn} \frac{\hat{z}_j}{\lambda + p_j} \quad (4.4)$$

where $p_j = \sum_{s=-\infty}^{\infty} | B_{j+sn} |^2$. Now to calculate the minimizing Δg^0 and Δg^1 , this solution is substituted back into (4.3) to give

$$\lambda \sum | \hat{z}_j |^2 (\lambda + p_j)^{-1} + \lambda (\Delta g^1)^2 . \quad (4.5)$$

Minimizing this with respect to Δg^0 and Δg^1 amounts to solving two linear equations in two unknowns, and it may be seen that the solution is approximately

$$\Delta g^0 \cong \left\{ \text{Re} \sum \frac{\bar{y}_j}{\sqrt{n}} A_j^{(n)} (\lambda + p_j)^{-1} \right\} \left\{ \sum | A_j^{(n)} |^2 (\lambda + p_j)^{-1} \right\}^{-1} , \quad (4.6)$$

$$\Delta g^1 \cong - \left\{ \Delta f^1 + \text{Re} \sum \frac{\bar{y}_j}{\sqrt{n}} B_j^{(n)} (\lambda + p_j)^{-1} \right\} \left\{ 1 + \sum | B_j^{(n)} |^2 (\lambda + p_j)^{-1} \right\}^{-1} . \quad (4.7)$$

We next consider the integrated squared variance, which is the sum of the variance of the Fourier coefficients of g . Now, from above,

$$\begin{aligned}
 g_{j+sn} &= \Delta g^0 \left[a_{j+sn} - \frac{|b_{j+sn}|^2 \bar{w}_{j+sn} A_j^{(n)}}{\lambda + p_j} \right] \\
 &+ \Delta g^1 \left[b_{j+sn} - \frac{|b_{j+sn}|^2 \bar{w}_{j+sn} B_j^{(n)}}{\lambda + p_j} \right] \\
 &+ \frac{|b_{j+sn}|^2 \bar{w}_{j+sn}}{\lambda + p_j} \frac{\hat{y}_j}{\sqrt{n}}.
 \end{aligned} \tag{4.8}$$

Via approximations similar to those in Section 3, it may be seen that the first two terms contribute a net variance of order $n^{-1} \lambda^{-2\beta/(2\beta+4)}$ whereas the third term contributes the dominating variance, which is of order $n^{-1} \lambda^{-(2\beta+1)/(2\beta+4)}$.

If we write the Fourier coefficients of f as

$$f_k = \Delta f^0 a_k - \Delta f^1 b_k + m_k \tag{4.9}$$

and take expectations in (4.8), the bias of the $(j+sn)^{\text{th}}$ Fourier coefficient may be expressed as

$$\begin{aligned}
 E g_{k+sn} - f_{k+sn} &= (E \Delta g^0 - \Delta f^0) \left[a_{k+sn} - \frac{|b_{k+sn}|^2 \bar{w}_{k+sn} A_k^{(n)}}{\lambda + p_k} \right] \\
 &- (E \Delta g^1 - \Delta f^1) \left[b_{k+sn} - \frac{|b_{k+sn}|^2 \bar{w}_{k+sn} B_k^{(n)}}{\lambda + p_k} \right] \\
 &+ \frac{|b_{k+sn}|^2 \bar{w}_{k+sn} M_k^{(n)}}{\lambda + p_k} - m_{k+sn} b_{k+sn}.
 \end{aligned} \tag{4.10}$$

ON THE THEORY OF
OF POOR QUALITY

As in section 3 for $|k| \leq n/2, k \neq 0$

$$\begin{aligned}
 E g_k^{-f_k} &\cong (E \Delta g^0 - \Delta f^0) \frac{\lambda a_k}{\lambda + |B_k|^2} \\
 &\quad - (E \Delta g^1 - \Delta f^1) \frac{\lambda b_k}{\lambda + |B_k|^2} \\
 &\quad + \Delta f^2 \frac{\lambda a_k b_k}{\lambda + |B_k|^2} \\
 &\quad - \Delta f^3 \frac{\lambda b_k^2}{\lambda + |B_k|^2} \\
 &\quad + \frac{\lambda f_k^{(4)} b_k^2}{\lambda + |B_k|^2}.
 \end{aligned} \tag{4.11}$$

If we let

$$D_j(t) = \lambda \sum_k \frac{k^{4-j}}{\lambda + |B_k|^2} \exp(-2\pi i k t) \quad j = 0, 1, 2, 3 \tag{4.12}$$

(note that $\|D_j\|^2 \sim \lambda^{(7-2j)/(4+2\beta)}$) then

$$\begin{aligned}
 E g(t) - f(t) &\cong (E \Delta g^0 - \Delta f^0) D_3(t) - (E \Delta g^1 - \Delta f^1) D_2(t) \\
 &\quad + \Delta f^2 D_1(t) - \Delta f^3 D_0(t) \\
 &\quad + \lambda \sum_k \frac{f_k^{(4)} b_k^2}{\lambda + |B_k|^2} \exp(-2\pi i k t).
 \end{aligned} \tag{4.13}$$

The function $D_j(t)$ play the role of the functions $B_j(t)$ of section 3. Although their exact analytical forms depend on w , they are, like the B_j 's, successively odd and even, and are increasingly peaked near 0 and 1 as $\lambda \rightarrow 0$.

We now consider the individual terms in (4.13). From (4.6) it follows that

$$E \Delta g^0 - \Delta f^0 = \frac{\sum m_k B_k \bar{A}_k (\lambda + p_k)^{-1}}{\sum |A_k|^2 (\lambda + p_k)^{-1}}.$$

The denominator can be estimated to be $\sim \lambda^{3/(2\beta+4)}$. If $\Delta f^2 \neq 0$, the numerator is

$$\cong \Delta f^2 \sum |B_k|^2 (\lambda + p_k)^{-1} \sim \lambda^{-1/(2\beta+4)}.$$

In combination with D_3 this gives a net contribution to the integrated squared bias which is $\sim \lambda^{5/(2\beta+4)}$. If $\Delta f^2 = 0$ the numerator is $\cong (f^{(3)}(1) + f^{(3)}(0))/2$, giving a net contribution of order $\lambda^{7/(2\beta+4)}$. If $f^{(k)}(0) = f^{(k)}(1) = 0$ $k = 2, 3$ the net contribution is $O(\lambda^2)$.

Next,

$$E \Delta g^1 - \Delta f^1 \cong \frac{\Delta f^1 + \sum m_j |B_j|^2 (\lambda + p_k)^{-1}}{1 + \sum |B_j|^2 (\lambda + p_k)^{-1}}.$$

The denominator is $\sim \lambda^{-1/(2\beta+4)}$ and if $f^{(2)}(1)$ or $f^{(2)}(0) \neq 0$ the numerator is $\cong (f^{(2)}(1) + f^{(2)}(0))/2$. This gives a net contribution to

the integrated squared bias of order $\lambda^{5/(2\beta+4)}$. If both second derivatives are zero the numerator is

$$\cong \lambda \Delta f^3 \sum \frac{b_k}{\lambda + |B_k|^2} \sim \lambda^{1/(2\beta+4)}$$

giving a net contribution of order $\lambda^{7/(2\beta+4)}$. If both second and third derivatives vanish at 1 and 0 the net contribution is $O(\lambda^2)$.

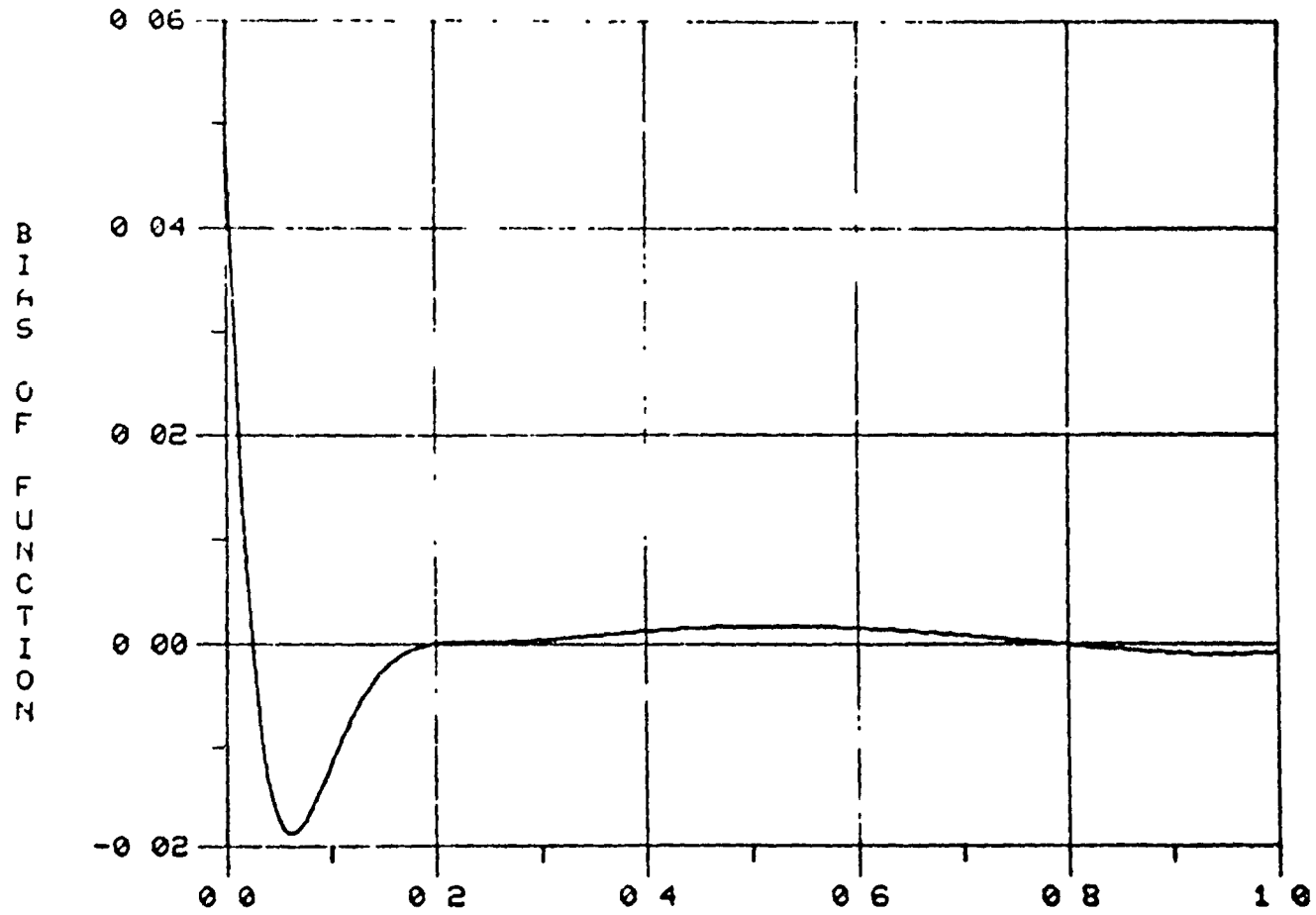
The last term in (4.11) can be estimated and makes a contribution to the integrated squared bias of order λ^2 .

References

- Anderssen, R.S. and Bloomfield, P. (1974). Numerical differentiation procedures for non-exact data. *Numer. Math.* 22, 157-182.
- de Boor, C. (1978). *A Practical Guide to Splines*. Springer-Verlag.
- Cogburn, R. and Davis, H. (1974). Periodic splines and spectral estimation. *Ann. Stat.* 2, 1108-1126.
- Craven, P. and Wahba, G. (1979). Smoothing noisy data with spline functions. *Numer. Math.* 31, 377-403.
- Cullum, J. (1971). Numerical differentiation and regularization. *SIAM J. Num. Anal.* 8, 254-265.
- Oldenburg, D.W. (1981). A comprehensive solution to the linear deconvolution problem. *Geophys. J. R. Astr. Soc.* 65, 331-357.
- Powell, M.J.D. (1981). *Approximation Theory and Methods*. Cambridge University Press.
- Reinsch, C. (1967). Smoothing by spline functions. *Numer. Math.* 24, 383-393.
- Rice, J. and Rosenblatt, M. (1981). Integrated mean square error of a smoothing spline. *J. Approx. Th.*, to appear.
- Rosenblatt, M. (1976). Asymptotics and representation of cubic splines. *J. Approx. Th.* 17, 332-343.
- Schoenberg, I.J. (1964). Spline functions and the problem of graduation. *Proc. Nat. Acad. Sci. U.S.A.* 52, 947-950.
- Speckman, P. (1982). Minimax estimates of linear functionals in a Hilbert space. *Ann. Stat.*, to appear.
- Tikhonov, A.N. and Arsenin, V.Y. (1977). *Solutions of Ill-posed Problems*. V. H. Winston and Sons.
- Wahba, G. (1975). Smoothing noisy data with spline functions. *Num. Math.*, 24, 309-317.
- Wahba, G. (1980). Automatic smoothing of the log periodogram. *J. Amer. Stat. Assoc.* 75, 122-132.
- Whittaker, E. (1923). On a new method of graduation. *Proc. Edinburgh Math. Soc.*, 41, 63-75.

Figure 1

LAMBDA = 0.102E-05



$$\cos(2\pi t) + 4 \cos \pi t$$

Figure 2

LAMBDA = 0.100E-05

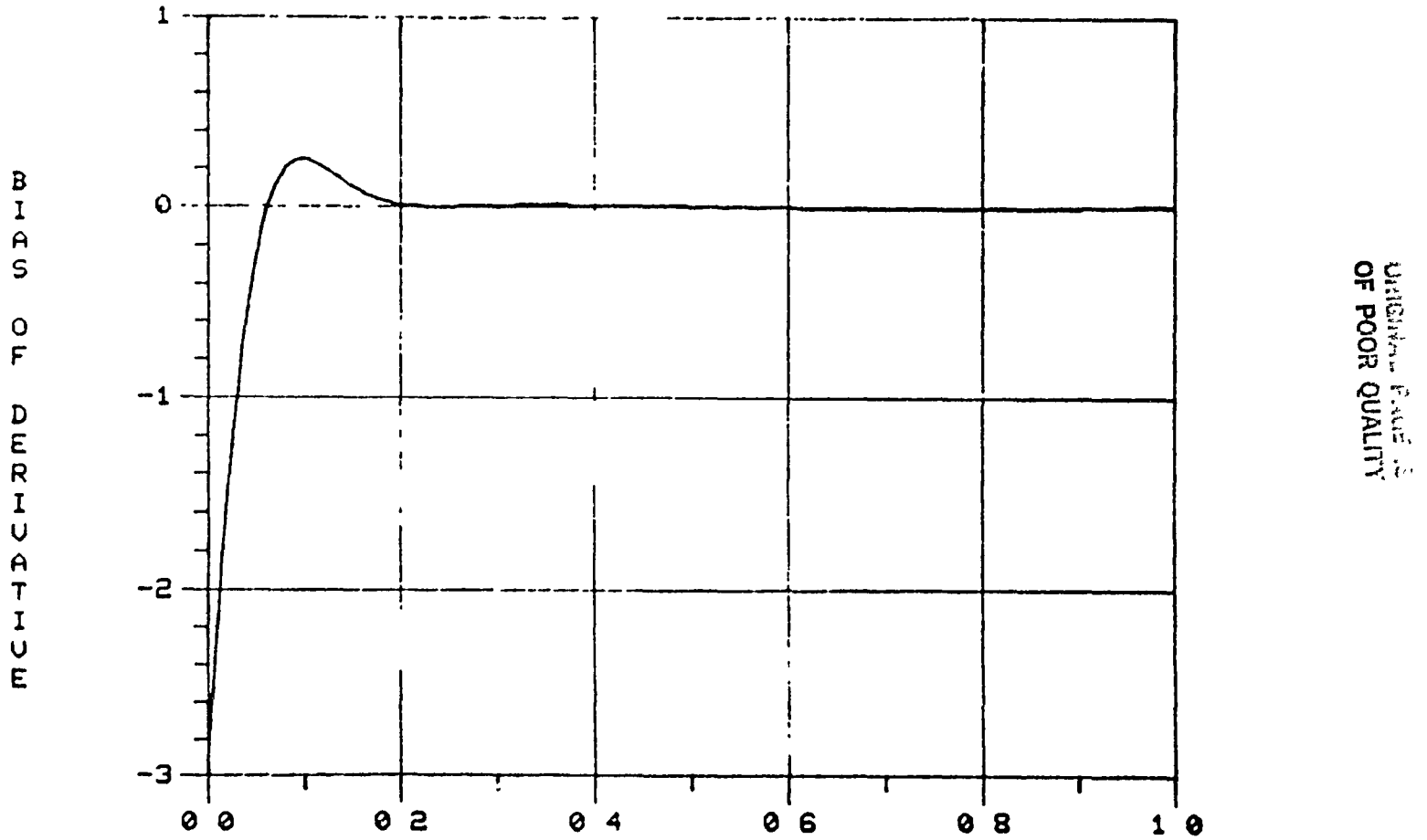
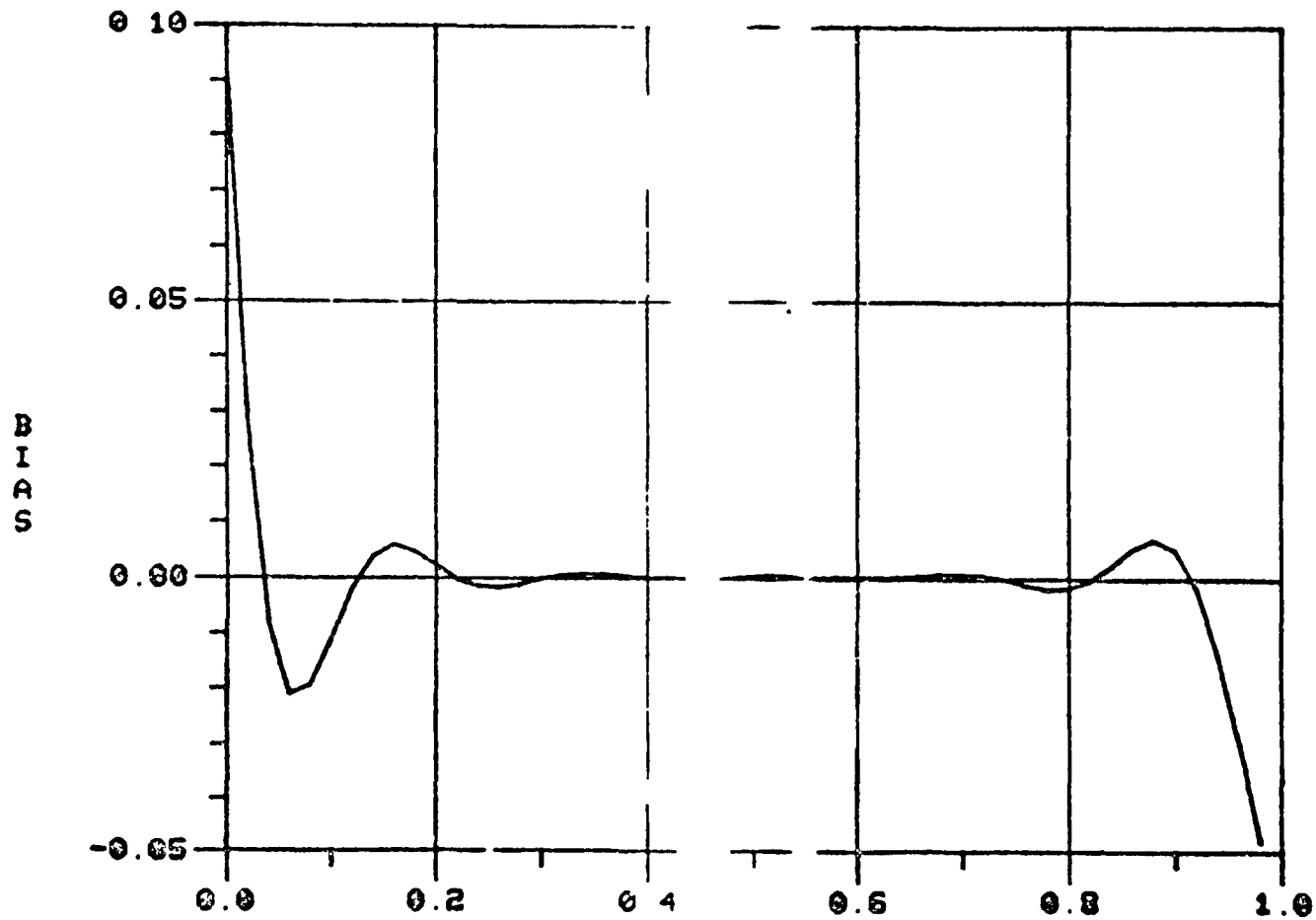


Figure 3

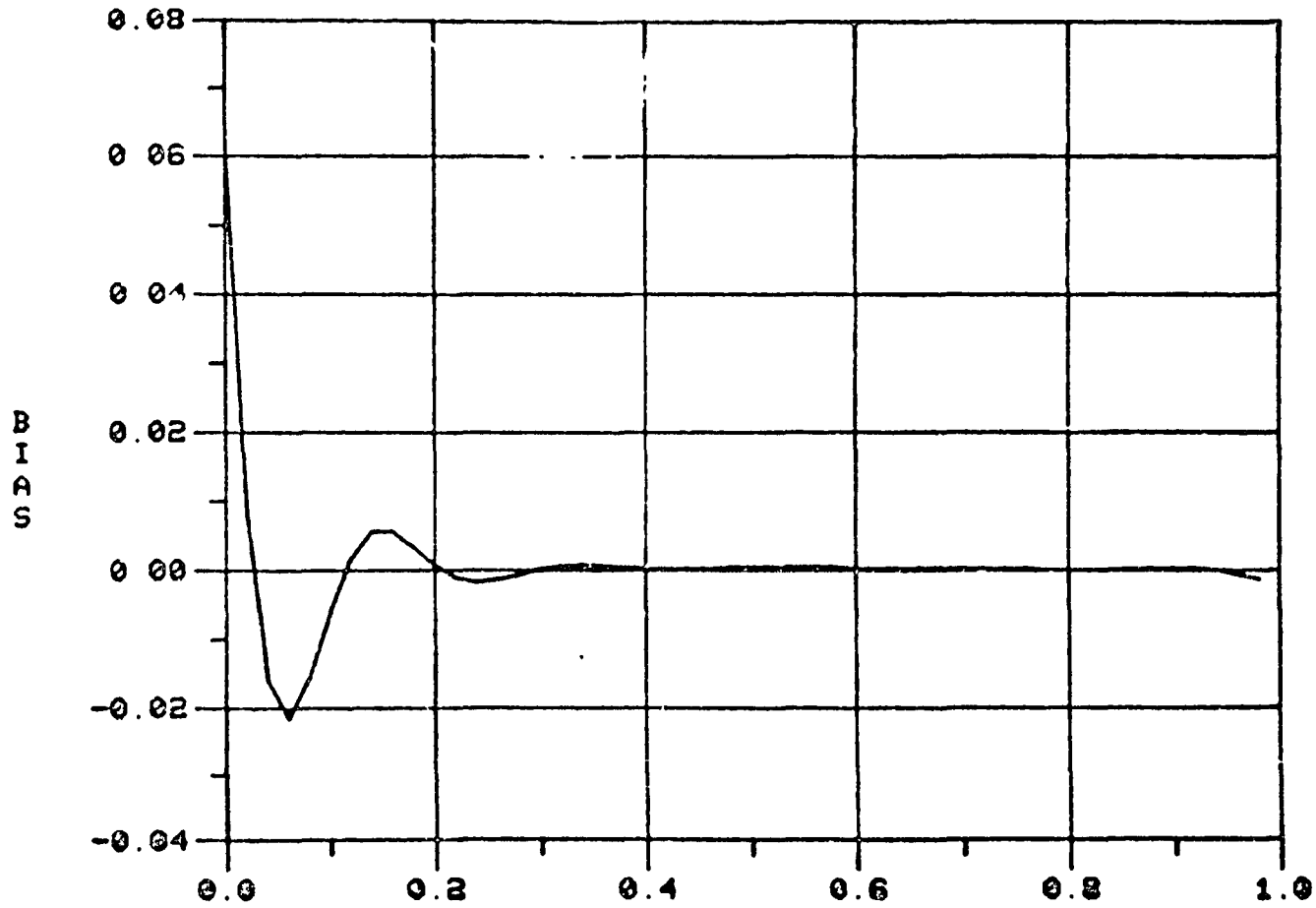
LAMBDA - 3 30E-07



ORIGINAL PROJECT
OF POOR QUALITY

Figure 4

LAMBDA = 0.100E-07



ORIGINAL PAGE IS
OF POOR QUALITY

N83

15782

UNCLAS

Eugene F. Schuster
Department of Mathematical Sciences
The University of Texas at El Paso
El Paso, Texas 79968

Abstract

One criterion proposed in the literature for selecting the smoothing parameter(s) in Rosenblatt-Parzen nonparametric constant kernel estimators of a probability density function is a leave-out-one-at-a-time nonparametric maximum likelihood method. In empirical work with this estimator in the univariate case, we found that it worked quite well for short-tailed distributions. It produced estimators which differed little from those produced by an intuitively appealing maximum likelihood method, depending on a random split of the data, which we had proposed earlier in unpublished work. However, both of these methods drastically oversmoothed for long-tailed distributions. In fact, we have shown that these nonparametric maximum likelihood methods will not select uniformly consistent estimates of the density for long-tailed distributions such as the double exponential or the Cauchy distribution when the kernel has compact support. A remedy we found for estimating long-tailed distributions was to apply the nonparametric maximum likelihood procedures to a variable kernel class of estimators considered by Breiman et al (Technometrics, 19, No. 2, May 1977, 135-143).

In addition to constant and variable kernel estimators we investigated the maximum likelihood criterion applied to a histogram family of estimators and report our experience with some modifications of the above procedures.

Our experience with these estimators includes numerous univariate case studies. This paper reports on the methods as applied to two univariate data sets of one hundred samples (one Cauchy, one normal). Finally, we discuss our limited experience in the multivariate case.

During the past decade there has been much work in the area of non-parametric density estimation. Unfortunately, most of the results have been of the large sample type and little guidance exists as to the practical implementation of the estimators proposed. This is the primary reason why these estimation procedures are used extensively by only a few applied statisticians. One criterion for selecting one out of a family of non-parametric density estimators, which has been mentioned in the literature is the maximum likelihood (ML) criterion. Habbema et al. (1974) and Duin (1976) mention the same ML procedure in the context of Rosenblatt-Parzen kernel type estimation. The form of these estimators is

$$\hat{f}(x) = \hat{f}(x|\theta; x_1, \dots, x_n) = \frac{1}{n} \sum_{i=1}^n K_{\theta}(x-x_i) \quad (1)$$

where x_1, \dots, x_n is a random sample from the density $f(x)$ and K_{θ} is a density with smoothing parameter θ (the quantities x , x_i and θ may be multidimensional).

In the univariate case

$$K_{\theta}(x) = K\left(\frac{x}{\theta}\right) \frac{1}{\theta}, \quad \theta > 0 \quad (2)$$

where $K(\cdot)$ is a fixed density. Choosing θ to maximize the non-parametric likelihood

$$\prod_{i=1}^n \hat{f}(x_i|\theta; x_1, \dots, x_n) \quad (3)$$

is useless; (3) is unbounded as $\theta \rightarrow 0$. To avoid this degeneracy problem, Habbema et al. and Duin consider replacing $\hat{f}(x_i|\theta; x_1, \dots, x_n)$ in (3) by the kernel estimator of $f(x_i)$ based on the data with x_i removed. That is, they chose θ to maximize the criterion

$$\prod_{i=1}^n \hat{f}(x_i|\theta; x_1, \dots, x_{i-1}, x_{i+1}, \dots, x_n) \quad (4)$$

of course this method may be applied to any family of estimators based on some smoothing parameter θ . We designate by ML1 this leave-out-one-at-a-time maximum likelihood procedure as a general method of choosing θ .

Wahba (1978) refers to (4) as a "cross-validation likelihood function." We think that term better describes the new maximum likelihood method which we now propose. Suppose we have a family of estimators $\{\hat{f}(x|\theta; x_1, \dots, x_n)\}$ for certain values of a smoothing parameter θ . Let π_1 be a subset of $\{x_1, \dots, x_n\}$ and $\pi_2 = \{x_1, \dots, x_n\} - \pi_1$. Denote by $\hat{f}(x|\theta_1; \pi_1)$ the estimator determined by the data values in π_1 using the smoothing parameter θ_1 . We use the data in $\pi_1(\pi_2)$ to choose $\theta_2(\theta_1)$ as follows:

$$\theta_1 \text{ is chosen to maximize} \\ \prod_{x \in \pi_j} \hat{f}(x|\theta_1; \pi_1) \quad , \quad i, j \in \{1, 2\} \quad . \quad (5) \\ i \neq j$$

The natural density estimator based on the data split (π_1, π_2) is

$$\hat{f}(x) = \frac{n_1 \hat{f}(x|\theta_1, \pi_1) + n_2 \hat{f}(x|\theta_2, \pi_2)}{n_1 + n_2} \quad (6)$$

where n_1 is the number of elements in π_1 . In Section 3 we consider this estimator for "equal" splits, $n_1 = [n/2]$. A permutation invariant estimator of $f(x)$ is the average of estimators (6) over all equal splits of the data. This estimator is computationally not feasible for moderate n . We suggest averaging (6) over several random splits of the data. In our experience there has been little change in the estimator after averaging over only a small number of random splits (one, two or three). We designate by ML2 this split-sample procedure as a general method for choosing θ . In Section 3 a single likelihood value is utilized as a measure of overall performance. For the ML1 method the single

$$\prod_{i=1}^n \hat{f}(x_i) \quad (7)$$

where \hat{f} is given by (6).

In empirical work with these estimators we found that they both worked quite well and were in close agreement for short tailed distributions. However, both methods drastically oversmoothed for long tailed distributions. In section 2 we discuss the nonconsistency of these methods when using kernels with compact support to estimate densities with tails as long as the double exponential or Cauchy distributions. A remedy which we found for estimating these long tailed distributions was to apply the non-parametric maximum likelihood procedures to a variable kernel class of estimators considered by Breiman et al (1977). This remedy is discussed in section 3 where we analyze two univariate data sets, one from the standard normal and one from the Cauchy.

In section 4 we briefly discuss some of our experience in the multivariate case for the ML2 method. Finally, in section 5 we give some comments and conclusions.

2. Nonconsistency of the ML Procedure.

By nonconsistency we mean that $\sup_x |f_n(x) - f(x)| \not\rightarrow 0$ in probability. This nonconsistency will be demonstrated for a wide class of densities f and kernels k , but we make no attempt to state results for as wide a class as possible. For the sake of argument, attention is placed on the left tail of the distribution and we consider only the ML1 estimators.

Let F denote the cdf of the density f and let $h(u) = u/fF^{-1}(u)$,

$0 < u < 1$. We assume

h is continuous and $\lim_{u \rightarrow 0^+} h(u) = h(0^+)$ exists, possibly infinite. (8)

We say f has a long left tail if $h(0^+) > 0$. Assume for the present only that k has finite support. Without loss of generality we suppose the support is contained in $[-1, 1]$, i.e.

$$k(u) = 0 \text{ if } |u| > 1. \quad (9)$$

A basic observation concerning the smoothing parameter $\theta = \theta_n$ which maximizes (4) is that for each x_i , $|x_i - x_j| \leq \theta_n$ for some x_j with $j \neq i$. In particular for the MLE estimator

$$x_{2n} - x_{1n} \leq \theta_n, \quad (10)$$

where $x_{1n} < x_{2n} < \dots < x_{nn}$ are the order statistics of the sample.

Let $u_{in} = F(x_{in})$, $i=1, \dots, n$. Then $x_{2n} - x_{1n} = F^{-1}(u_{2n}) - F^{-1}(u_{1n}) = h(u_n^*)(u_{2n} - u_{1n})/u_n^*$, where $u_{1n} \leq u_n^* \leq u_{2n}$. From (10) it follows that

$$h(u_n^*)(u_{2n} - u_{1n})/u_{2n} \leq \theta_n. \quad (11)$$

Using uniformity of $(u_{2n} - u_{1n})/u_{2n}$ and standard arguments (11)

$$P(\theta_n < b\epsilon) \leq \epsilon + P(h(u_n^*) < b), \quad b, \epsilon > 0. \quad (12)$$

Lemma 1. Under (8) and (9), $h(0^+) > 0$ implies $\theta_n \xrightarrow{p} 0$. Furthermore $h(0^+) = \infty$ implies $\theta_n \xrightarrow{p} \infty$.

Proof: Choose $0 < b < h(0^+)$ in (12).

Lemma 2. If $\theta_n \xrightarrow{p} \infty$ and $\sup_u |k(u)| < \infty$ then $\sup_x f_n(x) \xrightarrow{p} 0$.

Since k is bounded the proof follows from (1).

Now Lemmas 1 and 2 combine to give the nonconsistency result for distributions like the Cauchy where $h(0^+) = \infty$. There is no difficulty here; the density estimate flattens out entirely. It is more difficult to establish the nonconsistency for boundary cases where $0 < h(0^+) < \infty$. The double exponential density is one of these and is covered by the following lemma. In addition to (9) we will assume that the kernel k is left continuous and of bounded variation on $(-\infty, \infty)$.

Lemma 3. Let θ_n maximize $L_n(\theta)$ of (4) for each n . Suppose f is unimodal and $h(0^+) = a$ where $0 < a \leq \infty$. Then $\sup_x |f_n(x) - f(x)| \not\xrightarrow{p} 0$

in probability.

The proof can be found in Schuster and Gregory (1981).

The following table gives the left tail behavior of some common distributions.

distribution	density	$\lim_{u \rightarrow 0^+} u/fF^{-1}(u)$
Normal	$\frac{1}{\sigma\sqrt{2\pi}} \exp \left\{ -\frac{1}{2} \left(\frac{x-u}{\sigma} \right)^2 \right\}$	0
Double exponential	$\left(\frac{\lambda}{2}\right) \exp \{-\lambda x - u \}$	1/λ
Cauchy	$\left(\frac{\sigma}{\pi}\right) \{\sigma^2 + (x - u)^2\}^{-1}$	∞
Finite support		0

3. Two Univariate Data Sets Analyzed.

Two pseudo-randomly generated data sets, one from a Cauchy distribution and another from a normal distribution, are investigated in this section. General implications are summarized in Section 5.

We first consider the Cauchy example. Table 1 shows $n=100$ order statistics of a pseudo-random sample from a standard Cauchy distribution, $f(x) = \pi^{-1}(1+x^2)^{-1}$. The asterisks (*) indicate a division into two subsamples to be discussed later.

We consider two types of kernel estimators, the constant kernel type - given by equations (1) and (2) where we write $\theta = (\sigma)$, and the variable type -

$$\begin{aligned} \hat{f}(x) &= \hat{f}(x|\theta; x_1, \dots, x_n) \\ &= \frac{1}{n} \sum_{i=1}^n (\alpha d_{ik})^{-1} K\left(\frac{x-x_i}{\alpha d_{ik}}\right) \end{aligned} \quad (13)$$

where $\theta = (k, \alpha)$, $k \in \{1, \dots, n\}$
 $\alpha > 0$

and d_{ik} is the k th nearest neighbor to x_i in the sample $\{x_1, \dots, x_n\}$. For the analysis we chose a kernel K similar to the standard normal but one involving less computing time;

K is the $t(29)$ density.

In our experience the choice of the kernel among those with infinite support, seemed to matter little. However, for long-tailed distributions (such as our present example) kernels with finite support perform poorly.

Consider first the method ML1 for these two types of kernel estimators. We consider the types together as one family and let the maximum of the likelihood (4) choose between them. Notice that for the variable kernel estimators the maximum of (4) is sought over a two-dimensional space $\{(k, \alpha)\}$. The range of α (as well as σ) used for our likelihood calculations is .1(.1)5.5 (ie. from .1 to 5.5 in steps of .1) and that for k is 15(5)45. The constant kernel estimate picked by the ML1 method is useless, being much too flat (oversmoothed). In fact the estimate has a maximum of only .15 and possesses extremely long tails. In the combined family the ML1 method picked the variable kernel estimator with $k=30$ and $\alpha=2.6$. Figure 1 shows this estimator, as well as others, superimposed over a graph of the theoretical Cauchy density. Breiman et al. (1977), page 136, consider three error measures, percent variance not explained (PVNE), mean absolute error (MAE), and mean percent error (MPE), for comparing an estimated density f to a theoretical density f (in this case the Cauchy density which was the model for the pseudo-random samples). The error measures are defined as

$$\begin{aligned} \text{PVNE} &= \frac{1}{\sigma_f^2} \frac{1}{n} \sum_{i=1}^n (f(x_i) - \hat{f}(x_i))^2 \times 100 \\ \text{MAE} &= \frac{1}{n} \sum_{i=1}^n |f(x_i) - \hat{f}(x_i)| \times 100, \text{ and} \\ \text{MPE} &= \frac{1}{n} \sum_{i=1}^n \frac{|f(x_i) - \hat{f}(x_i)|}{f(x_i)} \times 100 \end{aligned} \quad (14)$$

1.0 0.5 0.0 -0.5 -1.0

where $\hat{\mu}_f = \frac{1}{n} \sum_1^n f(x_i)$, and $\hat{\sigma}_f^2 = \frac{1}{n} \sum_1^n (f(x_i) - \hat{\mu}_f)^2$.

In Figure 2 these are plotted as a function of α for the case $k=30$, chosen by the ML1 method. Superimposed on the graph is a plot of transformed likelihood values (4), plotted against α for the case $k=30$. The particular transformation plotted, $(-\log \{\text{expression(4)}\} - 240)/2$, is of no significance; the only intent was to bring the values into the range of the error percentages. It is seen that the ML1 method chooses a value of α (2.6) which is near the minimizing value for each error measure. Note that since all error measures in(14) depend on the unknown density, they could not be used in selecting the smoothing parameters.

Breiman et al. (1977), page 140, used a goodness-of-fit criterion to choose the smoothing parameters of kernel estimators in fitting two bivariate data sets. We investigated this goodness-of-fit criterion for the present univariate data set. The one dimensional value of $V(r)$ in the Breiman paper is $2r$. With $V(r) = 2r$ the goodness-of-fit criterion did not work; over a range of values k , the likelihood values were increasing at $\alpha = .1$ as α decreased. For α this small, the estimators were already too rough. It seemed to us that perhaps in one dimension one should use $V(r)=r$. However, this change gave no better results.

Consider now the method ML2 applied to the constant and variable kernel estimators. The application of the method applied to the constant kernel estimators is straightforward. However, for the variable kernel case, the strict application of the method might lead to an estimator which is a mixture of two with different values of k , which we view as undesirable. We make a distinction between the parameters k and α similar to that made by Wahba (1978) in another setting: α is the primary smoothing parameter and k is a secondary shape parameter. The value of k may be chosen first as follows. (i) Choose at random a partition (π_1, π_2) . (ii) For each of several values of k calculate a value for

the overall criterion (7) where the ML2 method has been applied to the smoothing parameter α only. (iii) Choose the value k which maximizes (7). We repeated the above procedure over three random partitions for the values $k=9(3)21$. In each case the value $k = 15$ was selected. Then with k selected at 15 we chose α based on a new random split. The asterisks in Table 1 indicate the resulting split. Say that a value is in π_1 if it has an asterisk and in π_2 otherwise. Searching over $\alpha = .1(.1)5.5$ the ML2 method chose $\alpha_1 = 2.7$ and $\alpha_2 = 2.5$. The resulting estimator is shown in Figure 1 and is very close to the estimator chosen by the ML1 method. Notice that the value of k chosen by the ML1 method is 30% of the total sample size and that the k chosen by the ML2 method is 30% of the size of each split sample.

We also considered a histogram estimator from Van Ryzin (1973),

$$\hat{f}(x|\theta; x_1, \dots, x_n) = \begin{cases} \frac{\theta}{n(x_{(j+\theta)} - x_{(j)})} & \text{if } x_{(j)} \leq x < x_{(j+\theta)} \\ & j = 1, \theta+1, 2\theta+1, \dots, r \\ \frac{n-r}{n(x_{(n)} - x_{(r)})} & \text{if } x_{(r)} \leq x < x_{(n)} \\ 0 & \text{if } x < x_{(1)} \text{ or } x \geq x_{(r)} \end{cases} \quad (15)$$

where $r = \lfloor \frac{n-1}{\theta} \rfloor + 1$,

with $\lfloor \cdot \rfloor$ the largest integer function, and $x_{(1)} \leq \dots \leq x_{(n)}$ the ordered sample.

Now θ is an integer valued smoothing parameter. We applied the ML2 procedure to this estimator with the following modification. Since at least one of the quantities in (5) would be identically zero due to the finite support of \hat{f} , we modified (5) in this case so that only those x 's for which $\hat{f} \neq 0$ entered into

the product. Averaging (6) over several random splits has a smoothing effect on these estimators. An estimate averaged over five random splits appears in Figure 1. The computation time required to generate the histogram estimate was very small when compared to the kernel estimates.

A similar analysis was carried out on 100 pseudo-random samples from the standard normal density $f(x) = e^{-x^2/2} / \sqrt{2\pi}$. The sample values appear in Table 2 and the density estimates appear in Figure 3. In applying the ML1 method we used the ranges σ and $\alpha = .100(.015).910$ and $k = 5(5)85$. The ML1 method picked the constant kernel estimate with $\sigma = .460$ but only barely so, over the case with $k = 15$. The constant kernel estimate is smooth (see Figure 3) and quite satisfactory while the estimate with $k = 15$ is very rough near the center. This pattern persisted in other examples we investigated; indeed the estimate corresponding to small k was often chosen over the constant kernel estimate. It is seen that the ML1 method, which worked well for long-tailed Cauchy data sets, has instability at small k for the short-tailed normal distribution. The error measures (14) are graphed in Figure 4. The transformation of (4) which is superimposed is $10(-\log \{\text{expression (4)}\} - 135)$. The ML1 method worked very well in picking the smoothing parameter σ of the constant kernel estimator close to the minimizing value for each error measure.

As described previously for the Cauchy data we first choose k for the ML2 method based on values (7) and several random splits. The ranges chosen were $k = 5(5)45$ and σ and $\alpha = .100(.015).910$. Instability was noted here also in the choice of k , different random splits indicating in turn the constant kernel estimator and variable kernel estimators with different k values. The ML2 method seemed to guard against the choice of an extremely small k better than the ML1 method. The use of repeated random splits, which at first glance is a drawback

of the ML2 method, gave the following useful observation. For each random split the constant kernel estimate gave a value for the logarithm of (7) close to the maximum. In fact averaging the logarithms of the likelihood over four random splits showed the constant kernel estimator to be the best. Based on this the constant kernel form was chosen; then three additional random splits were used to give three estimates of the form (6) whose average appears in Figure 3. We mention in passing that the logarithms of (7) used in making the choices among variable and constant kernel forms were often very close together, differing sometimes only in the fifth significant digit. To check for round-off inaccuracies we reprogrammed all calculations in double precision but none of the selections was changed.

The ML2 method was applied to the histogram family (15). The average of 25 estimates of the form (6) appears in Figure 3.

We checked the goodness-of-fit criterion, used by Breiman et al. (1977), on the normal data set. The same results occurred here as reported for the Cauchy data set.

Since the problem with the estimation of long tailed densities was in the oversmoothing caused by the extreme observations, we trimmed observations and considered the natural modified empirical likelihood representing an estimate of the joint density of the order statistics $X_{(L+1)}$ through $X_{(n-L)}$. The smoothing parameter chosen initially decreased drastically with increasing L for long tailed distributions and the estimate of the Cauchy density continued to improve as L increased. However, since the maximizing $\hat{\theta}_n$ was nearly decreasing as L increased we are not able to give any guidance as to how many observations to trim.

OF HOCR QUALITY

4. Multivariate Case.

In the immediate generalization of the univariate maximum likelihood procedure to the multivariate case one would need to choose a shape factor for each coordinate. For simplicity we restrict ourselves to the bivariate case where the bivariate kernel estimator based on the random sample $(x_1, y_1), \dots, (x_n, y_n)$ from a bivariate density is

$$f_n(x, y; a, b) = (nab)^{-1} \sum_{i=1}^n k\{(x-x_i)/a, (y-y_i)/b\}$$

where k is a bivariate density. A common oversimplification in case studies of applications of bivariate (and multivariate) kernel estimators has been to take the same shape factor for each coordinate. Our empirical work has been limited to the split sample ML2 method. We again assume n is even, say $n=2m$. Our bivariate procedure randomly splits the bivariate data into two groups of ordered pairs, say the first and the last m pairs, which we refer to as the x 's and the y 's. The x 's would be used in defining the functional form of the kernel estimator f_m of f and the y 's would be used to find the shape factors (a_1, b_1) which maximizes the "empirical probability" of observing the y 's (as in the univariate case). In the same fashion a second estimator would be constructed which uses the y 's in defining the functional form of f_m and the x 's would be used to find the shape factors (a_2, b_2) which maximize the "empirical probability" of observing the x 's. The resulting two estimators, say $f_{m,1}(\cdot, \cdot; a_1, b_1)$ and $f_{m,2}(\cdot, \cdot; a_2, b_2)$, are then averaged to obtain the bivariate estimator f_n of f .

of solving the bivariate maximization problem. As a good initial guess we used what we refer to as the marginal solutions. Basically, the marginal solution consists of finding the shape parameters which work "best" for each coordinate. To be more specific, if our data pairs were $(x_1, y_1), \dots, (x_n, y_n)$, where $n=2m$, we use (x_1, \dots, x_n) as in section 3 to obtain two estimators $f_{m,1}(x; \hat{a})$ and $f_{m,2}(x; a^*)$ of the common density of x_1, \dots, x_n . Similarly we use (y_1, \dots, y_n) as in section 3 to obtain two estimators $h_{m,1}(y; \hat{b})$ and $h_{m,2}(y; b^*)$ of the common density of y_1, \dots, y_n . The first approximation to (a_1, b_1) was (\hat{a}, \hat{b}) and the first approximation to (a_2, b_2) was (a^*, b^*) . Of course, one could just estimate the bivariate f by averaging $f_{m,1}(\cdot, \cdot, \hat{a}, \hat{b})$ and $f_{m,2}(\cdot, \cdot, a^*, b^*)$. We call this our marginal solution. Although somewhat more irregular in the bivariate normal cases we have studied, this marginal solution is less time-consuming to compute and seems to be adequate for many purposes. In figures 5-10 we picture the actual density, the marginal estimator, and the nonparametric maximum likelihood estimator for two case studies of samples of 400 from bivariate normal densities. In case 1 we are sampling from the bivariate normal density with mean vector 0 and unit covariance pictured in figure 5. Figure 7 gives the marginal estimators for this case and figure 8 gives the nonparametric maximum likelihood estimator. The second case study is a sample of 400 from the bivariate normal density 0 mean, and covariance matrix $A = \begin{bmatrix} 1 & 1/2 \\ 1/2 & 1 \end{bmatrix}$ pictured in figure 6. Figures 9 and 10 give the marginal and nonparametric maximum likelihood estimators for this case. The bivariate kernel used was a product of standard normal densities.

On the Use of Maximum Likelihood Estimation

The results reported here are only a small part of our experience in applying non-parametric maximum likelihood techniques. However, our multivariate experience with simulated data has been limited to the split sample ML2 method. Below is a list of observations and recommendations.

- a. We find the ML2 data-splitting method to be very attractive, both conceptually and in application. There is some instability in the choice of k for variable kernel estimators of short tailed densities but we judge it to be less than for the ML1 estimators. The answer here might be to have a preliminary classification of the density as long or short tailed and then to suitably restrict the k values considered. For short tailed densities only large k values and the constant kernel case would be considered. A drawback of the ML2 method for kernel estimation is the considerable computation time involved. The ML1 method offers only moderate improvement in computation time. We feel that the randomized nature of the ML2 method may prove very useful in future work in guarding against bad estimates both of densities and functionals of densities. An attempt was made to remove this randomized component by dividing the sample into the even and odd statistics. This approach failed; the density estimates were too rough.
- b. The ML1 method has as noted above instability in the estimation of short tailed densities. However, in cases where this was not a problem the ML1 and ML2 methods tended to agree closely and for constant kernel estimation, to coincide almost exactly. We view this as justification of the ML1 procedure which on the surface does not impart

- ... WOULD BE THE END OF THE ROAD.
- c. The histogram estimators are disappointing in their lack of smoothness but within that class of estimators we judge that the maximum likelihood method worked well. Computation time was fast. Perhaps future work will develop a "quick and dirty" way of using a histogram estimate as a preliminary in choosing the smoothing parameters of kernel estimators.
 - d. We have determined that maximum likelihood techniques are also applicable to multidimensional density estimation. In multivariate kernel estimation with a product kernel a marginal distribution technique is to choose the smoothing parameter associated with each variable by considering only the univariate marginal distribution of each variable. This is to be contrasted with a multidimensional search of the likelihood surface. In the multivariate case computation time is very important. In this direction our empirical work was limited to several multivariate normal data sets using the ML2 constant kernel method. Although the estimators were quite reasonable there was some tendency to oversmooth.
 - e. The use of maximum likelihood techniques on families of estimators other than those considered here, should be investigated. In particular this includes the orthogonal series estimators of Wahba (1978).

Acknowledgement. The work reported in this paper is primarily a summary of results in the papers by Schuster and Gregory (1978, 1979, 1981).

Table 1.
Cauchy Data

-23 376*	-17.511*	-11.159	-10 315
-3 726*	-4.828*	-3 836	-3.319*
-2 576*	-2 750*	-2 449	-2.417*
-2 363*	-2.314	-2.114*	-2 058*
-1 772	-1 623*	-1 543*	-1.401
-1 177	-1 236*	-1.178*	-1 083
-1 14*	-1 052*	-1.043	-0 987*
-0 74*	-0 756	-0 525*	-0 480
-0 429*	-0 455	-0 454	-0 434*
-0 270*	-0.421*	-0 328	-0 318
-0 157*	-0 265	-0 212*	-0.165
0 063*	-0 006	0 051*	0 056*
0 157	0 084*	0 116*	0 117
0 266*	0 186*	0 226*	0 253
0 370*	0 284	0 337	0.343
0 474*	0 381*	0 402	0.444
0 578*	0 590	0 607	0.684
0 682*	0.728	0 807*	0.846
0 786*	0 835	0 892	0.926*
1 039	1 067*	1 269	1 306*
1 374*	1.501	1 573*	1 638
1 657	1 925	2 053	2 451*
2 622*	3 019*	3 130	3.772
4 173	4 833*	5 093*	6 609
6 636	7.594	12 152	16 504*

Table 2.
Normal Data

-2.313	-2.232	-1.957	-1.641
-1.595	-1 529	-1.429	-1.424
-1.157	-1.145	-1 129	-1.065
-1.039	-1 018	-0.956	-0.908
-0.905	-0 885	-0 861	-0.851
-0 814	-0.757	-0.692	-0 690
-0.647	-0 631	-0.621	-0.611
-0.598	-0 536	-0.532	-0.502
-0.491	-0 465	-0 447	-0.432
-0 421	-0 420	-0.384	-0.361
-0.297	-0 259	-0.244	-0.197
-0.173	-0 112	-0 090	-0.047
-0 031	-0 029	-0 015	-0.004
-0 001	0 014	0.094	0.116
0.134	0 146	0.163	0.166
0 167	0 178	0.222	0 242
0 263	0.276	0 326	0.366
0 380	0 409	0 416	0.423
0 500	0 594	0 643	0 692
0 695	0 737	0.740	0.773
0.780	0 799	0.805	0 830
0 844	0 880	0 934	0 949
0.963	1 004	1 009	1 084
1 115	1.250	1.303	1.489
1.534	1 601	1.836	2.311

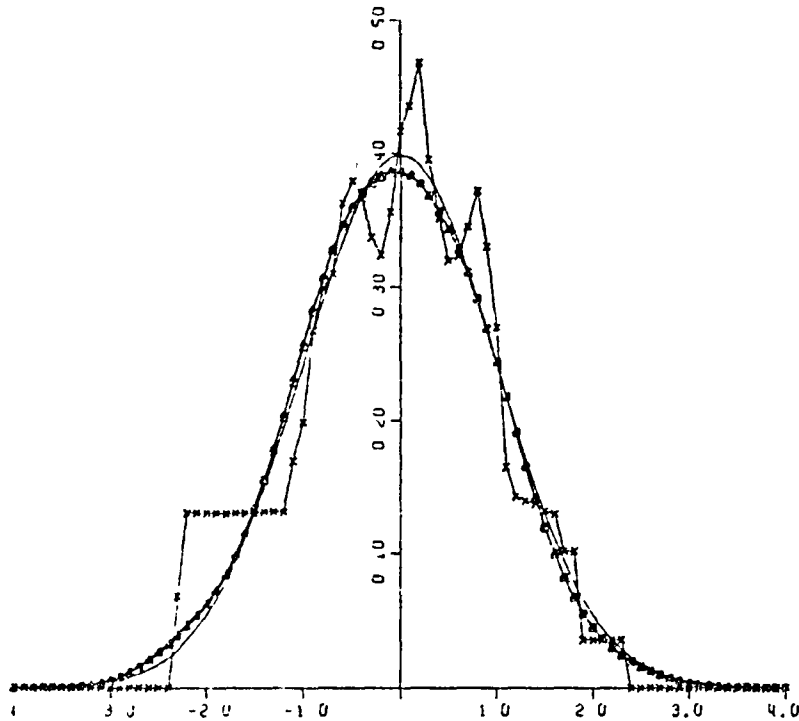


Figure 3. Normal Estimation Δ method ML1 with kernel estimators
 \square method ML2 with kernel estimators
 \times method ML2 with histogram estimators
 no symbol theoretical density

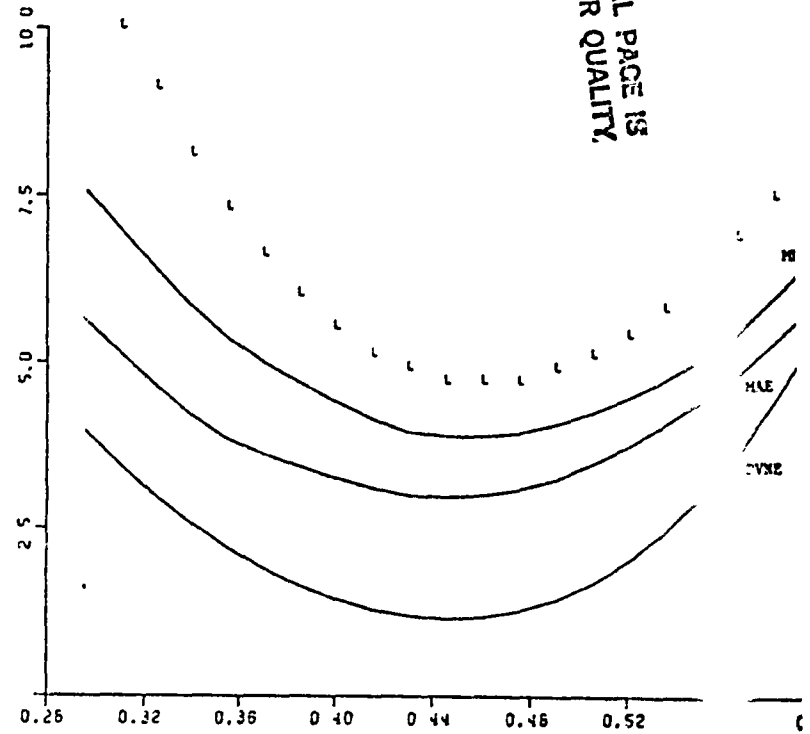


Figure 4. Error Measures for ML1 Estimation. Symbol L's transformed likelihood (see text).

ORIGINAL PAGE IS
 OF POOR QUALITY.

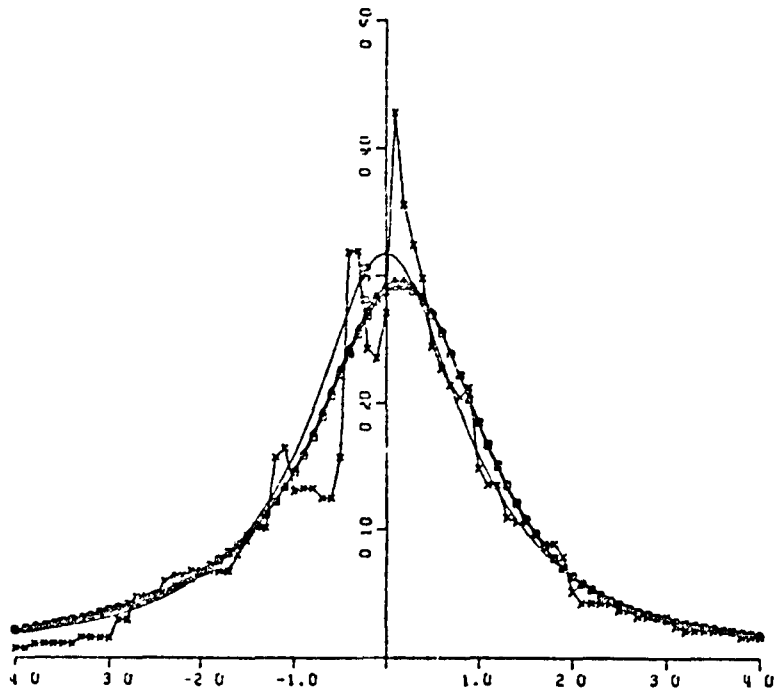


Figure 1. Cauchy Estimation. Δ method M1 with kernel estimators
 \square method M2 with kernel estimators
 \times method M2 with histogram estimators
 no symbol theoretical density

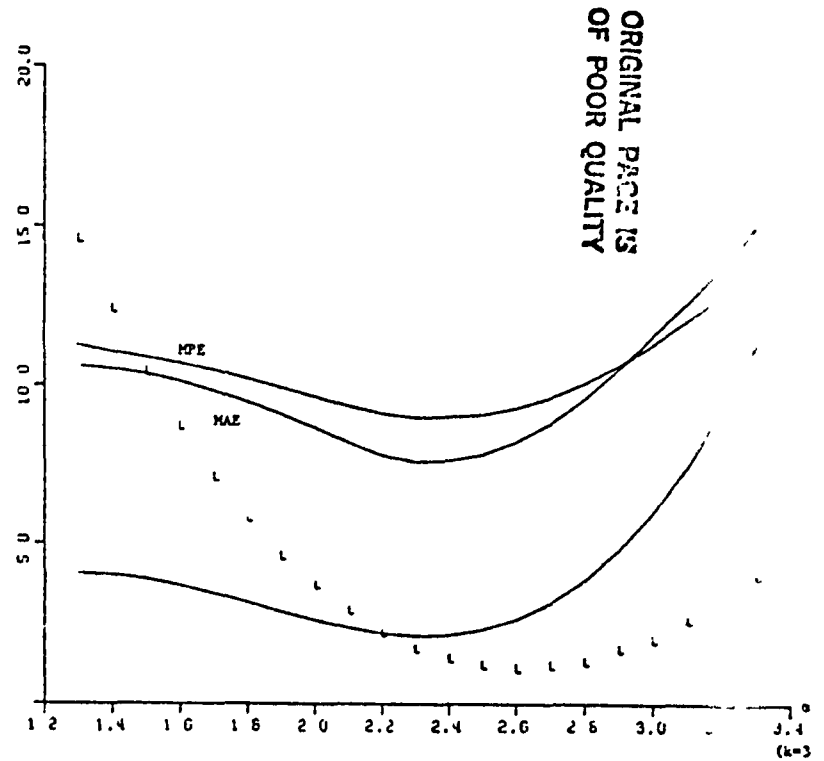


Figure 2. Error Measures for M1 Cauchy Estimation. Symbol L shows transformed likelihood values (see text).

ORIGINAL PAGE IS
 OF POOR QUALITY

Figure 5.
Bivariate normal density with mean $\underline{0}$ and unit covariance matrix.

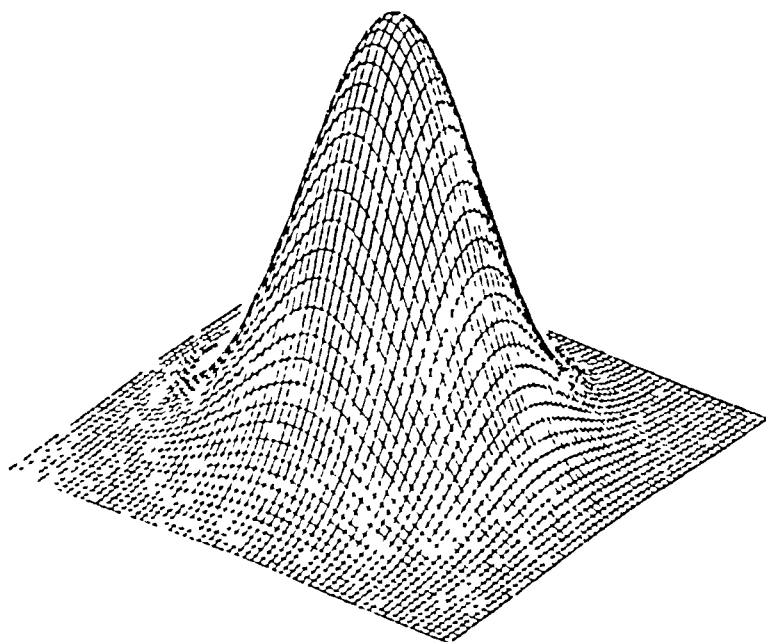


Figure 6.
Bivariate normal density with mean $\underline{2}$ and covariance matrix $A = \begin{bmatrix} 1 & 0 \\ 0 & 2 \end{bmatrix}$.

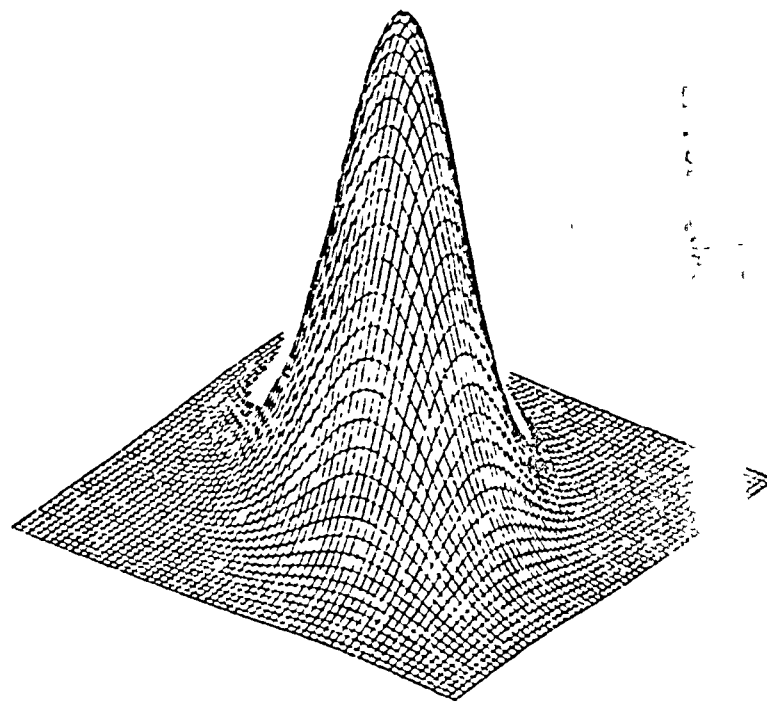


Figure 7.
Marginal estimator of the bivariate normal of figure 5.

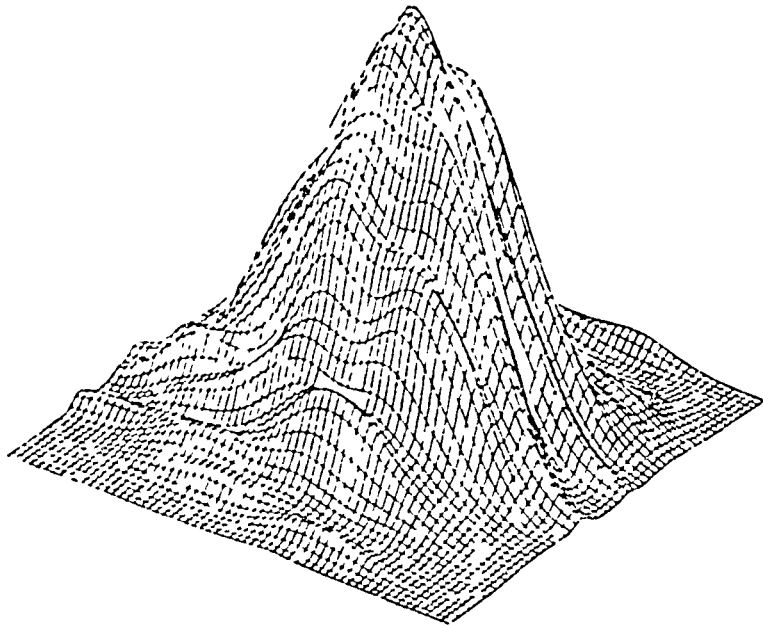
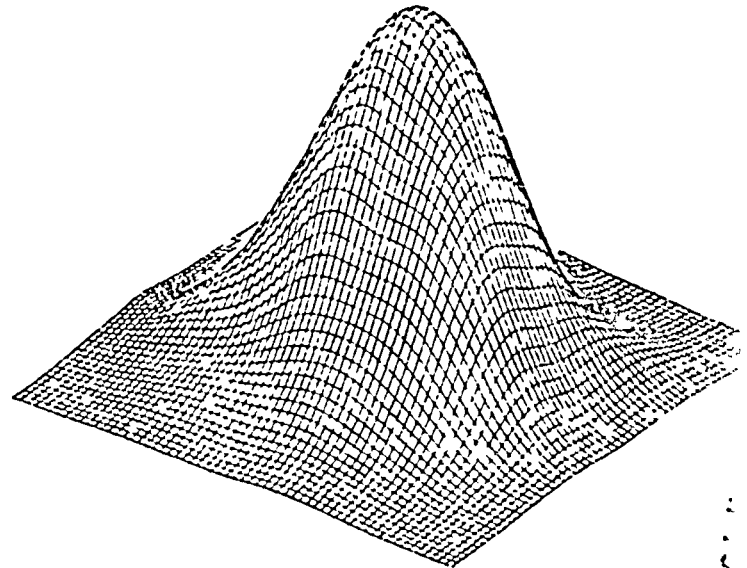


Figure 8.
Nonparametric maximum likelihood estimator of the bivariate normal of figure 5.



LOW QUALITY

Figure 9.
Marginal estimator of the bivariate normal of figure 6.

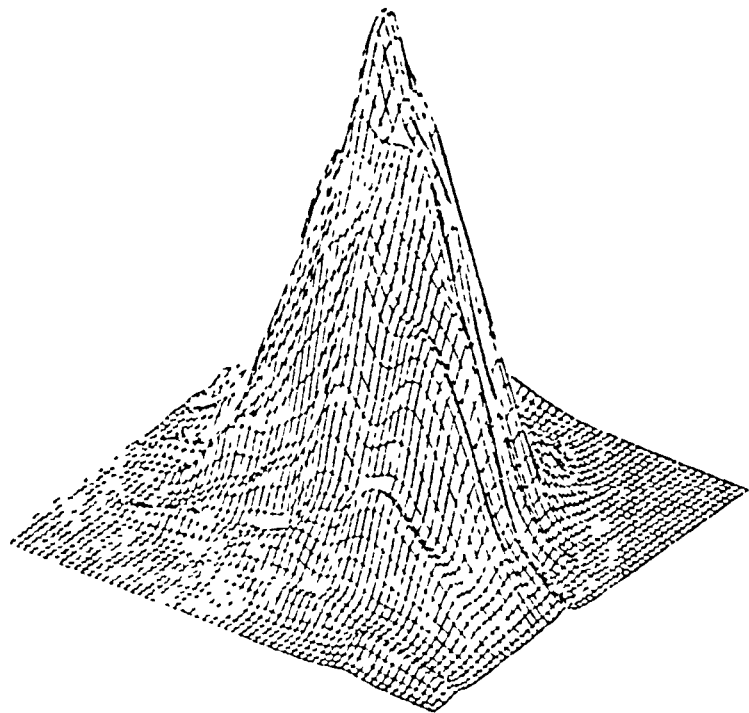
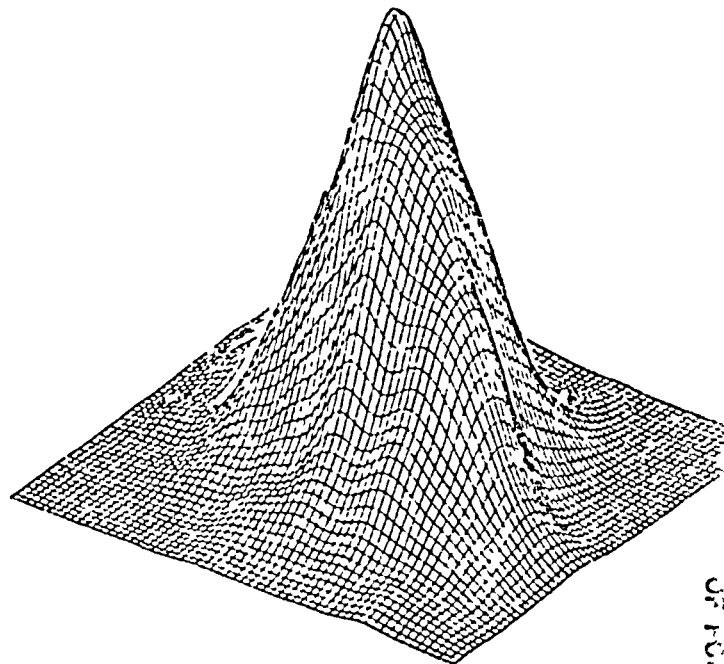


Figure 10.
Nonparametric maximum likelihood estimator of the bivariate normal of figure 6.



OF POOR QUALITY.

References

- Breiman, L., Meisel, W., Purcell, E. (1977). Variable kernel estimates of multivariate densities and their calibration. Technometrics. Vol. 19, No. 2, 135-144.
- Duin, Robert P. W. (1976). On the choice of smoothing parameters for Parzen estimators of probability density functions. IEEE Transactions on Computers 1176-1179.
- Gregory, C. G. and Schuster, E. F. (1979). Contributions to non-parametric maximum likelihood methods of density estimation. Proceedings of Computer Science and Statistics: 12th Annual Symposium on the Interface, ed. Jane F. Gentleman, 427-431, University of Waterloo, Waterloo, Ontario, Canada.
- Habbema, J. D. F., Hermans, J. and van den Brock, K. (1974). A stepwise discriminant analysis program using density estimation. Compstat, 1974, Proceedings in Computational Statistics, Wien, Physics Verlag, 101-110.
- Schuster, E. F. and Gregory, C. G. (1978). Choosing the shape factor(s) when estimating a density. Bulletin of the Institute of Mathematical Statistics, Vol. 7, No. 5, 292. Presented at the Annual Statistical Meeting, San Diego, Calif., August 1978.
- Schuster, E. F. and Gregory, C. G. (1981). On the nonconsistency of Maximum Likelihood Nonparametric Density Estimators, Computer Science and Statistics: Proceedings of the 13th Symposium on the Interface, ed. William F. Eddy, 295-298, Springer-Verlag, New York.
- Van Ryzin, J. (1973). A histogram method of density estimation. Communications in Statistics. Vol. 2, No. 6, 493-506.
- Wahba, Grace (1978). Data-based optimal smoothing of orthogonal series density estimates. Technical Report No. 509, Department of Statistics, University of Wisconsin, Madison, Wisconsin.

N83

15783

UNCLAS

Review of Some Results in Bivariate Density Estimation

by

David W. Scott
Rice University

Presented: NASA workshop on "Density Estimation and Function Smoothing,"
Texas A&M University, March 11-13, 1982

Summary: In this paper, we review some recent results for choosing
smoothing parameters for some bivariate density estimators.
Some univariate results are reviewed as well.

Acknowledgment: Supported by ARO under grant DAAG-29-82-K-0014

Review of Some Results in Bivariate Density Estimation

1. Introduction

For representing and examining data in up to several dimensions, nonparametric density estimation provides an analytic tool that is simultaneously exploratory and confirmatory. Unusual data features that may be discovered or explored include multiple modes or clusters, as well as unusual isolated points. At the same time, nonparametric density estimators are confirmatory since they provide a consistent estimator of the true underlying sample density function under mild restrictions.

The family of nonparametric density estimators is diverse, including the histogram, frequency polygon, kernel estimator, series estimator, and penalized-likelihood estimators to name a few important choices. Each of these methods has one or more calibration or design parameters commonly referred to as smoothing parameters. The bin width for an equally-spaced histogram plays the role of the smoothing parameter; too wide a bin width gives an oversmoothed estimate while too narrow a bin width results in an undersmoothed or rough-looking figure. In the terminology of Tukey's exploratory data analysis, in the first case we see too much of the forest (i.e., the smooth) and in the latter case we see too many trees in the forest (too rough).

Much theoretical and some practical work has appeared on how to choose the smoothing parameter to provide the best approximation to the underlying density function. It is also the case that the smoothing parameter has a certain exploratory nature, where we dynamically adjust how much forest and how many trees we wish to see.

2. Univariate Density Estimation

2.1 Some Graphical Interactive Approaches

In recent years there have appeared some interesting algorithms that automatically pick a smoothing parameter appropriate for a given data set for a particular nonparametric density estimation procedure. Prior to the evolution of these algorithms, statisticians learned how to pick good smoothing parameters through simulation experiments and interactive graphical methods. These latter methods will be important to use even with the automatic methods for validation purposes, data exploration purposes, and in cases where the automatic methods return occasionally bad smoothing parameter values.

We can illustrate several graphical methods, some known, some not with a nonparametric kernel density estimator

$$\hat{f}(x) = \frac{1}{nh} \sum_i K\left(\frac{x_i - x}{h}\right) \quad (1)$$

The first method is to pick a decreasing sequence of smoothing parameters (h 's) and look at the corresponding sequence of density estimates. For some simulated Gaussian mixture data with $n = 300$, we show a sequence of estimates in Figures 1(A)-(C). It is important to start with obviously oversmoothed estimates (large h) and look at the resulting sequence of estimates that shows increasing fidelity to the data and then finally becomes contaminated with noisy fluctuations.

The preceding interactive approach is not very sensitive for discriminating among several apparently "good" estimate-pictures. A similar problem exists in curve fitting. Tukey points out that plots of the residuals = data - fit provide a greatly enhanced ability to compare

the goodness of cover. A better modification is to examine plots of \hat{f}'' rather than of \hat{f} itself. The second derivative is much more sensitive to small changes in h than the density function itself. This is the procedure advocated by Silverman (1978), which results in pictures he calls "test graphs." With a little more experience and experimentation, we can go through a sequence of test graphs and accept a test graph with an desired amount of noisy fluctuations. In Figure 2, we reproduce three test graphs presented by Silverman.

A third procedure provides a useful shortcut and sometimes welcome relief to the previous methods. A possible choice of a measure of the roughness contained in a univariate test graph for a Gaussian kernel is

$$\int \hat{f}''(x)^2 dx = \frac{3}{8n^2 h^9} \sum_{ij} [h^4 - (x_i - x_j)^2 h^2 + \frac{1}{12} (x_i - x_j)^4] e^{-\frac{(x_i - x_j)^2}{4h^2}} \quad (2)$$

We simply plot the logarithm of equation (2) as a function of h . The graph has slope near zero for values of h corresponding to moderate oversmoothing and very large slope for values of h corresponding to rough estimates approaching Dirac spikes at the sample points. In Figure 3 we show six examples for various simulated data sets of this so-called "h-rough" plot. Also shown in each h-rough graph is a point labelled "best h ." This is the particular choice of h for that sample that minimized the integrated squared error (ISE) between the sampling density f and the estimate \hat{f} and is given by

$$ISE = \int (\hat{f}(x) - f(x))^2 dx.$$

It is clear that good choices of h lie in the region where the slope of

the h-roughness graph is about $h \approx 0.01$.

... out the line $y = x$.

Other useful approaches are based on rules of thumb derived from asymptotic theoretical results. For example, Scott (1979) proved that the optimal bin width h for an equally spaced histogram density estimator is given by

$$h = \left[\frac{6}{\int f'(x)^2 dx} \right]^{1/3} n^{-1/3}. \quad (3)$$

The rule of thumb he proposed was to choose

$$h^* = 3.5 s_x n^{-1/3}, \quad (4)$$

a formula based on using equation (3) and data moment estimators assuming the sampling data is $N(\mu, \sigma^2)$. He also provided multiplicative correction factors based on higher order sample moments such as the skewness. In Figure 4, we show 3 histograms of the same simulated Normal data with $n = 1000$. These figures also illustrate the usefulness of the integrated squared error criterion upon which equation (3) is based. Also notice how the sequential interactive approach works well here.

One automatic method for picking a kernel smoothing parameter is called the "quasi-optimal" procedure (Scott, 1976). It is based upon the well-known theoretically optimal choice for

$$h = h^* = \beta(f) = \beta \left(\int f''(x)^2 dx \right) \quad (5)$$

For a particular choice of h , we have a ready estimate for the right-hand side of equation (5) using equation (2) for a Gaussian kernel. The quasi-optimal smoothing parameter is the largest stationary point of

the right- and left-hand sides of equation (5) as a function of h . Stationary points are marked by arrows and occur when the lines intersect. This and several other automatic procedures have recently been compared by Monte Carlo methods (Scott and Factor, 1981).

2.2 Other Univariate Procedures

A new density estimator was proposed (Scott, Tapia & Thompson, 1979) based on the maximum penalized-likelihood criterion:

$$\ln L(f) = \sum_1 \ln[f(x_i)] - \alpha \int f''(t)^2 dt . \quad (6)$$

If we optimize (6) over the class of continuous piecewise-linear functions defined on a given mesh we obtain the DMPLE - the discrete maximum penalized-likelihood estimator, a code for which exists in the IMSL library (1982, NDMPLE). Here α , the penalty or roughness weight, plays the role of the smoothing parameter. While consistency of the DMPLE is well-known, we have few theoretical results on actual convergence behavior. Extensive numerical simulations indicate that the rate of convergence is $n^{-4/5}$, the same as for many other techniques (except, for example, for the histogram, which is $n^{-2/3}$). However, these same simulations indicate that the DMPLE is very efficient for the sampling densities examined.

In Table I, we examine sample sizes required to achieve an average integrated squared error of $1/400$ for Gaussian sample data $N(0,1)$. A complete picture of the general behavior of the DMPLE for various penalty functions and sampling densities is an open area of research.

Next the frequency of ... values by line segments. It is infrequently used. Note the DNPLE has the same form if not origin as the frequency polygon. However, I have recently shown that the frequency polygon properly constructed actually shares the same approximation properties as the kernel methods rather than the poorer properties of the histogram; that is, it converges at the rate $n^{-4/5}$; see Scott (1982). This observation was recently made independently by David Freedman at SLAC. However, this is a whole paper in itself. But notice how the frequency polygon behaves in Table I. This is generally the case.

With the above as background, we can look more closely at some corresponding two-dimensional results.

3. Bivariate Density Estimation

3.1 Need for Two Smoothing Parameters

The bivariate kernel estimator is given by

$$\hat{f}(x,y) = \frac{1}{n} \sum_i K_{x,y}(x_i, y_i), \quad (7)$$

that is, the kernel varies from point to point. A more useful form is

$$\hat{f}(x,y) = \frac{1}{nh_x h_y} \sum K_0 \left[\frac{x_i - x}{h_x}, \frac{y_i - y}{h_y} \right] \quad (8)$$

where K_0 is a bivariate density function with certain restrictions, but whose exact form is secondary in importance to the choices of h_x and h_y .

by usually K_0 is a non-degenerate kernel satisfying the following

symmetry condition: $K_0(x,y) = K_0(-x,-y)$.

Cacoullus (1964) examined this general case and, in fact, proposed the simpler product kernel

$$K_0\left[\frac{x_1-x}{h_x}, \frac{y_1-y}{h_y}\right] = K_1\left[\frac{x_1-x}{h_x}\right] K_1\left[\frac{y_1-y}{h_y}\right] \quad (9)$$

This form has certain computational advantages especially when the univariate kernel K_1 has finite support. Cacoullus wrongly proves in his last theorem that optimally for product kernels we should restrict ourselves to

$$h_x = h_y \quad (10)$$

Nezames (1980) has considered this question (and much of the following material comes from her thesis). Suppose f is bivariate Normal with covariance matrix

$$\Sigma = \sigma^2 \begin{pmatrix} 1 & 0 \\ 0 & q^2 \end{pmatrix} \quad (11)$$

Then in Table II we look at the inefficiency with respect to average ISE for the restricted optimization problem satisfying equation (10) versus the unrestricted optimization problem. The results given in the Table emphasize how large this inefficiency can be. The obvious fix is to standardize the data so $s_x = s_y$. However, the behavior shown in Table II is the result of complicated functions of second order derivatives of f and not simply functions of the moments, in general.

For other bivariate methods, the above results emphasize the need for having at least one smoothing parameter for each variable.

3.2 Optimal Kernels

For univariate kernel estimation, Epanechnikov (1969) proved that the optimal kernel was of the form

$$K_1^*(x) = \frac{3}{4} (1-x^2) \quad -1 < x < 1 \quad . \quad (12)$$

Nezames has proven the following:

Theorem: The optimal bivariate kernel K_0 is given by

$$K_0^*(x,y) = \frac{1}{3\pi} \left[1 - \frac{1}{6}(x^2+y^2) \right] \quad x^2+y^2 < 6 \quad . \quad (13)$$

This kernel looks like kernel K_1^* swept 360° about the z-axis. The increase in efficiency of the optimal kernel (13) compared to other kernels is not large, a situation similar to the one-dimensional case investigated by Epanechnikov; see Table III. Notice how the product kernels are only slightly inefficient. The Gaussian product kernel is perhaps surprisingly inefficient.

3.3 Picking the Smoothing Parameters h_x and h_y

All of the one-dimensional methods described in section 2.1 may be directly extended to the 2-dimensional case. Direct sequential bivariate iterations are much more time consuming and difficult to perform repeatedly. The test graph approach is less easy to visualize than the density estimate (using contours, say) because the test graph will have contours corresponding

to negative values. However, the test graph is still a more sensitive instrument than the direct interactive approach.

There also exists a bivariate quasi-optimal algorithm implementation that has been evaluated in some simple cases by Nezames.

There are also bivariate extensions of rules of thumb based on theoretical results. For example, for bivariate histograms with rectangular bins of size h_x by h_y , the data-based rules are

$$\begin{aligned} h_x &= 3.5 s_x n^{-1/4} & \text{and} \\ h_y &= 3.5 s_y n^{-1/4}, \end{aligned} \tag{14}$$

expressions virtually the same as the one-dimensional result in equation (4), except for the exponent on n . The first correction to equation (14) is based on the sample correlation coefficient r .

Equation (14) should be divided by

$$(1 - r^2)^{3/8} . \tag{15}$$

Higher order moment corrections could be developed.

We next consider the smoothing of a bivariate series estimator using the cross-validation algorithm of Wahba (1981). The smoothing parameters used minimize a certain generalized cross-validation functional.

Depending upon the exact form of the initial series estimator, you get either the algorithm given by Wahba (1981) or a slightly different version developed by Nezames (1980). First, for $n = 50$, and $\rho = .80$ with bivariate Gaussian sample data, we show contours of the cross-validation functions for the two approaches, see Figures 6 and 7. The two corresponding estimates are shown in Figures 8 and 9.

The edge effects of series estimators are well-known to exist but are always somewhat surprising to see. In Figure 10 we show an estimate for $n = 100$, $\rho = 0$, bivariate Gaussian data. Notice how the periodic nature of the solution is clear.

3.4 Rates of Convergence

In Table IV, we summarize the rates of convergence of the various density estimators. The frequency polygon again performs very well. Notice that the two-dimensional kernel methods have the same convergence rate as the one-dimensional histogram. As an aside, the bivariate frequency polygon may best be constructed using histograms with base bins in the shape of hexagons; that is, a shape capable of tiling the plane and approximating a circle.

3.5 Bivariate DMPLE

Nezames has implemented the Bivariate DMPLE for the class of piecewise constant functions. As an example, $n = 200$, $\rho = 0$ bivariate Gaussian sample, the histogram is shown in Figure 10 and the corresponding DMPLE in Figure 11. Notice the reduction in noise and false peaks in Figure 11.

3.6 Scatter Diagrams or Density Estimate Contours?

One thing statisticians are supposed to do well is examine scatter diagrams, such as those from residual plots. It has been my experience that the naked scatter diagram is a difficult object to "see." For example,

consider the blood lipid (fat) data shown in Figure 12 (Scott, et al, 1978). These data represent the cholesterol and triglyceride values of 320 males with angiographically demonstrated coronary artery disease. Now look at the contours of a kernel estimate of the same data shown in Figure 13. The bimodal feature was an important undiscovered feature in previous analyses of these data.

- Cacoullos, T. (1966). Estimation of a multivariate density. Ann. Inst. Statist. Math., Tokyo, 18, 179.
- Epanechnikov, V.A. (1969). Nonparametric estimates of a multivariate probability density. Theor. Prob. Appl., 14, 153.
- International Mathematical and Statistical Libraries, Inc. (1981). Houston, TX.
- Nezames, D.D. (1980), "Some Results for Estimating Bivariate Densities by Kernel, Orthogonal Series and Penalized Likelihood Procedures," Unpublished Ph.D. thesis, Rice University, Houston, TX.
- Scott, D.W. (1976), "Nonparametric Probability Density Estimation by Optimization Theoretic Techniques," Unpublished Ph.D. thesis, Rice University, Houston, TX.
- Scott, D.W. (1979). "On Optimal and Data-Based Histograms." Biometrika 66:605-610.
- Scott, D.W. (1982), "Frequency Polygons: Theory and Application," in preparation.
- Scott, D.W. and L.E. Factor (1981). "Monte Carlo Study of Three Data-Based Nonparametric Density Estimators." J. American Statistical Association 76:9-15.
- Scott, D.W., A.M. Gotto, J.S. Cole, and G.A. Gorry (1978). "Plasma Lipids as Collateral Risk Factors in Coronary Artery Disease: A Study of 371 Males with Chest Pain." Journal of Chronic Diseases 31:337-345.
- Scott, D.W., R.A. Tapia, and J.R. Thompson (1980). "Nonparametric Probability Density Estimation by Discrete Maximum Penalized-Likelihood Criteria." Annals of Statistics 8:820-832.

...ability," Biometrika, 65:1-11.

Tapia, R.A. and J.R. Thompson (1976). Nonparametric Probability Density Estimation. Baltimore: Johns Hopkins Press.

Tukey, J.W. (1977). Exploratory Data Analysis. Reading, Mass: Addison-Wesley.

Wahba, G. (1981), "Data-Based Optimal Smoothing of Orthogonal Series Density Estimates," Ann. Statist.

ORIGINAL PAGE IS
OF POOR QUALITY

... for an i.i.d. = 1/400 with $i(0,1)$ data (See Text)

Estimator	Equivalent Sample Size
$K(\bar{x}, 1)$	57
$K(\bar{x}, s^2)$	100
Epanechnikov kernel	431
Boxcar kernel	463
Frequency Polygon	546
Histogram	2,256
D.F.P.L.U.	160

Table II

Table 2.2.1

q^2	r
∞ or 0	∞
$\pm 10^3$ or $\pm 10^{-3}$	72.11
± 100 or $\pm .01$	15.54
± 50 or $\pm .02$	9.79
± 10 or $\pm .1$	3.35
± 5 or $\pm .2$	2.13
± 2 or $\pm .5$	1.23
± 1	1.00

ORIGINAL PAGE IS
OF POOR QUALITY

OF POOR QUALITY

Table 2.3.1

$K(x,y)$	Support	A	R
$K^*(x,y)$	$x^2 + y^2 \leq 6$	0.1710	1
$K_1 = \frac{9}{16} (1 - x^2)(1 - y^2)$	$ x , y \leq 1$	0.1731	1.0121
$K_2 = \frac{225}{256} (1 - x^2)^2(1 - y^2)^2$	$ x , y \leq 1$	0.1741	1.0179
$K_3 = \frac{1}{2\pi} \exp\{-\frac{x^2}{2} - \frac{y^2}{2}\}$	$ x , y < \infty$	0.1850	1.0819
$K_4 = 1/4$	$ x , y < 1$	0.1908	1.1157

Table IV

	univariate	bivariate
histogram	$n^{-2/3}$	$n^{-1/2}$
kernel, series, freq polygon	$n^{-4/5}$	$n^{-2/3}$

UNIVERSITY OF CALIFORNIA, BERKELEY QUANTILE METHOD $N(N) = 2.0$

Figure 1(A)

ORIGINAL PLOT IS
OF POOR QUALITY

UNIMODAL MINIMAL WITH PLANS = 1 1.50 VARIANCE OF LEFT = 1 WITH WEIGHT AND VARIANCE OF RIGHT = 0.2500 0.1111
 SAMPLE SIZE = 300 WITH 75 SAMPLES ON RIGHT VS.
 KERNEL ESTIMATOR WITH KERNEL = GAUSS WITH H = 0.00000000
 INTEGRATED MEAN SQUARE ERROR 0.527520661 E-02
 INTEGRATED SQUARE ERROR 0.268321602 E-01
 MAXIMUM ABSOLUTE DIFFERENCE 0.176553645 E+00

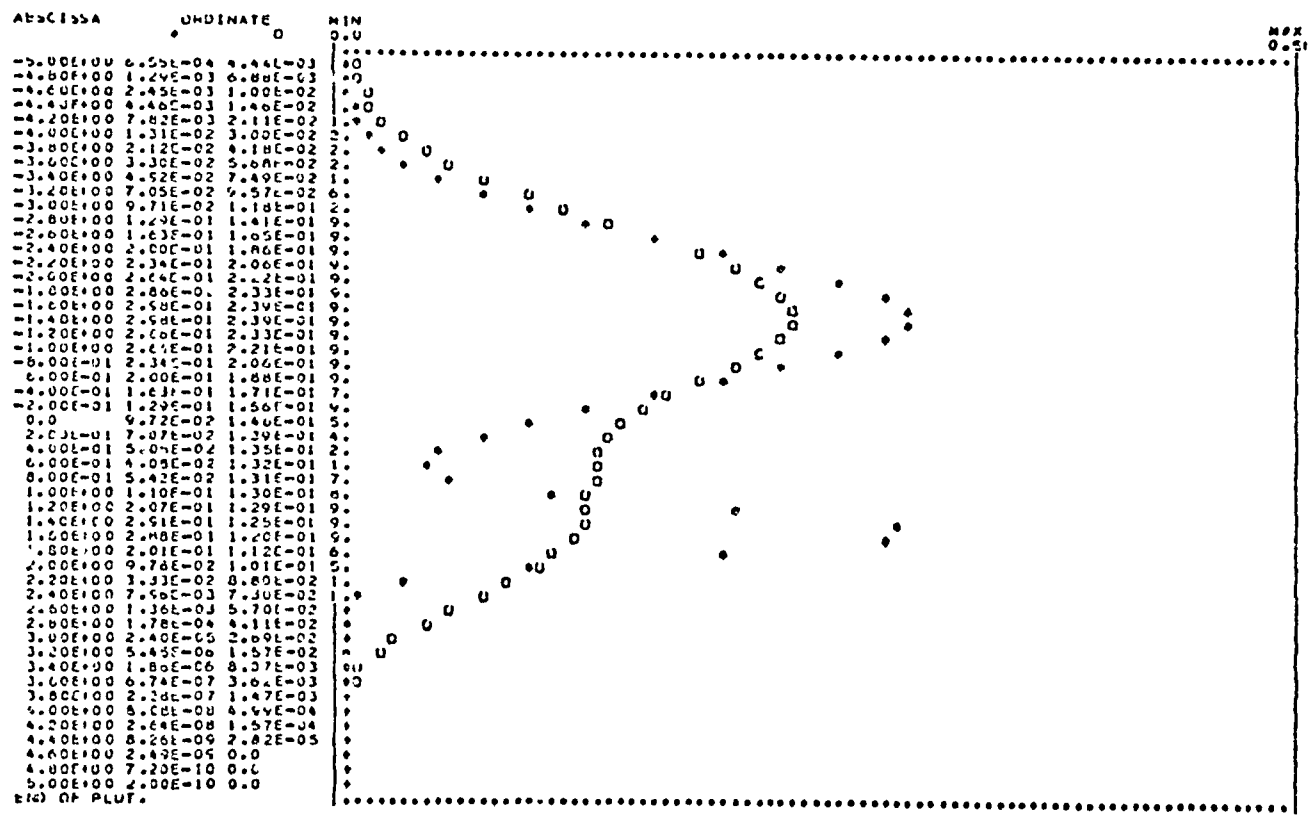
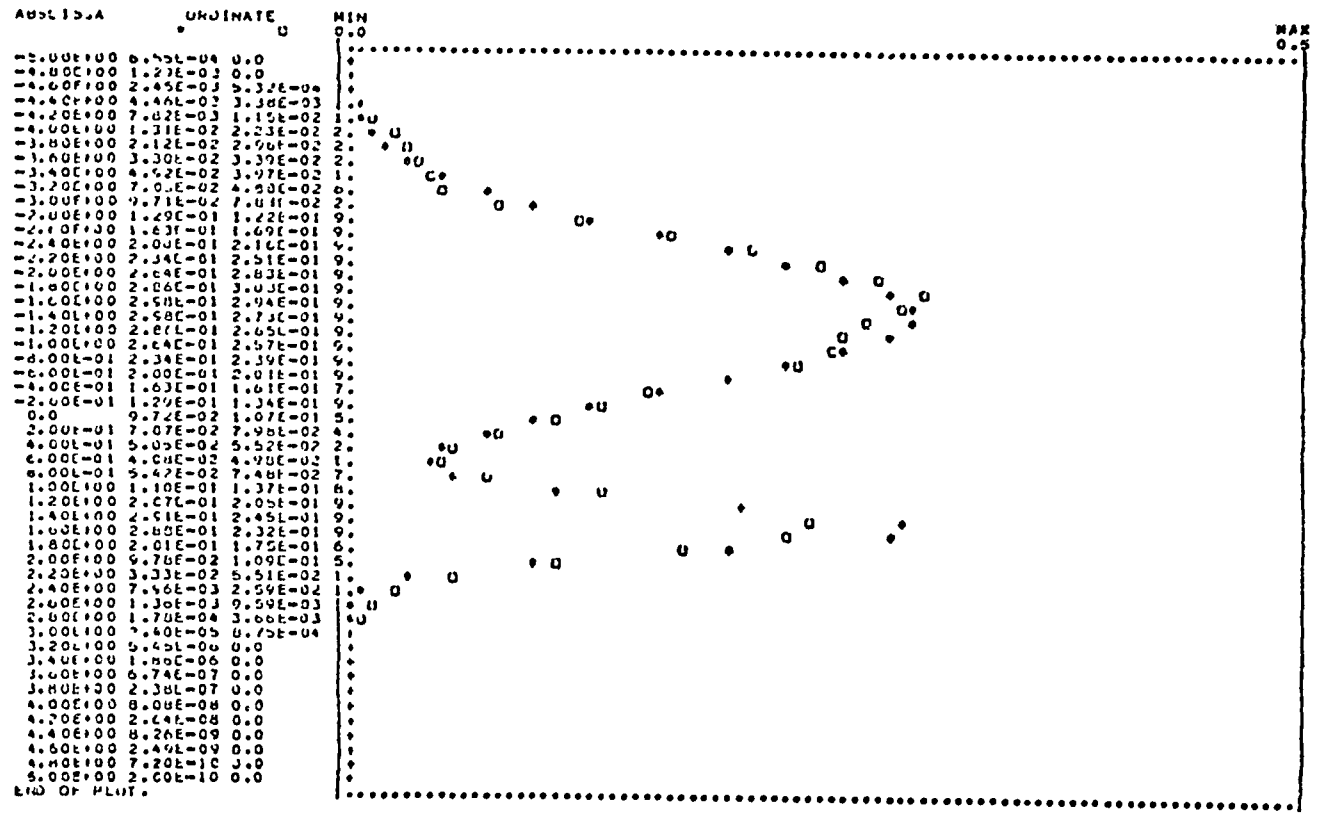


FIGURE 113

ORIGINAL PAGE IS
OF POOR QUALITY

UNIFORM NORMAL WITH MEAN = 0.50 VARIANCE OF LEFT = 1 WITH WEIGHT AND VARIANCE OF RIGHT = 0.2000 0.1111
 SAMPLE SIZE = 100 WITH 75 SAMPLES ON RIGHT VS.
 KERNEL ESTIMATOR WITH KERNEL = GAUSS WITH H = 0.60000002
 INTEGRATED MEAN SQUARE ERROR 0.303594056 E-03
 INTEGRATED SQUARE ERROR 0.236094631 E-02
 MAXIMUM ABSOLUTE DIFFERENCE 0.57475073 E-01



C-3

BINOMIAL MINIMAX WITH MEANS $\lambda = 1.50$ VARIANCE OF LEFT $\lambda = 1$ WITH WEIGHT AND VARIANCE OF RIGHT $\lambda = 0.0000$ 0.0111
 SAMPLE SIZE = 100 WITH 75 SAMPLES ON RIGHT VS.
 KERNEL ESTIMATION WITH KERNEL λ QUAD WITH $h = 0.1666666$
 INTEGRATED SQUARE ERROR 0.14030804 E-02
 INTEGRATED SQUARE ERROR 0.68230109 E-02
 MAXIMUM ABSOLUTE DIFFERENCE 0.936030044 E-01

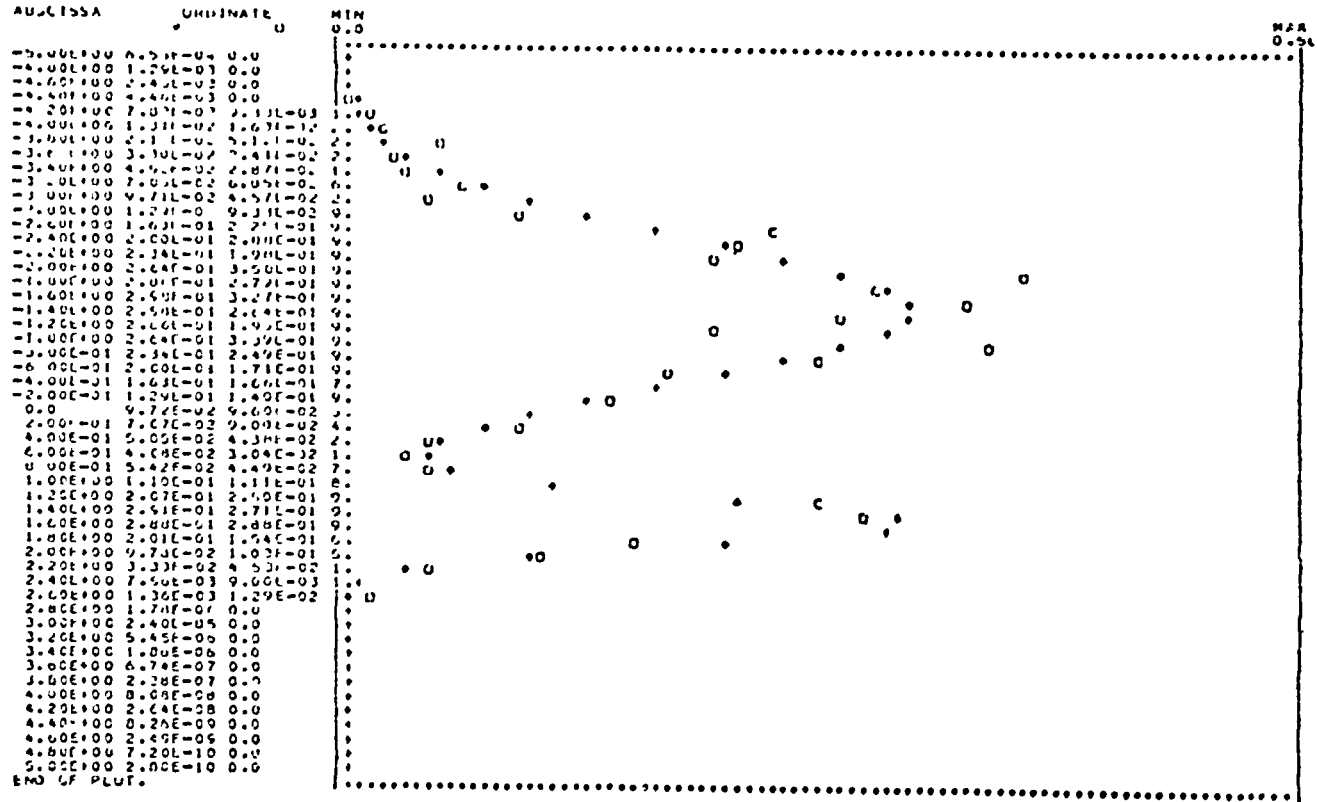


Figure 1.5
 ORIGINAL PAGES IS
 OF POOR QUALITY

Choosing the window width when estimating a density

FIGURE 2.

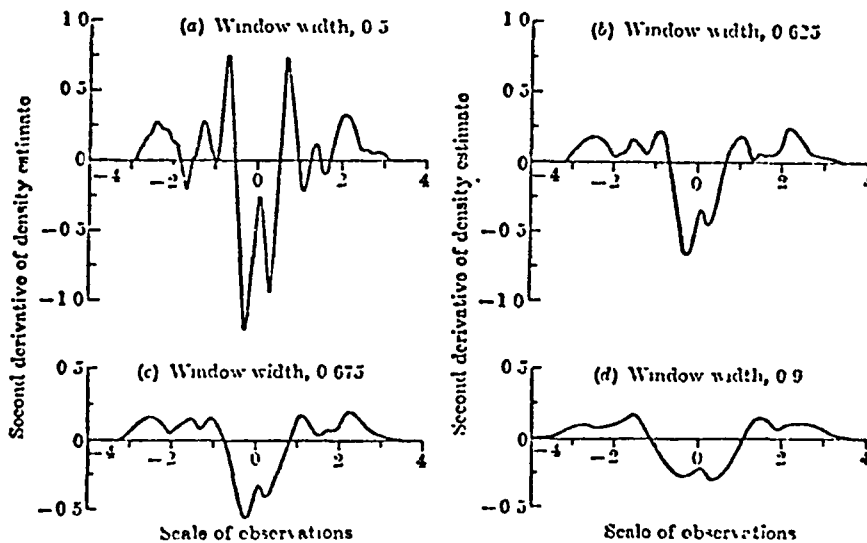
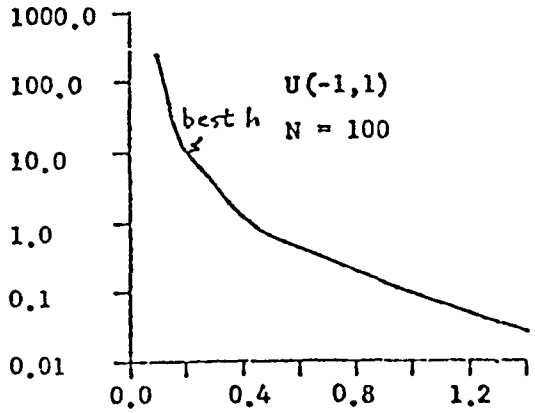
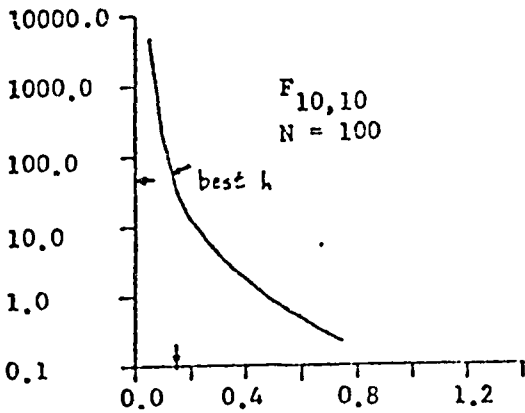
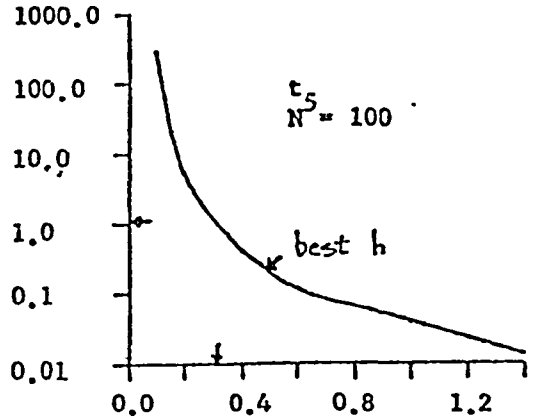
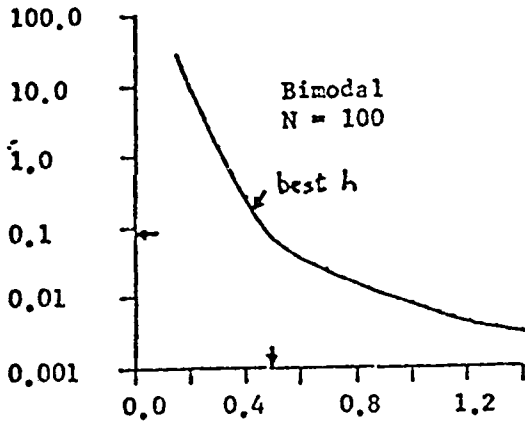
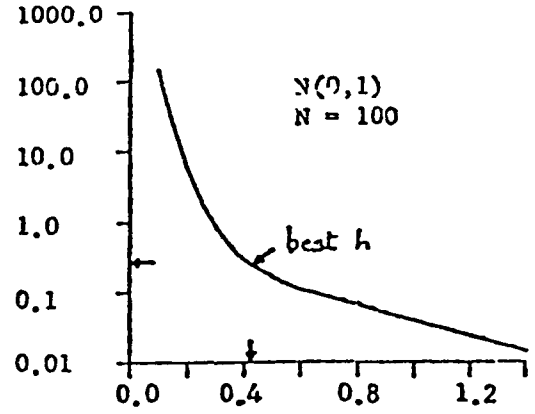
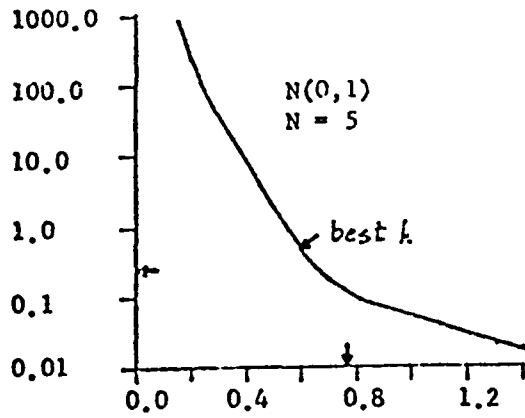


Fig. 1. Test graphs for 100 normal observations for window widths, 0.5, 0.625, 0.675 and 0.9.

FIGURE 2.5.1. Graphs of Equation (2.5.1) vs. h .



$f = N(0,1)$

$N = 1000$

FIGURE 4

ORIGINAL PAGE IS OF POOR QUALITY

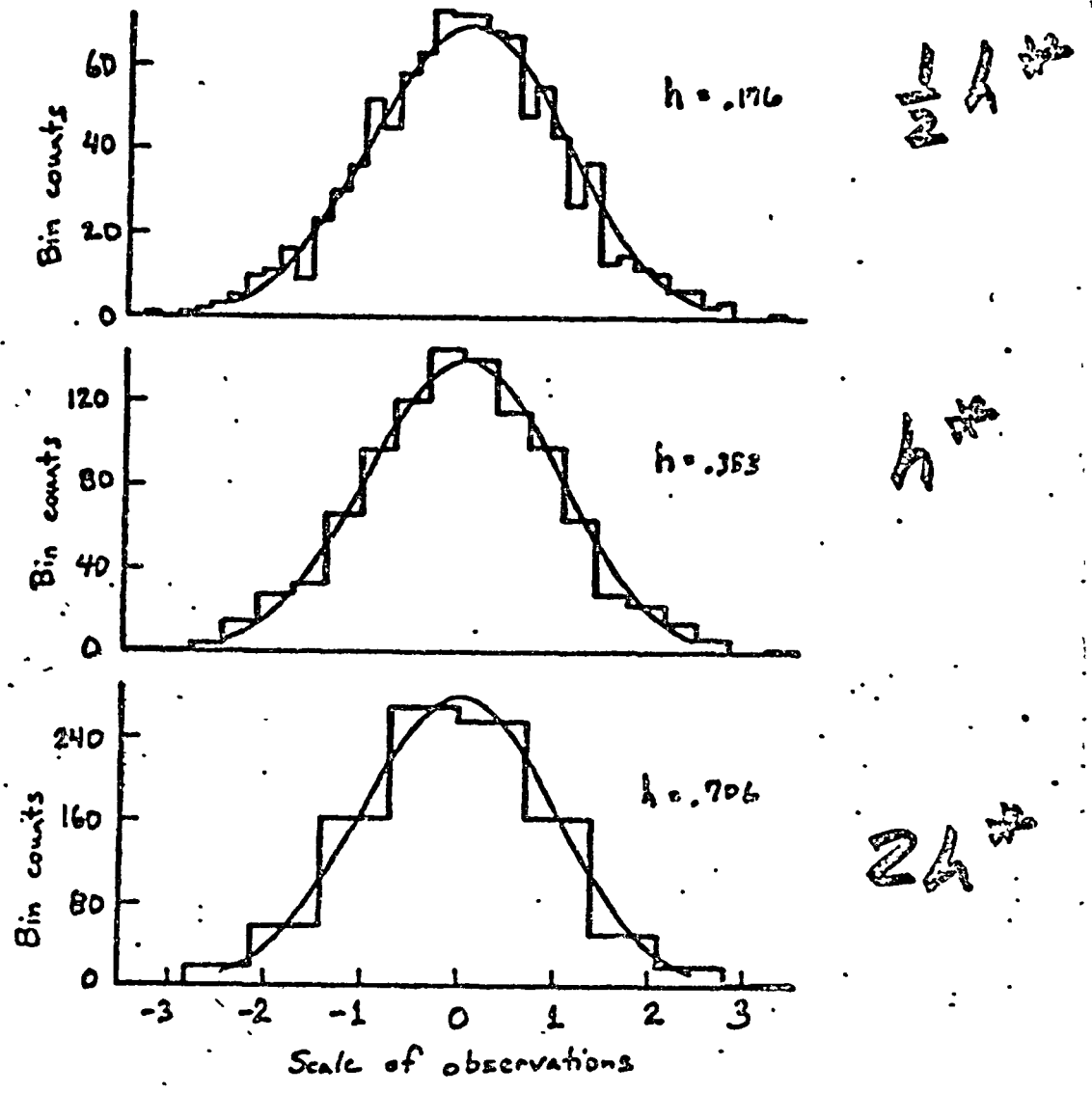


Figure 24 (see p.10)

FIGURE 5.0.1. PLOTS OF PROBABILITY DENSITY FUNCTIONS
FOR SEVERAL SAMPLES

FIGURE 5

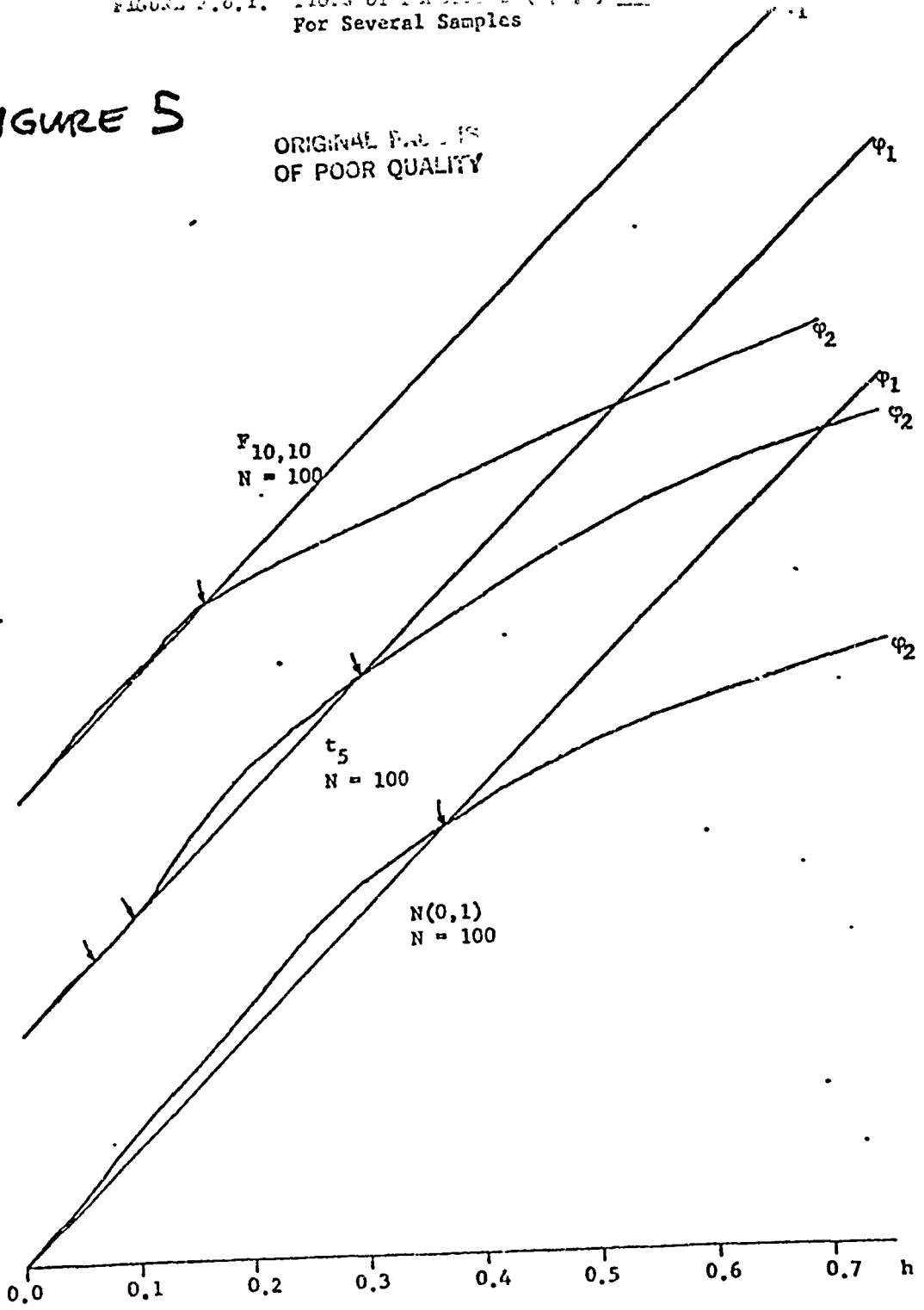
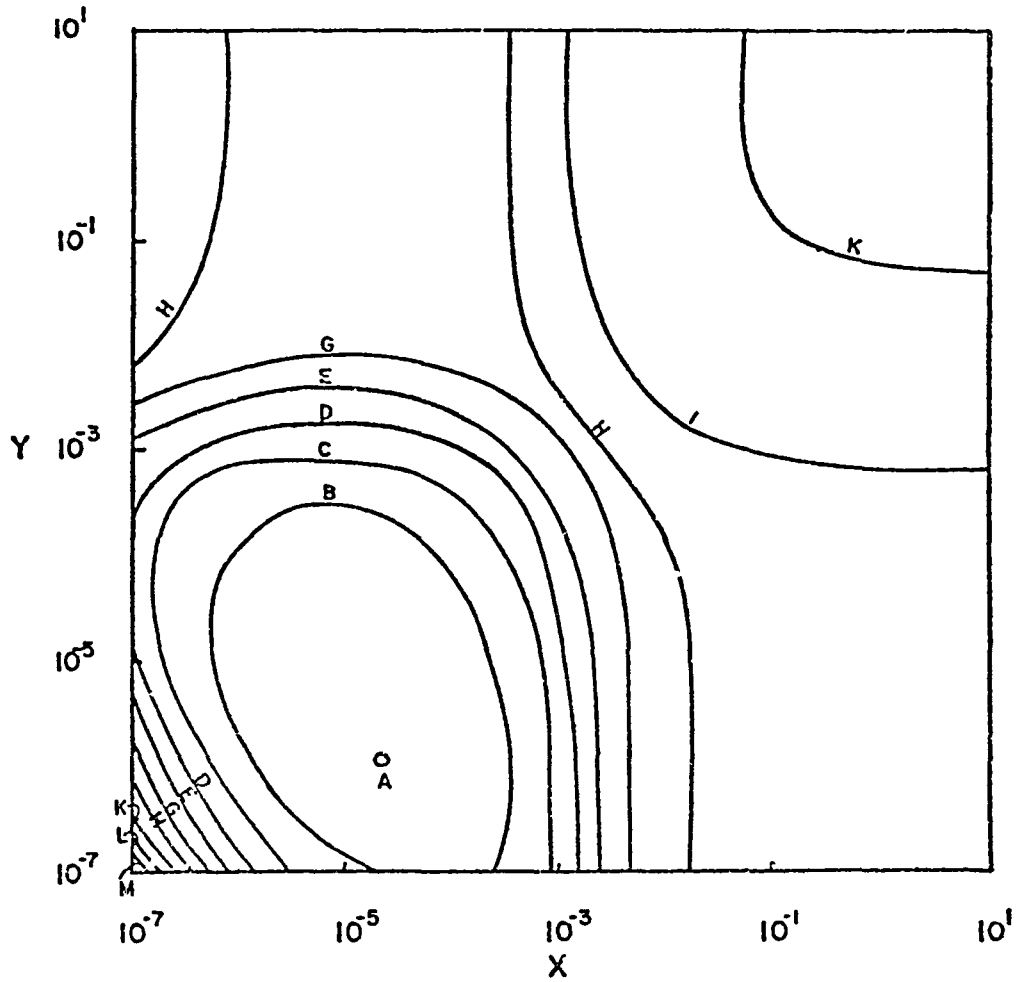


Figure 3.2.1

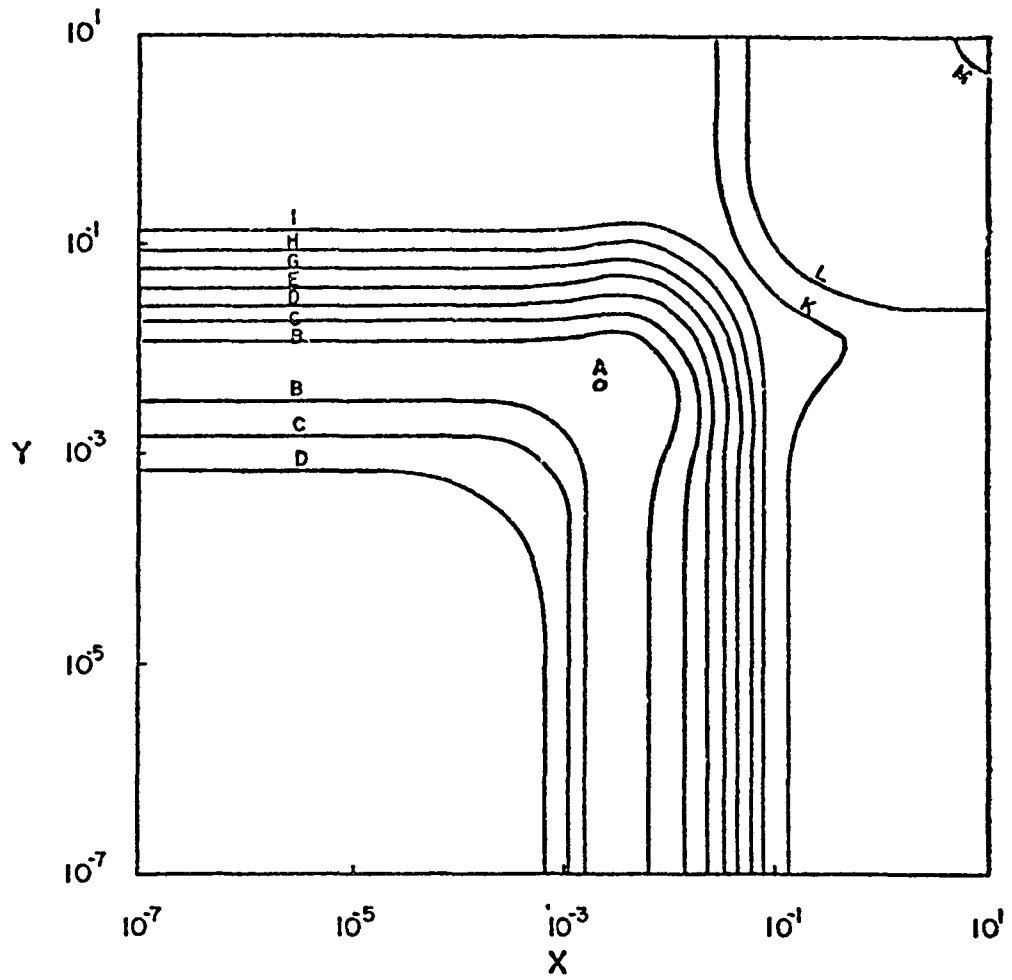
CONTOUR PLOTS
OF POOR QUALITY



A = -148	H = 1.03
B = -1.06	I = 1.45
C = -.65	K = 1.87
D = -.23	L = 2.29
E = .19	M = 2.71
G = .61	

NEZAMES' CRITERION FUNCTION

FIGURE 3.2.1

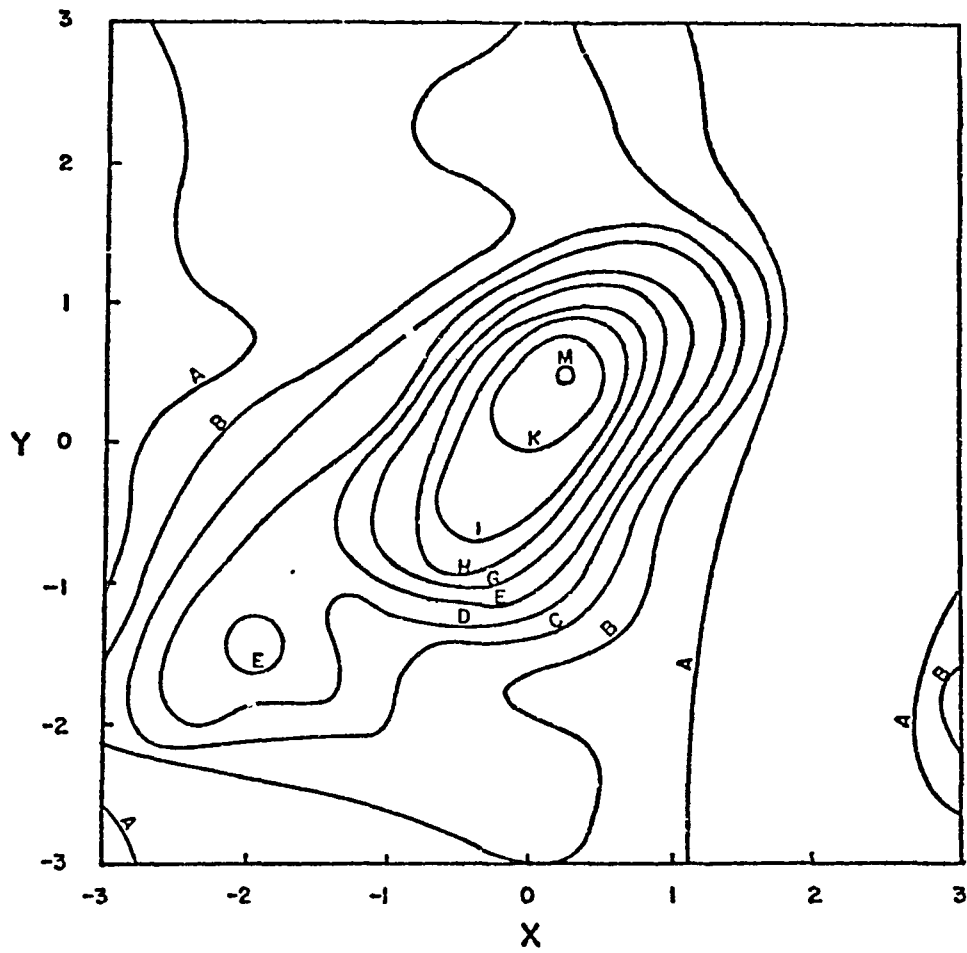


A = .58	H = 2.54
B = .91	I = 2.86
C = 1.24	K = 3.19
D = 1.56	L = 3.51
E = 1.89	M = 3.84
G = 2.21	

WAHBA'S CRITERION FUNCTION

FIGURE 3.2.2

OF POOR QUALITY.



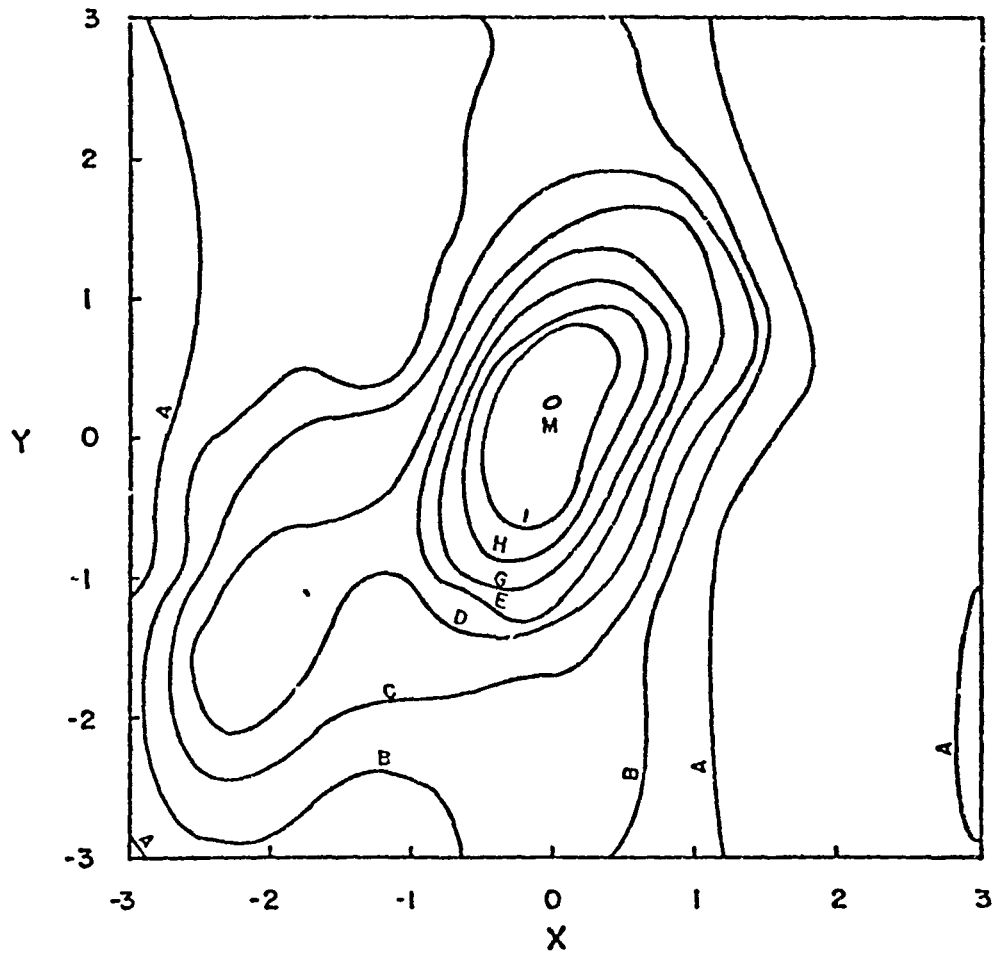
A=	-.009	G=	.135
B=	.020	H=	.154
C=	.049	I=	.193
D=	.078	K=	.222
E=	.106	M=	.251

NEZAMES' DENSITY ESTIMATE

FIGURE 3.2.3

Figures 1

ORIGINAL FILE IS
OF POOR QUALITY



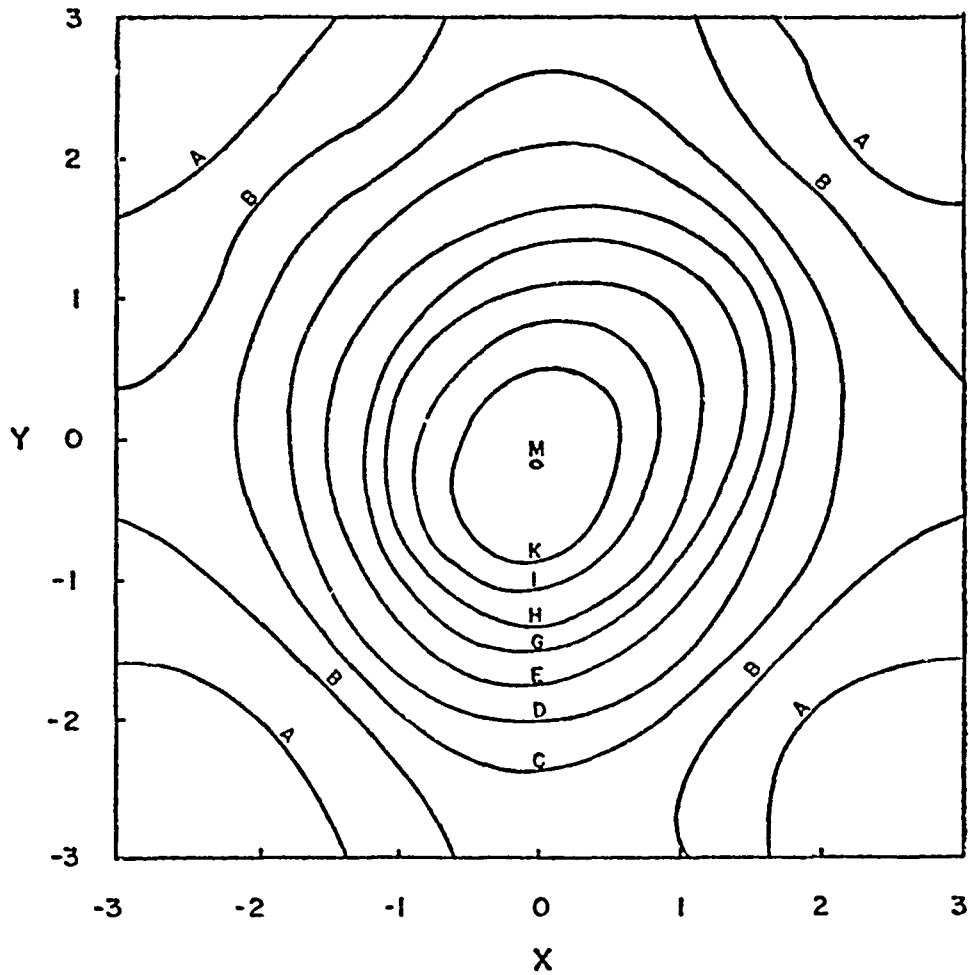
A = -.008	G = .137
B = .021	H = .166
C = .050	I = .196
D = .079	M = .254
E = .108	

WAHBA'S DENSITY ESTIMATE

FIGURE 3.2.4

FIGURE 10

ORIGINAL PAGE IS
OF POOR QUALITY



A = -.030	G = .093
B = -.006	H = .118
C = -.019	I = .142
D = .042	K = .167
E = .068	M = .191

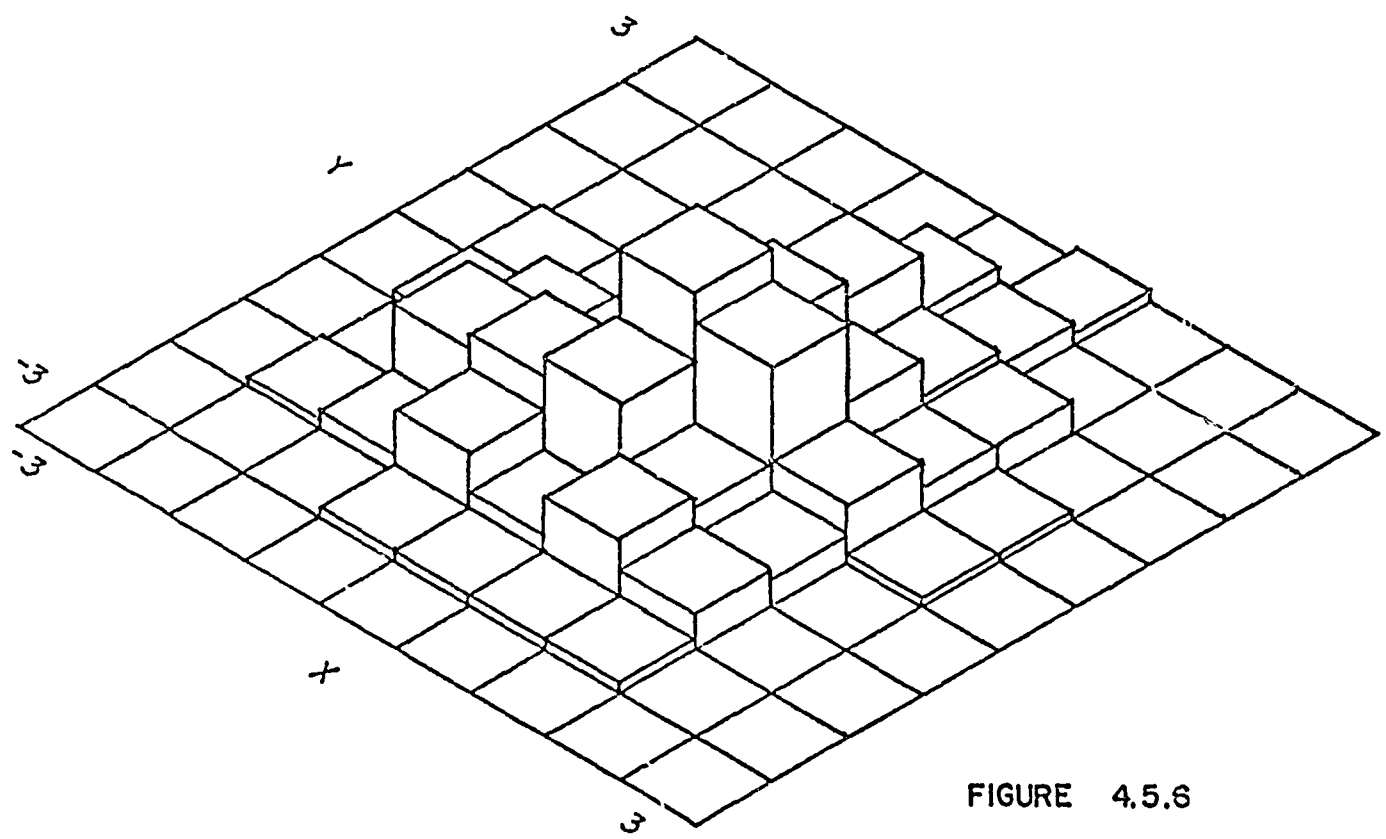
NEZAMES' DENSITY FUNCTION

$$v: |v| \leq 15$$

FIGURE 3.2.9

FIGURE 11 A

HISTOGRAM (M = 9)



ORIGINAL PAGE IS
OF POOR QUALITY

FIGURE 4.5.6

FIGURE 11 B

DISCRETE MAXIMUM PENALIZED LIKELIHOOD DENSITY ESTIMATE
(ALPHA = 1000, M=9)

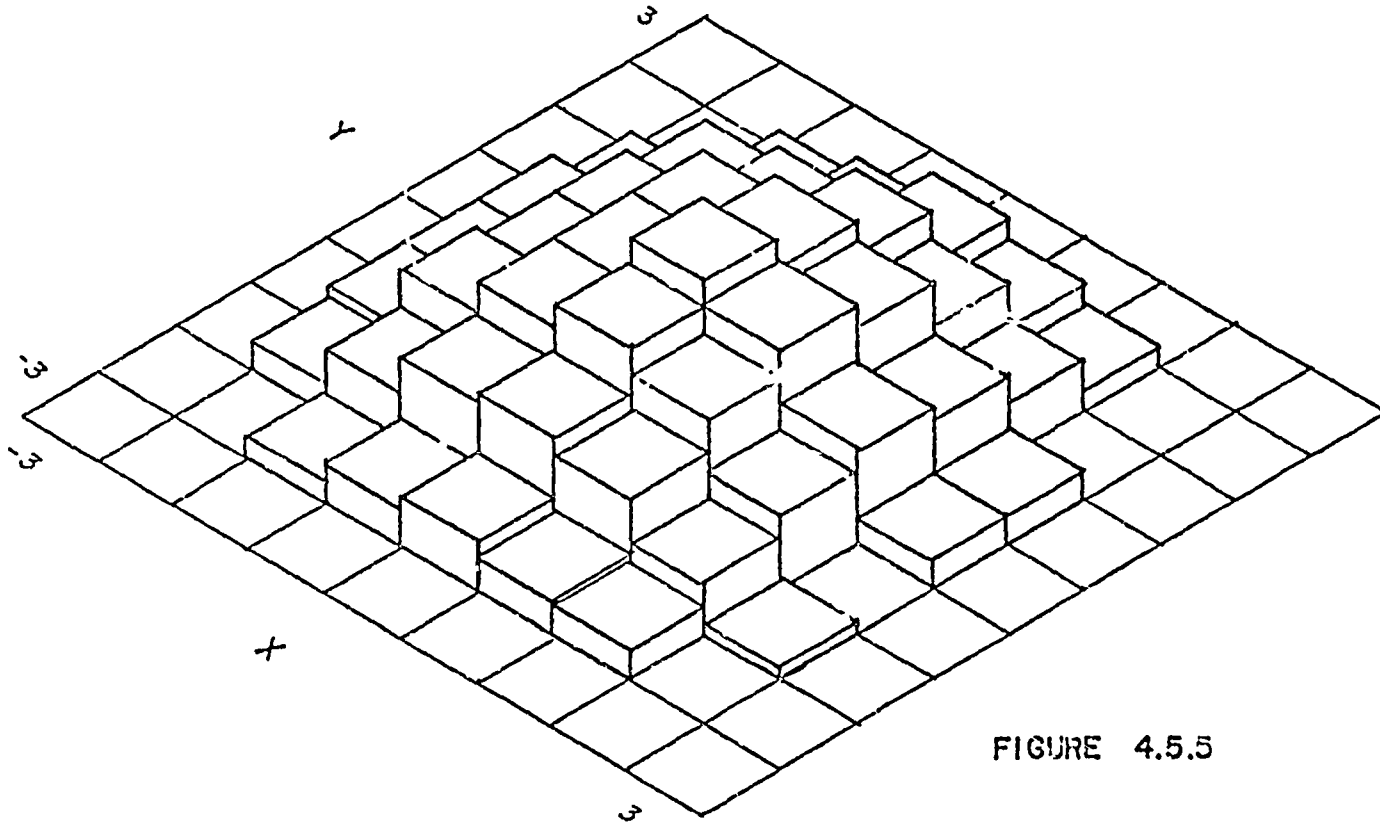


FIGURE 4.5.5

ORIGINAL PAGE IS
OF POOR QUALITY

N83

15784

UNCLAS

72
ORIGINAL COPY
OF POOR QUALITY

A BOOTSTRAP APPROACH TO BUMP HUNTING

B.W. Silverman
School of Mathematics
University of Bath
Claverton Down,
Bath, BA2 7AY.
England.

Presented at the NASA Workshop on "Density Estimation and Function Smoothing", Texas A. & M. University, 11 March 1982.

1. INTRODUCTION

An important question in cluster analysis and pattern recognition is the determination of the number of clusters into which a given population should be divided. Frequently, particularly when certain specific clustering methods are being used, the number of clusters is taken to be equal to the number of modes, or local maxima, in the probability density function underlying the given data set; in some applications this question is of direct interest in its own right.

Investigation of the number of local maxima in a density or its derivative has been considered by several authors, for example Cox (1966) and Good and Gaskins (1980). Most methods seem to depend on some arbitrary implicit or explicit choice of the scale of the effects being studied; see the remarks of Silverman (1980). The simple approach based on kernel density estimates described in this paper has the virtue of making this choice in an automatic and natural way.

The use of kernel density estimates in mode estimation was originated by Parzen (1962). The 'gradient method' of cluster analysis is based on clustering towards modes in the estimated density; see, for example, Andrews (1972), Fukunaga and Hostetler (1975), and Bock (1977).

In Section 2 below the test statistic to be used is defined, and a bootstrap technique for assessing significance is given in Section 3. An illustrative application is given in Section 4. In Sections 5 and 6 the asymptotic behaviour of the test statistic is discussed.

For a published version of this work, see Silverman (1981a) and Silverman (1983).

2. THE CRITICAL WINDOW WIDTH

A possible test statistic for hypotheses concerning the number of modes in the density can be obtained by constructing kernel density estimates of the data. The kernel density estimate (Rosenblatt, 1956) for window width h based on univariate observations X_1, \dots, X_n is defined by

$$\hat{f}(t; h) = n^{-1} h^{-1} \sum_{i=1}^n K\{h^{-1}(t - X_i)\}, \quad (1)$$

where K is a kernel function, which we shall assume throughout to be the normal density function. Apart from the theoretical advantages of this choice, the use of a normal kernel has strong computational advantages; see Silverman (1981b).

The window width h controls the amount by which the data are smoothed to obtain the kernel estimate. Thus, for example, if the data are strongly bimodal a large value of h will be needed to obtain a unimodal estimate. Suppose that we wish to test the null hypothesis that the density f underlying the data has k modes, against the alternative that f has more than k modes; often $k = 1$. Define the k -critical window width h_{crit} by

$$h_{\text{crit}} = \inf \{h; \hat{f}(\cdot, h) \text{ has at most } k \text{ modes}\}. \quad (2)$$

Large values of h_{crit} will reject the null hypothesis. Silverman (1978) used a critical value of a smoothing parameter in a somewhat different context. The computation of h_{crit} in practice is facilitated by the following theorem and corollary.

Theorem Given any fixed X_1, \dots, X_n , define $\hat{f}(t, h)$ as in (1) above, using a normal kernel K . For each integer $m \geq 0$, the number of maxima as t varies in $\partial^m \hat{f} / \partial t^m$ is a right continuous decreasing function of h .

The following corollary follows at once.

Corollary. Defining h_{crit} as in (2) above, $\hat{f}(\cdot; h)$ has more than k modes if and only if $h < h_{\text{crit}}$.

The corollary shows that h_{crit} can be found by a binary search procedure, since for any value of h we can tell at once whether or not $h < h_{\text{crit}}$ by counting the number of modes in $\hat{f}(\cdot; h)$. The result is also used in Section 3 below.

This section is concluded with the proof of the theorem, which makes use of the theory of total positivity; see, for example, Karlin (1968). Let $v_{m+1}(h)$ denote the number of sign changes in $\hat{f}^{(m+1)}(\cdot, h)$. Since $(-t)^{m+1} \hat{f}^{(m+1)}(t, h)$ is, for all $m \geq 0$ and h , eventually positive as $t \rightarrow -\infty$ and as $t \rightarrow \infty$, it suffices to show that v_{m+1} is decreasing and right continuous. For $h_2 > h_1 > 0$, $\hat{f}^{(m+1)}(\cdot, h_2)$ is the convolution of $\hat{f}^{(m+1)}(\cdot, h_1)$ with a $N(0, h_2^2 - h_1^2)$ density, and $\hat{f}^{(m+1)}(\cdot, h_1)$ is continuous and bounded. Thus, by Theorem 2 of Schoenberg (1950), $v_{m+1}(h_2) \leq v_{m+1}(h_1)$ so that v_{m+1} is decreasing. Now suppose, for given $h_0 > 0$, there exist $a_1 < b_1 < a_2 < \dots < a_r < b_r$ such that $\hat{f}^{(m+1)}(a_i, h_0) > 0$ and $\hat{f}^{(m+1)}(b_i, h_0) < 0$ for all i . By the continuity of $\hat{f}^{(m+1)}(t, \cdot)$, for all sufficiently small ε and all i , $\hat{f}^{(m+1)}(a_i, h_0 + \varepsilon) > 0$ and $\hat{f}^{(m+1)}(b_i, h_0 + \varepsilon) < 0$. Hence $\liminf_{h \downarrow h_0} v_{m+1}(h) \geq v_{m+1}(h_0)$, the right continuity of v_{m+1} follows from the fact that v_{m+1} is known to be decreasing.

Note that Schoenberg's theorem does not apply for general kernels. Indeed, the convolution of unimodal densities need not be unimodal, see Feller (1966, p. 164).

3. ASSESSING SIGNIFICANCE

For any particular k -modal simple null hypothesis, it is easy to assess, by simulation, the significance of the value of the critical window width obtained from the data. Suppose the null hypothesis is that the true density is g and that the value of h_{crit} obtained from the data is h_0 . Then the theory of Section 2 implies that

$$\text{pr}_g(h_{crit} > h_0) = \text{pr} \{ \hat{f}(\cdot; h_0) \text{ has more than } k \text{ modes} \mid \{X_1, \dots, X_n\} \text{ is drawn from } g \}.$$

Thus, in order to assess the significance of h_0 for sample size n , it is only necessary to calculate the single density estimate $\hat{f}(\cdot; h_0)$ for each sample of size n generated from g ; there is no need to find h_{crit} for each replication.

The hypothesis that the true density is at most k -modal is of course a compound hypothesis. To provide a conservative assessment of the significance of h_0 , an appealing choice of the representative g_0 from which to simulate is obtained by rescaling $\hat{f}(\cdot, h_0)$, as constructed from the data, to have variance equal to the sample variance. The theory of Section 2 shows that g_0 is indeed at most k -modal; it is, in a sense, the most extreme k -modal density consistent with the data. It is extremely easy to simulate from g_0 ; Efron (1979) pointed out that independent observations y_i from g_0 are given by

$$y_i = (1 + h_0^2/\sigma^2)^{-1/2}(X_{I(i)} + h_0 \varepsilon_i),$$

where $X_{I(i)}$ are sampled uniformly, with replacement, from the data X_1, \dots, X_n , σ^2 is the sample variance of the data, and ε_i is an independent sequence of standard normal random variables.

Simulating from g_0 to assess significance is an example of a smoothed bootstrap procedure as defined by Efron (1979). However, Efron's procedure contains an implicit arbitrary choice of smoothing parameter, since his σ_2^2 is essentially arbitrary. In our case, the amount of smoothing is automatically determined in a natural way.

Finally, it should be pointed out that the theory and procedure of finding a critical window width and simulating from a rescaled density estimate constructed using this window width carries over immediately, *mutatis mutandis*, to the investigation of maxima in the first or higher derivative of the data. Both Cox (1966) and Good and Gaskins (1980) show a preference for seeking maxima in the density derivative.

ORIGINAL PAGE IS
OF POOR QUALITY

4. AN APPLICATION

We illustrate the method by analysing a small data set of observations on chondrite meteors. These data consist of 22 observations which are given in Table 2 of Good and Gaskins (1980).

TABLE I
*Chondrite data critical window widths and their
estimated significance levels*

<i>Number of modes</i>	<i>Critical window width</i>	<i>P</i>
1	2.39	0.08
2	1.83	0.05
3	0.68	0.79
4	0.47	0.93

The data have been considered by several authors, see Good and Gaskins (1980) for details. In this analysis the raw values of the observations were used. Table 1 gives critical window widths and significance levels for tests of the null hypothesis that the underlying density has at most k modes against the alternative that it has more than k modes. The p -values are computed by simulating from a critical density as described above; 100 replications of 22 observations were used in each case.

These results must of course be interpreted as a hierarchical set of significance tests. All other things being equal, considerations of parsimony perhaps suggest that we should test successively for an increasing number of modes until we find a number that is accepted. Particularly bearing in mind the small sample size, the results clearly indicate the trimodal nature of the population; Good and Gaskins (1980) also arrived at this conclusion.

OF 1988

5. ASYMPTOTIC BEHAVIOUR OF THE CRITICAL WINDOW WIDTH :
INTRODUCTION

In Section 2 above it was stated heuristically that large values of h_{crit} will tend to reject the null hypothesis. The results of this section show that this procedure does indeed lead to a consistent test.

Subject to certain regularity conditions, it is shown that, under the null hypothesis, h_{crit} converges stochastically to zero, while this is not the case under the alternative hypothesis. The exact rate of convergence of h_{crit} to zero under the null hypothesis is found. It is perhaps interesting that this rate of convergence has precisely the same order as the rate of convergence for the optimum choice of window width for the uniform estimation of the density given, for example, by Silverman (1978b).

In the smoothed bootstrap procedure given in section 3, the representative of the null hypothesis constructed from the data is obtained from the density estimate with window width h_{crit} ; the estimate is rescaled, as suggested by Efron (1979), to have variance equal to the sample variance of the data. The remarks above show that $f_n(., h_{crit})$ is, in a certain sense, optimally uniformly consistent as an estimate of the true density f . It follows that, on the null hypothesis, the bootstrap procedure is likely, at least for large samples, to provide an estimate of the true underlying density which is accurate in the uniform norm. A possible drawback for small samples is the fact that the implied constant in the rate of convergence does not necessarily take its optimum value.

An interesting open question raised by this discussion is the possibility of using $h_{crit}(k)$ for some value of k in developing an automatic method for choosing the smoothing parameter in density estimation. Boneva, Kendall and Stefanov (1971) suggested choosing the window width where 'rabbits' or rapid fluctuations just started to appear. Such a window width would perhaps correspond to $h_{crit}(k)$ for some $k > j$; since $h_{crit}(k)$ converges to zero at the optimum rate for all $k > j$, a suitable formalization of the Boneva-Kendall-Stefanov procedure would give estimates which converged at the optimal rate, though not necessarily with the optimal constant multiplier.

The fact that $h_{\text{crit}}(k)$ has the same rate of convergence for all $k > j$ provides some explanation for the observation made by Boneva, Kendall and Stefanov that the estimate seems suddenly to become noisy as the window width is reduced.

6. ASYMPTOTIC RESULTS

In this section, the main results on the asymptotic behaviour of h_{crit} are stated and proved. It is convenient to use the convention throughout that all limits and implied limits are taken as n tends to infinity. Varying conventions will apply to unqualified suprema and infima in Propositions 1 and 2 below, and these will be introduced where necessary. The notations $p \lim inf$ and $p \lim sup$ will be used to signify the corresponding limits in probability as n tends to infinity, and \underline{O}_p and \overline{O}_p will denote probability orders of magnitude.

Define, for $h > 0$,

$$a(h) = h^{-5} \log(h^{-1}) \quad (1)$$

The main results are all contained in the following theorem.

Theorem

Suppose f is a bounded density with bounded support $[a,b]$, and suppose that the following conditions are satisfied:

- (i) f is twice continuously differentiable on (a,b)
- (ii) f has exactly j local maxima on (a,b)
- (iii) $f'(a+) > 0, f'(b-) < 0$
- (iv) $\min_{\{z: f'(z)=0\}} \frac{f''(z)^2}{f(z)} = c_0 > 0$.

Let $h_{crit}(k)$ be the k -critical window width constructed from an i.i.d.

sample of size n from f . Then, if $k > j$, defining a as in (1) above,

$$p \lim inf n^{-1} a(h_{crit}(k)) > \frac{2}{3} \sqrt{2} c_0 \quad (2)$$

and
$$p \lim sup n^{-1} a(h_{crit}(k)) < \infty \quad (3)$$

while if $k < j$ then there exists a constant $h_0(f,k)$ such that

$$P(h_{crit}(k) > h_0) \rightarrow 1 \quad (4)$$

Note that condition (iv) is equivalent, in the presence of the other conditions, to the condition that f is strictly positive on $[a,b]$ and f' has no multiple zeroes on $[a,b]$.

It is convenient to prove the various assertions of the theorem separately. Except where otherwise stated, the conditions of the theorem on f will be assumed to be true throughout. The first proposition facilitates the proof of (2).

Proposition 1. Given any c_1 with

$$0 < c_1 < \frac{2}{3} \pi \sqrt{2} c_0 ,$$

suppose the sequence of window widths h_n satisfies

$$n^{-1} \alpha(h_n) \rightarrow c_1 . \quad (5)$$

Then the number of maxima of f_n tends in probability to ν .

It follows from Proposition 1 and the ~~Theorem of Gnani~~ that, for all $k > \nu$, provided (5) holds,

$$P\{h_{\text{crit}}(k) < h_n\} \rightarrow 1$$

and hence that (2) is satisfied.

The proof of Proposition 1 makes use of several lemmas, the first of which shows that, under certain conditions, maxima and minima of f_n can, eventually, only occur arbitrarily close to those of f .

Lemma 1. Let I be any closed interval contained in $[a, b]$, such that I contains none of the zeroes of f' . Then, provided $h_n \rightarrow 0$ and $n^{-1} h_n^2 \alpha(h_n) \rightarrow 0$, it will follow that

$$P\{f_n \text{ monotonic on } I \text{ in the same sense as } f\} \rightarrow 1 .$$

Proof. By slight adaptation of the results of Silverman (1978a), it can be seen that, provided f is bounded, we will have, if h_n satisfies the assumptions of Proposition 1,

$$\begin{aligned} \left(\sup_{x \in I} |f'_n - E f'_n| \right) &= O_p \left\{ n^{-\frac{1}{2}} h_n^{-1} \alpha(h_n)^{\frac{1}{2}} \right\} \\ &= o_p(1) . \end{aligned} \quad (6)$$

In Silverman (1978a) the uniform continuity of f was additionally assumed, but careful examination of the proofs of that paper shows that the derivation of the rate of stochastic convergence, though not of the exact constant implied in the O_p , goes through under the assumption of bounded f .

Supposing without loss of generality that f is increasing on I , it follows from the continuity of f' on $[a,b]$ that f' is bounded away from zero on I and is non-negative on a neighborhood of I , and hence by elementary analysis that

$$\liminf_I \inf_n E f'_n > 0 . \quad (7)$$

Combining (6) and (7) completes the proof of Lemma 1.

The next lemma shows that, under suitable conditions, f_n will eventually have exactly one maximum and no minima near each maximum of f , and exactly one minimum and no maxima near each minimum of f .

Lemma 2. Suppose $f'(z) = 0$ and f has a local maximum (respectively minimum) at z . Suppose $h_n \rightarrow 0$ and

$$n^{-1} \alpha(h_n) + c_2 \in (0, \frac{2}{3} \pi \sqrt{2} f''(z)^2 / f(z)) . \quad (8)$$

Then, for all sufficiently small $\epsilon > 0$, the probability that f'_n has exactly one zero in $(z-\epsilon, z+\epsilon)$, and that this zero is a maximum (respectively minimum) of f_n , tends to one as n tends to infinity.

Proof. Only the case of a local maximum will be considered. The proof for a minimum proceeds very similarly and is omitted. Throughout this proof unqualified infima and suprema will be taken to be over x in $[z-\epsilon, z+\epsilon]$.

By the continuity of f and f'' , choose ϵ sufficiently small that

$$\frac{\inf f''(x)^2}{\sup f(x)} > \frac{3c_2}{2\pi\sqrt{2}} \quad (9)$$

and also $[z-\epsilon, z+\epsilon] \subseteq (a,b)$. It is then immediate that $f'(z-\epsilon) > 0$ and $f'(z+\epsilon) < 0$ since, by (9), f'' cannot cross zero in $(z-\epsilon, z+\epsilon)$. Since f' is continuous at $z \pm \epsilon$, by standard results on the consistency of f'_n (a combination of Parzen (1962) and Bhattacharya (1967))

$$P\{f'_n(z-\epsilon) > 0 \text{ and } f'_n(z+\epsilon) < 0\} \rightarrow 1 \quad (10)$$

Very slightly adapting the proofs of Silverman (1976 and 1978a) to cope with the fact that f'' is only uniformly continuous on a neighborhood of $[z-\epsilon, z+\epsilon]$ gives

$$\frac{1}{n} \frac{1}{2} \frac{1}{a(h)^2} \sup |f''_n(x) - Ef''_n(x)| \leq K_1$$

where

$$\begin{aligned} K_1^2 &= 2 \sup f \int \phi^{*2} \\ &= 3(2\pi/2)^{-1} \sup f \end{aligned}$$

Since, by elementary analysis, $\sup |Ef''_n(x) - f''(x)|$ converges to zero, it

follows from (8) that $p \lim_n \sup \sup |f''_n(x) - f(x)| < K_1 c_2^{\frac{1}{2}}$
< $\inf |f''(x)|$

by (9). It is immediate that

$$P\{f''_n(x) < 0 \text{ for all } x \text{ in } [z-\epsilon, z+\epsilon]\} \rightarrow 1 \quad (11)$$

Combining (10) and (11) completes the proof of Lemma 2.

To complete the proof of Proposition 1, note first that no maxima of f_n can occur outside the interval (a,b) . Let z_1, \dots, z_{2j-1} be the zeroes of f' in (a,b) and choose ϵ sufficiently small to satisfy the conclusion of

Lemma 2 for all z_j and to ensure that

$$a < z_1 - \epsilon < z_1 + \epsilon < z_2 - \epsilon < \dots < z_{2j-1} + \epsilon < b . \quad (12)$$

Applying either Lemma 1 or Lemma 2 as appropriate to each of the intervals in the partition (12) of the interval (a,b) completes the proof of Proposition 1.

The next proposition leads to the proof of assertion (3), in a similar way to the derivation of (2) from Proposition 1.

Proposition 2

Defining α as in (1) above, suppose that

$$n^{-1} \alpha(h_n) \rightarrow \infty \quad \text{and} \quad n^{-1} h_n^{-5} \rightarrow 0 . \quad (13)$$

Then the number of maxima in f_n tends in probability to infinity.

Given any k , it follows from this result and the corollary of Silverman (1981) that, provided (13) holds,

$$P\{h_{\text{crit}}(k) > h_n\} \rightarrow 1 ;$$

assertion (2) follows at once.

To prove Proposition 2, suppose without loss of generality that f has a maximum at 0 in (a,b) . Choose a sequence l_n which satisfies

$$l_n \rightarrow 0, \quad h_n^{-1} l_n = o(n^{-1} \alpha(h_n)) , \quad (14)$$

$$h_n^{-1} l_n \rightarrow \infty \quad \text{and} \quad |\log l_n| |\log h_n|^{-1} \rightarrow 1 .$$

The explicit dependence of h and l on n will often be suppressed. Let $I_{j,n}$ be the interval $[(j-1)l, jl]$ for integer $j > 0$.

Following Silverman (1978a) apply Theorem 3 of Komlos, Major and Tusnady (1975) to obtain

$$f'_n(x) = E f'_n(x) + h^{-1} n^{-\frac{1}{2}} \rho_1(x) + \epsilon'_n(x)$$

where ρ_1 is a Gaussian process with the same covariance structure as $\frac{1}{n^2}h(f'_n - Ef'_n)$ and ε'_n is a secondary random error. The process ρ_1 is obtained by putting $\delta(u)$ equal to $\phi'(u)$ in Proposition 1 of Silverman (1978a). By elementary analysis and the arguments of Silverman (1978a) we have, in a neighborhood of 0,

$$|Ef'_n(x) - f'(x)| = \underline{O}(h) ;$$

$$|\varepsilon'_n(x)| = \underline{O}(n^{-1}h^{-2} \log n) \quad \text{a.s.}$$

$$= \underline{O}(h^2) \quad \text{from (13) above ;}$$

$$\text{and } |f'(x)| = \underline{O}(x) ,$$

since $f'(0) = 0$ and f'' exists. It follows that, a.s.,

$$\begin{aligned} \sup |Ef'_n(x) + \varepsilon'_n(x)| &= \underline{O}(j\ell) + \underline{O}(h) \\ &= \underline{O}(n^{-1}h^{-5} \log(\ell/h))^{\frac{1}{2}} \end{aligned} \quad (15)$$

by (13) and (14) above, where we adopt the convention, here and subsequently in this proof, that unqualified suprema are taken to be over the interval $I_{j,n}$, and that a fixed j is being considered.

We slightly adapt the argument of Silverman (1976) pp. 138-140 to investigate $\sup \rho_1$. Define

$$\begin{aligned} \sigma^2(x) &= \text{var } \rho_1(x) = h^{-1}f(x) \int \phi'^2(1 + \underline{O}(1)) \\ &= h^{-1}f(0) \int \phi'^2(1 + \underline{O}(1)) \quad \text{for } x \text{ in } I_{j,n} , \end{aligned}$$

since the end points of $I_{j,n}$ both converge to zero. Analogously to (12) of Silverman (1976), given any λ in $(0,2)$,

$$\begin{aligned}
& P\left\{\sup \sigma^{-1} \rho_1 < \left(1 - \frac{1}{2} \lambda\right) \left(2 \log(h^{-1} \ell)\right)^{\frac{1}{2}}\right\} \\
& < \underline{O}(\ell^{-2}) \log(h^{-1} \ell) \\
& \quad \times \iint_{I_{j,n}} |\chi| \exp\left[2 \log(h^{-1} \ell) \left(1 - \frac{1}{2} \lambda\right)^2 |\chi| / (1 + |\chi|)\right]
\end{aligned} \tag{16}$$

where $\chi(x, y) = \text{corr}(\rho(x), \rho(y))$. Using a similar argument to that following (12) of Silverman (1976), but allowing the interval I to vary, shows that the expression in (16) is dominated by

$$\begin{aligned}
& \underline{O}(\ell^{-2}) \log(h^{-1} \ell) \left\{\sigma^2(0) + \underline{O}(1)\right\}^{-1} (h^{-1} \ell)^{\left(1 - \frac{1}{2} \lambda\right)^2} \underline{O}(\ell) \\
& = (h^{-1} \ell)^{-\lambda + \frac{1}{4} \lambda^2} \log(h^{-1} \ell) \rightarrow 0
\end{aligned}$$

by (14) above.

It follows that, setting $K = \left\{2f(0) \int \phi^2\right\}^{\frac{1}{2}}$,

$$P \lim \inf \sup \left\{h^{-1} \log(h^{-1} \ell)\right\}^2 \rho_1 > K \tag{17}$$

and that the same result holds if ρ_1 is replaced by $-\rho_1$, giving a corresponding result for $\inf \rho_1$. It follows from (15), (17) and the corresponding result for $\inf \rho_1$ that

$$P\left\{\rho_1 \text{ crosses } -\frac{1}{2} h(Lf'_n + c'_n) \text{ in } I_{j,n}\right\} \rightarrow 1,$$

and hence that

$$P\left\{f'_n \text{ crosses zero in } I_{j,n}\right\} \rightarrow 1. \tag{18}$$

Since (18) holds for all j , the number of maxima in f_n tends in probability to infinity, completing the proof of Proposition 2.

ORIGINAL PAGE IS
OF POOR QUALITY

The final proposition of this section deals with the case where the alternative hypothesis is true, and shows that h_{crit} will remain bounded away from zero.

Proposition 3

If $k < j$ then there exists a constant $h_0 > 0$, depending on f and k , such that

$$P(h_{crit}(k) > h_0) \rightarrow 1.$$

Proof

By arguments analogous to those of the proof of the theorem of *Section 2 above*, making use of the variation diminishing properties of the Gaussian kernel and the continuity properties of Ef_n , the number of maxima in $Ef_n(\cdot, h)$ is a right continuous decreasing function of h , for $h > 0$. By choosing h_0 sufficiently small, we can ensure that $Ef_n(\cdot, h_0)$ has independently of n , exactly j maxima. Because of the conditions imposed on f in the statement of the Theorem above, we can also ensure that $Ef_n''(\cdot, h_0)$ is non-zero at all stationary points of $Ef_n(\cdot, h_0)$.

The argument of Lemma 2.2 of Schuster (1969), which does not in fact require the convergence to zero of the sequence of window widths, then implies that, with probability one,

$$f_n'(x, h_0) - Ef_n'(x, h_0) \quad \text{and} \quad f_n''(x, h_0) - Ef_n''(x, h_0)$$

both converge to zero uniformly over x . By an argument similar to that used in Proposition 1 above, it follows that the number of maxima of $f_n(\cdot, h_0)$ on $[a, b]$ tends almost surely to j , the number of maxima of $Ef_n(\cdot, h_0)$.

Applying the corollary of *Section 2* completes the proof of Proposition 3.

ORIGINAL PAGE IS
OF POOR QUALITY

Discussion

It is natural to enquire to what extent the conditions of the theorem above can be relaxed without affecting the conclusions. In particular it seems intuitively clear that the condition of bounded support for the density f should be able to be replaced by some condition on the tails of f , though the present method of proof cannot deal with this case. Condition (iv) appears to be more fundamental to the result; if, for example, $f'(0) = f''(0) = 0 \neq f'''(0)$, then an examination of f_n and Ef_n near zero seems to indicate that, under suitable regularity conditions, there will be no maximum of f_n near zero provided $|f_n''' - Ef_n'''|$ remains small. A heuristic argument suggests that a result corresponding to the theorem of Section 2 can be proved, but with $\alpha(h)$ replaced by $h^{-7} \log(h^{-1})$, so that h_{crit} converges to zero more slowly. Even slower convergence will occur for higher order zeroes in f' .

The interest in this discussion lies in the fact that the bootstrap density constructed using the critical window width will not only have infinite tails of similar weight to those of the corresponding normal kernels but will also have a stationary point which is a point of inflexion. The slower convergence to zero of h_{crit} provides support for the remark in Section 3 that the bootstrap test may be conservative; it also bears out the intuition of P. Huber (private communication) that the bootstrap procedure may be excessively conservative, though the difference between $\frac{1}{5}$ and $\frac{1}{7}$ convergence is very slight in practice.

The methods of this paper can also be used to study the asymptotic properties of a corresponding test for the number of points of inflexion in the density. Both Cox (1960) and Good and Gaskins (1980) prefer to use points of inflexion as an indication that the density is a mixture. The critical

ORIGINAL PAGE IS
OF POOR QUALITY

window width will now be the smallest window width for which the density has k maxima. Under suitable conditions a result corresponding to the theorem of Section 2 can be proved, but again, among other changes, $\alpha(h)$ will be replaced by $h^{-7} \log(1/h)$ since f_n'' will be replaced by f_n''' in much of the argument of the proofs of Propositions 1 and 2.

REFERENCES

- Andrews, H.C. (1972), Introduction to mathematical techniques in pattern recognition. Wiley, New York.
- Bhattacharya, P.K. (1967), Estimation of a probability density function and its derivatives. *Sankhyā*, Series A, 29, 373-382.
- Böck, H.H. (1977), On tests concerning the existence of a classification. Proc. First International Symposium on Data Analysis and Informatics, Versailles, 1977, Institut de Recherche d'Informatique et d'Automatique, Domaine de Voulceau, Le Chesnay, France, 449-464.
- Boneva, L.I., Kendall, D.G. and Stefanov, I. (1971), Spline Transformations. *J. Roy. Statist. Soc. B*, 33, 1-70.
- Cox, D.R. (1966), Notes on the analysis of mixed frequency distributions. *Brit. J. Math. Statist. Psych.*, 19, 39-47.
- Efron, B. (1973), Bootstrap methods - another look at the jack-knife. *Ann. Statist.*, 7, 1-26.
- Feller, W. (1966). An Introduction to Probability Theory and its Applications, Volume 11. New York: Wiley.
- Fukunaga, K. and Hostetler, L.D. (1975). The estimation of the gradient of a density function with applications in pattern recognition. *IEEE Trans. Inform. Theory*, IT-21, 32-40.
- Good, I.J. and Gaskins, R.A. (1980). Density estimation and bump-hunting by the penalized likelihood method exemplified by scattering and meteorite data. *J. Amer. Stat. Assoc.*, 75, 42-56.
- Karlin, S. (1968). Total Positivity. Stanford: Stanford University Press.
- Komlos, J., Major, P. and Tusnady, G. (1975). An approximation of partial sums of independent random variables and the sample distribution function. *Z. Wahrscheinlichkeitstheorie und Verw. Gebiete*. 32, 111-131.
- Parzen, E. (1962). On estimation of a probability density function and mode. *Ann. Math. Statist.*, 33, 1065-1076.

- Rosenblatt, M. (1956). Remarks on some non-parametric estimates of a density function. *Ann. Math. Statist.* 27, 832-837.
- Schoenberg, I.J. (1950). On Polya frequency functions. II: Variation diminishing integral operators of the convolution type. *Acta Scientiarum Mathematicarum Szeged*, 12B. 97-106.
- Silverman, B.W. (1976). On a Gaussian process related to multivariate probability density estimation. *Math. Proc. Cambridge Philos. Soc.*, 80.
- Silverman, B.W. (1978a). Weak and strong uniform consistency of the kernel estimate of a density and its derivatives. *Ann. Statist.*, 6, 177-184.
- Silverman, B.W. (1978b). Choosing the window width when estimating a density. *Biometrika*, 65, 1-11.
- Silverman, B.W. (1978c). Density ratios, empirical likelihood and co \ddot{c} death. *Appl. Statist.*, 27, 26-33.
- Silverman, B.W. (1980) Comment on Good and Gaskins (1980). *J. Amer. Statist. Ass.*, 75, 67-68.
- Silverman, B.W. (1981a) Density estimation for univariate and bivariate data. In *Interpreting Multivariate Data* (V. Barnett, ed.) Chichester: Wiley.
- Silverman, B.W. (1981b). Using kernel density estimates to investigate multimodality. *J. Roy. Statist. Soc. B*, 43, 97-99.
- Silverman, B.W. (1983). Some properties of a test for multimodality based on kernel density estimates. To appear in a collection in honour of David Kendall, edited by J.F.C. Kingman and G.E.H. Reuter, pub. Cambridge University Press, Cambridge, England.

N83

15785

UNCLAS

A DATA BASED RANDOM NUMBER GENERATOR FOR A MULTIVARIATE DISTRIBUTION*
(USING STOCHASTIC INTERPOLATION)

James R. Thompson
Malcolm S. Taylor

Rice University
USA Ballistic Research Laboratory

ABSTRACT. Let X be a k -dimensional random variable serving as input for a system with output Y (not necessarily of dimension k). Given X , an outcome Y or a distribution of outcomes $G(Y|X)$ may be obtained either explicitly or implicitly. We consider here the situation in which we have a real world data set $\{X_j\}_{j=1}^n$ and a means of simulating an outcome Y . A method for empirical random number generation based on the sample of observations of the random variable X without estimating the underlying density is discussed.

INTRODUCTION. The manner of dealing with multivariate data depends upon the application at hand. For example, let us suppose that $\{X_j\}_{j=1}^n$ is a sample of size n of a k -dimensional random variable. We may be interested simply in estimating the mean μ . In such a case, we may complete our task by computing the sample mean \bar{X} . If we are interested in the interrelationships between the various vector components, we may find it desirable to compute the sample covariance matrix $\hat{\sigma}$.

At a greater level of complexity, we may be required to estimate the density of X nonparametrically [1,3]. Here, the representational difficulties are substantial--- particularly for $k > 2$, where our 3-dimensional intuitions are inadequate for graphing the density even if we knew it precisely on a discrete mesh. Indeed, it would appear that for increasing dimensionality, our estimation theoretic difficulties pale in comparison to those of representation.

* This research was supported in part by ARO Contract DAAG-29-82-K-0014 at Rice University. To appear in Proceedings of the Twenty-Seventh Conference on the Design of Experiments in Army Research Development and Testing.

Suppose we are given, for example, the task of estimating the density f at a point X_0 in k -space, based on a sample of size n . The naive nearest neighbor estimator

$$\hat{f}(X_0) = \frac{p}{n \cdot V_k(X_0, d(X_0, p))}$$

where $d(X_0, p)$ is the Euclidean distance from X_0 to the p^{th} nearest neighbor and $V_k(X_0, d(X_0, p))$ is the volume of the k -sphere centered at X_0 with radius $d(X_0, p)$, is likely to be quite satisfactory. But a problem occurs when we are asked for a usable summary of the unknown density over the space of non-negligible mass. If we know the functional form of the density $f(X; \theta)$, then we have a relatively easy task--- the estimation of θ . But in the highly ubiquitous nonparametric situation, in which we do not know the functional form of f , we are not so fortunate. We might decide, for example, to tabulate \hat{f} on a mesh of size 20 in each dimension. This would require 20^k pointwise estimations of f --- a tedious but manageable task. But how shall we scan this k -dimensional table to obtain a useful feel for the density? Other approaches, clearly are required. One of these is discussed in [2].

There are, happily, cases in which the density representational difficulties may be sidestepped when coping nonparametrically with data sets in higher dimensions. For example, let us suppose the k -dimensional random variable X is an input into a system with output Y (of whatever dimension). Given X , an outcome Y or a distribution of outcomes $G(Y|X)$ is obtained either explicitly or implicitly through an output data set. Let us suppose these outcomes fall into six categories: Very Good, Good, Fair, Poor, Very Bad, Catastrophically Bad. Suppose further that these sets are well-defined

in the Y-space. We are given a real world data set $\{X_j\}_{j=1}^n$. We have a means of simulating an outcome Y given the input X. We wish to determine the probability of arriving in each of the six category sets.

One way to achieve this result might be, simply, to sample from the n data points $\{X_j\}_{j=1}^n$. In many cases this will prove quite satisfactory. But let us suppose that "Catastrophically Bad" happens for $Y > 10$,

$$\text{where } Y = 1 / \sum_{i=1}^4 x_i^2 \quad \text{with } X = (x_1, x_2, x_3, x_4).$$

Then, if the x_i 's are (unbeknownst to us, but in actuality) independently distributed as $N(0,1)$, the chance of a "Catastrophically Bad" event is .0012. Let us suppose the size (n) of our data set is 100. The chance of none of these observations being in the "Catastrophically Bad" region is .887. So, a simulation which used only the 100 data points would, with probability .887, give us the information that "Catastrophically Bad" occurred with zero probability. We need to avoid this pitfall.

One procedure would be to estimate the density of X nonparametrically and then build a random number generator using the density. Such a scheme would run into the representational difficulties mentioned above. We can be much more efficient.

THE ALGORITHM. Let us consider the following situation: We have a random sample $\{X_j\}_{j=1}^n$ of size n from a multivariate distribution of dimension k, and we want to generate pseudorandom vectors from the underlying, but unknown, distribution that gave rise to the random sample. Since we do not know, and usually will never know, the form of this distribution, our attack

should be empirical. We shall endeavor to see to it that our pseudorandom vectors look very much like those in the original data set. In so doing, we will maintain the essential structural integrity of the problem.

We now direct our attention to the mechanics of the algorithm. After carrying out a rough rescaling to account for differing variances that may exist among the k variates, we select at random one of the n data points, say X_1 , from the data base and then proceed to determine its $m-1$ nearest neighbors. The nearest neighbors are determined under the ordinary Euclidean metric and the value of m will depend upon the sample size n , the characteristics of the data, and can best be determined after perusal of the data. A conservative estimate would be to choose $m = n/20$.

The vectors $\{X_j\}_{j=1}^m$ are now coded about the sample mean $\bar{X} = 1/m \sum X_i$ to yield $\{X'_j\} = \{X_j - \bar{X}\}_{j=1}^m$, and an independent random sample of size m is generated from the uniform distribution $U(1/m - \sqrt{\frac{3(m-1)}{m^2}}, 1/m + \sqrt{\frac{3(m-1)}{m^2}})$.

Now the linear combination

$$X' = \sum_{\ell=1}^m u_{\ell} X'_{\ell}$$

is formed, where $\{u_{\ell}\}_{\ell=1}^m$ is the random sample from the $U(1/m - \sqrt{\cdot}, 1/m + \sqrt{\cdot})$.

Finally the translation

$$X = X' + \bar{X}$$

restores the relative magnitude, and X is a pseudorandom vector which we propose to be representative of the multivariate distribution that provided the $\{X_j\}_{j=1}^n$.

To obtain the next pseudorandom vector we randomly select another of the n data points and proceed as above.

We will now attempt to motivate the algorithm by considering the mathematics that suggests the mechanics that we have just outlined. Consider the distribution of X_1 and its $m-1$ nearest neighbors:

$\{(x_{1\ell}, x_{2\ell}, \dots, x_{k\ell})\}_{\ell=1}^m = \{X_\ell\}_{\ell=1}^m$. Let us suppose that this "truncated set" of random observations has mean vector μ and covariance matrix σ . Let $\{u_\ell\}_{\ell=1}^m$ be an independent random sample from the uniform distribution $U(1/m - \sqrt{\cdot}, 1/m + \sqrt{\cdot})$. Then, $E(u_\ell) = 1/m$, $\text{Var}(u_\ell) = (m-1)/m^2$, and $\text{Cov}(u_i, u_j) = 0$, for $i \neq j$.

Forming the linear combination

$$Z = \sum_{\ell=1}^m u_\ell X_\ell$$

we have, for the r^{th} component $z_r = u_1 x_{r1} + u_2 x_{r2} + \dots + u_m x_{rm}$, the following relations

$$E(z_r) = m \cdot 1/m \cdot \mu_r = \mu_r,$$

$$\text{Var}(z_r) = \sigma_r^2 + (m-1)/m \cdot \mu_r^2,$$

$$\text{Cov}(z_r, z_s) = \sigma_{rs} + (m-1)/m \cdot \mu_r \mu_s.$$

Clearly, if the mean vector of X was $\mu = (0, 0, \dots, 0)^t$, then the mean vector and covariance matrix of Z would be identical to those of X . In the less idealized situation with which we are confronted, the translation to the sample mean of the nearest neighbor cloud should result in the pseudoobservation having very nearly the same mean and covariance structure as that of the

(truncated) distribution of the points in the nearest neighbor cloud, a conjecture borne out in many actual cases that have been considered. For m moderately large, our algorithm essentially samples from n Gaussian distributions with the means and covariance matrices corresponding to those of the n m nearest neighbor clouds.

EXAMPLES. For a substantial test case, we considered a mixture of three bivariate normal distributions. The first (N_1) has mean vector $\begin{pmatrix} -1 \\ -2 \end{pmatrix}$ and covariance matrix $\begin{pmatrix} 1 & -1/2 \\ -1/2 & 1 \end{pmatrix}$; the second (N_2) has mean vector $\begin{pmatrix} -2 \\ 3 \end{pmatrix}$ and covariance matrix $\begin{pmatrix} 1 & 1/2 \\ 1/2 & 1 \end{pmatrix}$; and the third (N_3) has mean vector $\begin{pmatrix} 2 \\ 3/2 \end{pmatrix}$ and covariance matrix $\begin{pmatrix} 1 & 1/10 \\ 1/10 & 1 \end{pmatrix}$. The corresponding mixing scalars are $\alpha_1 = 1/2$, $\alpha_2 = 1/3$, and $\alpha_3 = 1/6$, respectively. Representative contours of equal density are illustrated in Figure 1. To establish a data base, a sample of eighty-five points was generated from this distribution via Monte Carlo simulation; a sample of eighty-five pseudorandom values was then produced by the algorithm, and the combined sample is shown in Figure 2.

Notice that the structure of the data is maintained in that the modes are preserved; the algorithm has not attempted to fill in gaps where gaps belong; the algorithm has, however, generated some points outside the boundary of the convex hull of the data base, all of which are desirable properties. These observations lend credence to the term "structural integrity" mentioned previously.

An application of the algorithm to a real world data set is summarized in Figures 3 and 4. In Figure 3, a two-dimensional marginal of a set of 973 four-dimensional behind armor debris measurements is portrayed; in Figure 4, 973 simulated data points produced by our procedure. Once again, the salient features of the data set are preserved.

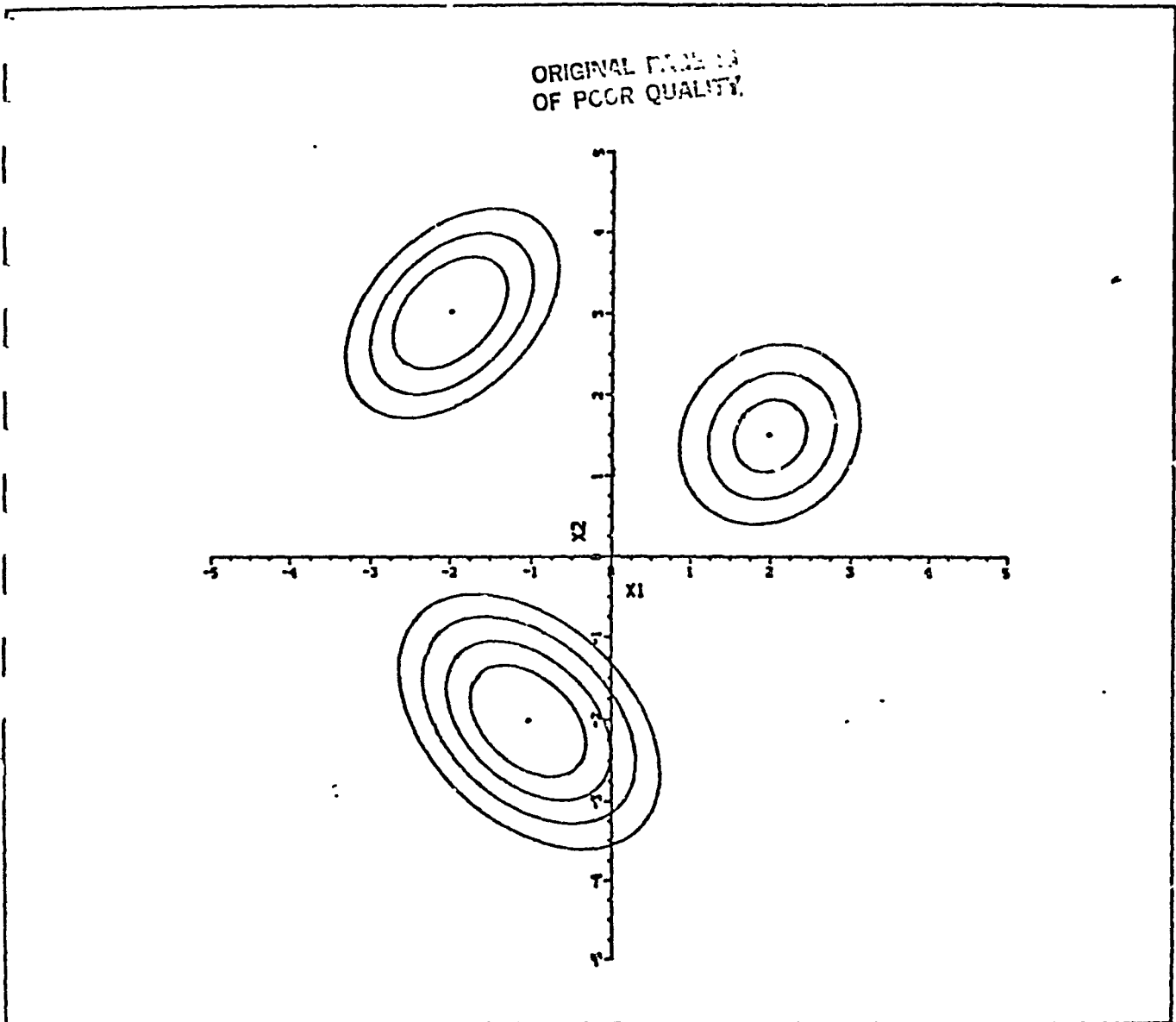


Fig. 1. Density contours for a mixture of three bivariate normal distributions.

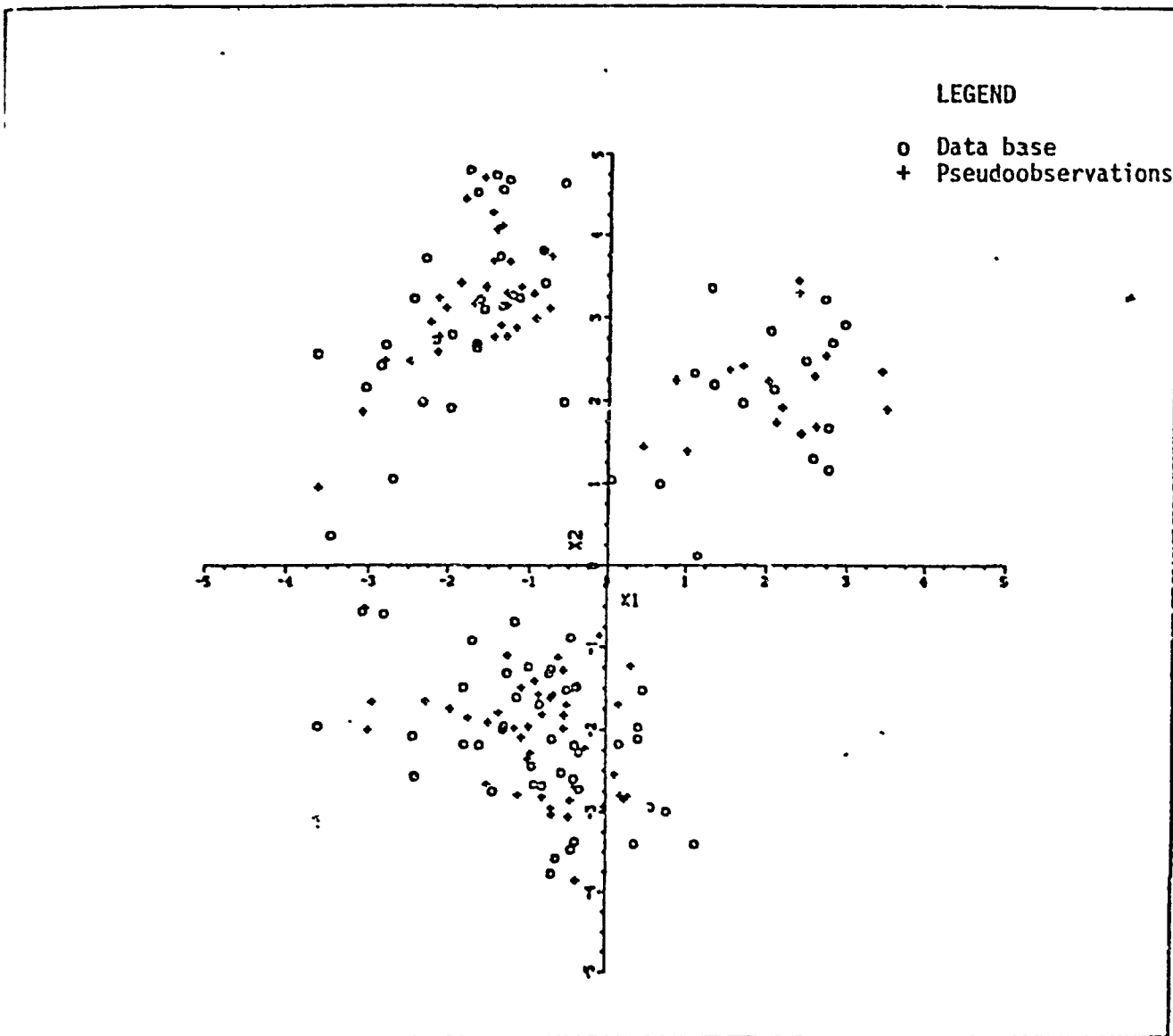


Fig. 2. Combined sample: Data base and Pseudoobservations.

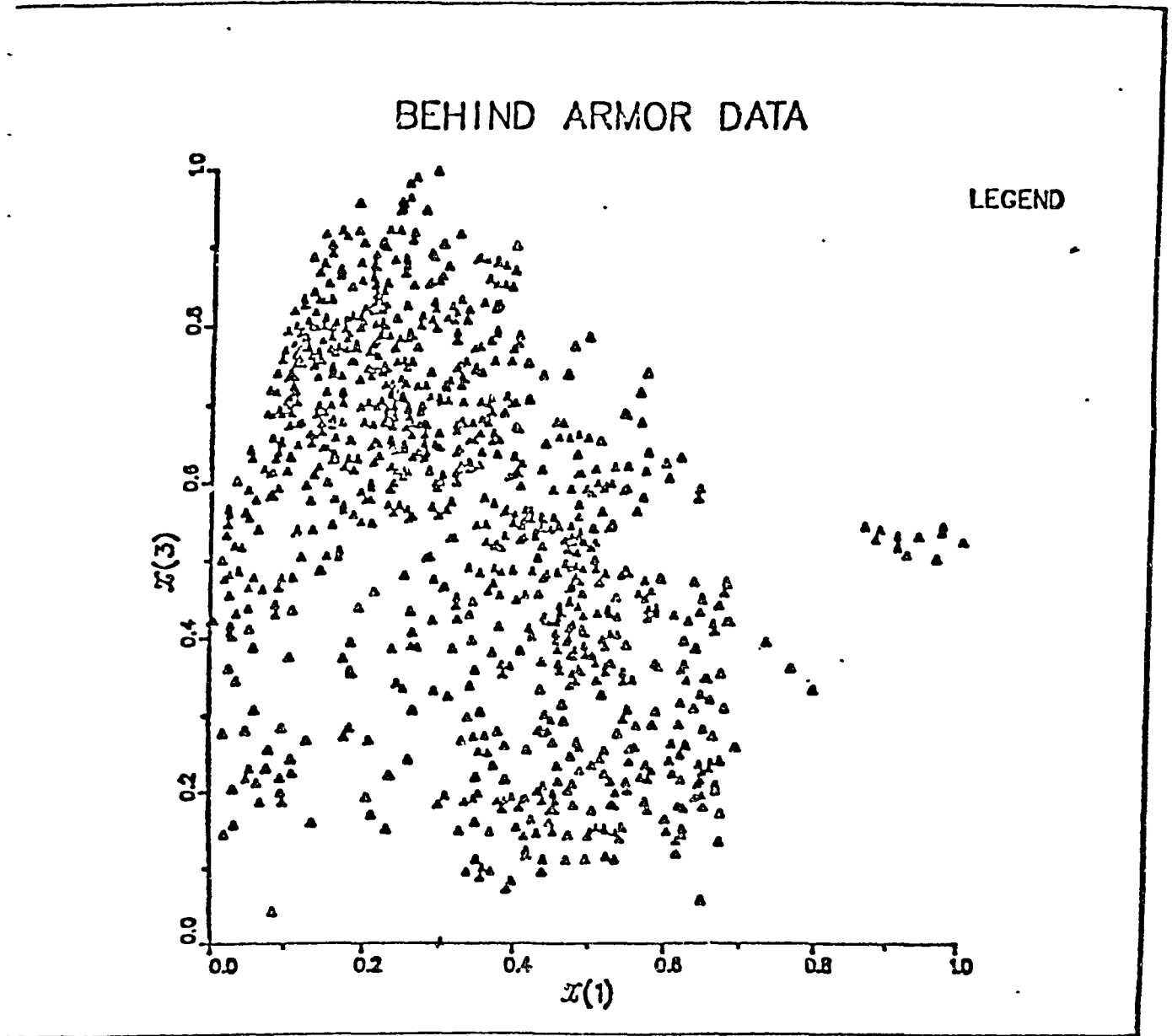


Fig. 3. Marginal data for 4-dimensional behind armor debris.

ORIGINAL PAGE IS
OF POOR QUALITY

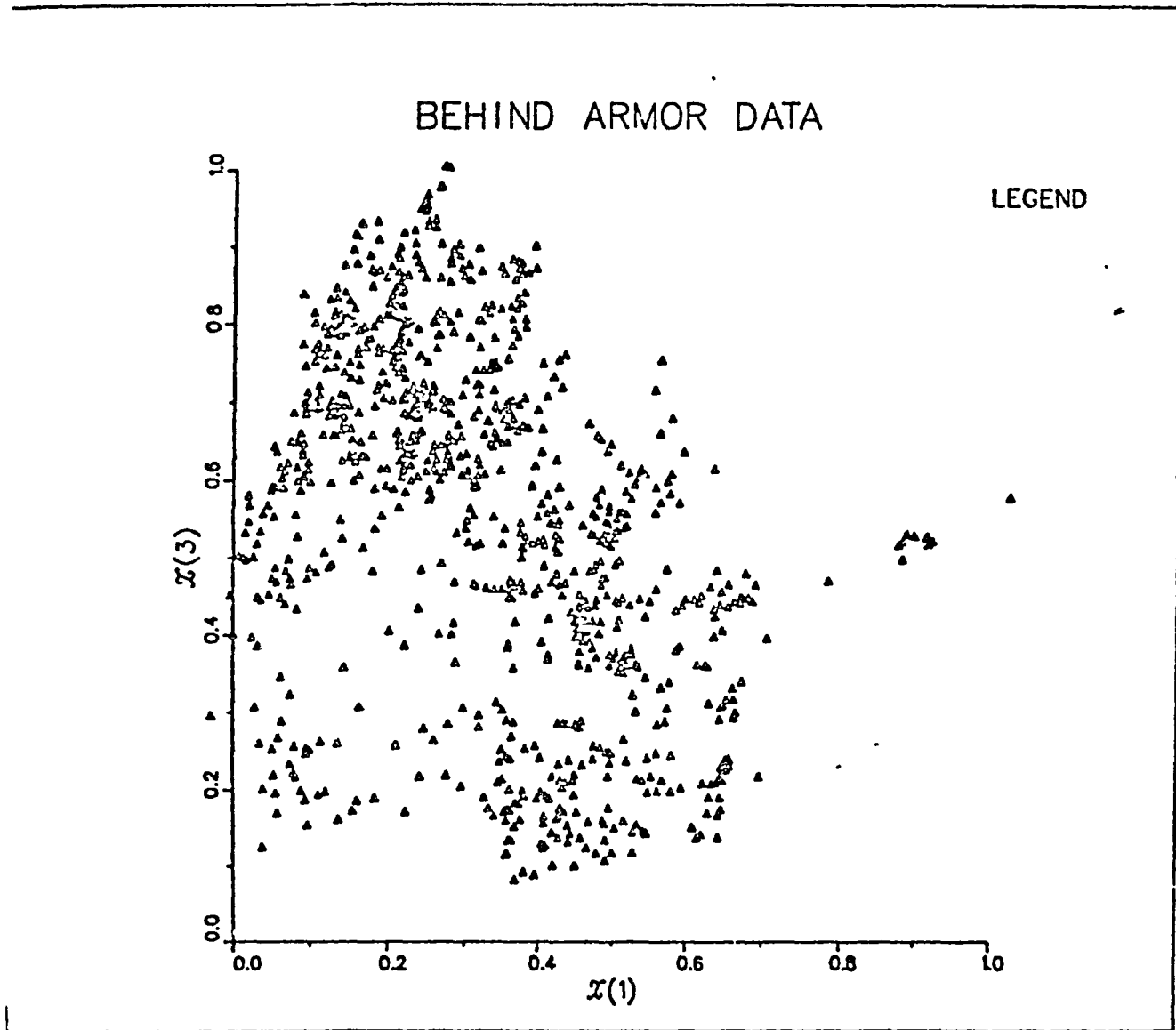


Fig. 4. Simulated behind armor debris.

CONCLUSIONS. We have demonstrated a means of empirical random number generation based on a sample of observations of a random variable X . No estimation of the underlying density is required. And, because of the local nature of the generation scheme, it is essentially free of assumptions on the underlying density of X . Naturally, any attempt to use this algorithm for generating bona fide new observations using the computer rather than producing real world data would be unwise. Rather, the algorithm operates somewhat like a smooth interpolator--- highly dependent on the quality of the data points on which it is based. It gives us a means of avoiding nonrobust conclusions due to "holes" in the data set at important points of the simulation model.

Also included in Thompson's presentation was a discussion of how alternatives to the usual (contour map) density estimators may be constructed based on stochastic interpolation.

ORIGINAL PAGE IS
OF POOR QUALITY

REFERENCES

1. Bean, Steven J. and Tsokos, Chris P. (1980). "Developments in nonparametric density estimation," International Statistical Review, v. 48, pp. 267-287.
2. Fwu, Chih-chy, Tapia, Richard A. and Thompson, James R. (1980). "The nonparametric estimation of probability densities in ballistics research," to appear in the Proceedings of the Twenty-Sixth Conference on the Design of Experiments in Army Research, Development and Testing.
3. Scott, David W., Tapia, Richard A. and Thompson, James R. (1978), "Multivariate density estimation by discrete maximum penalized likelihood methods," Graphical Representations of Multivariate Data, Peter C. Wang, ed., Academic Press, pp. 169-182.

N83

157886

UNCLAS

MIXTURE DENSITIES, MAXIMUM LIKELIHOOD, AND THE EM ALGORITHM

by

Richard A. Redner
Department of Mathematical Sciences
University of Tulsa
Tulsa, Oklahoma 74104

and

Homer F. Walker
Department of Mathematics
University of Houston
Houston, Texas 77004

Abstract: The problem of estimating the parameters which determine a mixture density has been the subject of a large, diverse body of literature spanning nearly ninety years. During the last two decades, the method of maximum-likelihood has become the most widely followed approach to this problem, thanks primarily to the advent of high-speed electronic computers. Here, we first offer a brief survey of the literature directed toward this problem and review maximum-likelihood estimation for it. We then turn to the subject of ultimate interest, which is a particular iterative procedure for numerically approximating maximum-likelihood estimates for mixture density problems. This procedure, known as the EM algorithm, is a specialization to the mixture density context of a general algorithm of the same name used to approximate maximum-likelihood estimates for incomplete data problems. We discuss the formulation and theoretical and practical properties of the EM algorithm for mixture densities, focussing in particular on mixtures of densities from exponential families.

Key words and phrases: Mixture densities, maximum-likelihood, EM algorithm, exponential families, incomplete data.

MIXTURE DENSITIES, MAXIMUM LIKELIHOOD, AND THE EM ALGORITHM

By

Richard A. Redner

Department of Mathematical Sciences

University of Tulsa

Tulsa, Oklahoma 74104

and

Homer F. Walker¹

Department of Mathematics

University of Houston

Houston, Texas 77004

1. Introduction

Of interest here is a parametric family of finite mixture densities, i.e., a family of probability density functions of the form

$$p(x|\Phi) = \sum_{i=1}^m \alpha_i p_i(x|\rho_i), \quad x = (x_1, \dots, x_n)^T \in R^n, \quad (1.1)$$

where each α_i is nonnegative and $\sum_{i=1}^m \alpha_i = 1$, and where each

p_i is itself a density function parametrized by $\rho_i \in \Omega_i \subseteq R^{n_i}$.

We denote $\Phi = (\alpha_1, \dots, \alpha_m, \rho_1, \dots, \rho_m)$ and set

1. The work of this author was supported by the U.S. Department of Energy under grant DE-AS05-76ER05046.

ORIGINAL PAGE IS
OF POOR QUALITY

$$\Omega = \{(\alpha_1, \dots, \alpha_m, \rho_1, \dots, \rho_m) : \sum_{i=1}^m \alpha_i = 1 \text{ and } \alpha_i > 0, \rho_i \in \Omega_i \\ \text{for } i = 1, \dots, m\} .$$

The more general case of a possibly infinite mixture density, expressible as

$$\int_{\Lambda} p(x|\Phi(\lambda))d\alpha(\lambda) , \quad (1.2)$$

is not considered here, even though much of the following is applicable with few modifications to such a density. For general references dealing with infinite mixture densities and related densities not considered here, see the survey of Blischke [12]. Also, it is understood that in determining probabilities, probability density functions are integrated with respect to a measure on R^n which is either Lebesgue measure, counting measure on some finite or countably infinite subset of R^n , or a combination of the two. In the following, it is usually obvious from the context which measure on R^n is appropriate for a particular probability density function, and so measures on R^n are not specified unless there is a possibility of confusion. It is further understood that the topology on Ω is the natural product topology induced by the topology on the real numbers. At times when it is convenient to determine this topology by a norm, we will regard elements of Ω as $(m + \sum_{i=1}^m n_i)$ - vectors and consider norms defined on such vectors.

Finite mixture densities arise naturally - and can naturally be interpreted - as densities associated with a statistical population which is a mixture of m component populations with associated component densities $\{p_i\}_{i=1, \dots, m}$ and mixing proportions $\{\alpha_i\}_{i=1, \dots, m}$. Such densities appear as fundamental models in areas of applied statistics such as statistical pattern recognition, classification, and clustering. (As examples of general references in the broad literature on these subjects, we mention Duda and Hart [44], Fukunaga [48], Hartigan [62], Van Ryzin [128], and Young and Calvert [138]. For some specific applications, see, for example, the Special Issue on Remote Sensing of the Communications in Statistics [32]). In addition, finite mixture densities often are of interest in life testing and acceptance testing (cf. Cox [34], Hald [60], Mendenhall and Hader [89], and other authors referred to by Blischke [12]). Finally, many scientific investigations involving statistical modeling require by their very nature the consideration of mixture populations and their associated mixture densities. The example of Hosmer [68] below is simple but typical. For references to other examples in Fishery studies, genetics, medicine, chemistry, psychology, and other fields, see Blischke [12], Everitt and Hand [45], and Hosmer [67].

Example: According to the International Halibut Commission of Seattle, Washington, the length distribution of Halibut of a given age is closely approximated by a mixture of two normal distributions corresponding to the length distributions of the

male and female subpopulations. Thus the length distribution is modeled by a mixture density of the form

$$p(x|\Phi) = \alpha_1 p_1(x|\rho_1) + \alpha_2 p_2(x|\rho_2) , x \in R^1 , \quad (1.3)$$

where for $i = 1, 2$,

$$p_i(x|\rho_i) = \frac{1}{\sqrt{2\pi}\sigma_i} e^{-\frac{(x-\mu_i)^2}{2\sigma_i^2}} , \rho_i = (\mu_i, \sigma_i^2)^T \in R^2 , \quad (1.4)$$

and $\Phi = (\alpha_1, \alpha_2, \rho_1, \rho_2)$. Suppose that one would like to estimate Φ on the basis of some sample of length measurements of halibut of a given age. If one had a large sample of measurements which were labeled according to sex, then it would be an easy and straightforward matter to obtain a satisfactory estimate of Φ . Unfortunately, it is reported in [68] that the sex of halibut cannot be easily (i.e., cheaply) determined by humans; therefore, as a practical matter, it is likely to be necessary to estimate Φ from a sample in which the majority of members are not labeled according to sex.

Regarding p in (1.1) as modeling a mixture population, we say that a sample observation on the mixture is labeled if its component population of origin is known with certainty; otherwise, we say that it is unlabeled. The example above illustrates the central problem with which we are concerned here, namely that of estimating Φ in (1.1) using a sample in which some or all of the observations are unlabeled. This problem is

referred to in the following as the mixture density estimation problem. (For simplicity, we do not consider here the problem of estimating not only Φ but also the number m of component populations in the mixture.) A variety of cases of this problem and several approaches to its solution have been the subject of or at least touched on by a large, diverse set of papers spanning nearly ninety years. We begin by offering in the next section a cohesive but very sketchy review of those papers of which we are aware which have as their main thrust some aspect of this problem and its solution. It is hoped that this survey will provide both some perspective in which to view the remainder of this paper and a starting point for those who wish to explore the literature associated with this problem in greater depth.

Following the review in the next section, we discuss at some length the method of maximum-likelihood for the mixture density estimation problem. In rough general terms, a maximum-likelihood estimate of a parameter which determines a density function is a choice of the parameter which maximizes the induced density function (called in this context the likelihood function) of a given sample of observations. Maximum-likelihood estimation has been the approach to the mixture density estimation problem most widely considered in the literature since the use of high speed electronic computers became widespread in the 1960's. In Section 3, the maximum-likelihood estimates of interest here are defined precisely, and both their important theoretical properties and aspects of their practical behavior are summarized.

The remainder of the paper is devoted to the subject of ultimate interest here, which is a particular iterative procedure for numerically approximating maximum-likelihood estimates of the parameters in mixture densities. This procedure is a specialization to the mixture density estimation problem of a general method for approximating maximum-likelihood estimates in an incomplete data context which was formalized by Dempster, Laird, and Rubin [38] and termed by them the EM algorithm (E for "expectation" and M for "maximization"). The EM algorithm for the mixture density estimation problem has been studied by many authors over the last two decades. In fact, there have been a number of independent derivations of the algorithm from at least two quite distinct points of view. It has been found in most instances to have the advantages of reliable global convergence, low cost per iteration, economy of storage, and ease of programming as well as a certain heuristic appeal. On the other hand, it can also exhibit hopelessly slow convergence in some seemingly innocuous applications. All in all, it is undeniably of considerable current interest, and it seems likely to play an important role in the mixture density estimation problem for some time to come.

We feel that the point of view toward the EM algorithm for mixture densities advanced in [38] greatly facilitates both the formulation of a general procedure for prescribing the algorithm and the understanding of the important theoretical properties of the algorithm. Our objectives in the following are to present

this point of view in detail in the mixture density context, to unify and extend the diverse results in the literature concerning the derivation and theoretical properties of the EM algorithm, and to review and add to what is known about its practical behavior.

In Section 4, we interpret the mixture density estimation problem as an incomplete data problem, formulate the general EM algorithm for mixture densities from this point of view, and discuss the general properties of the algorithm. In Section 5, the focus is narrowed to mixtures of densities from the exponential family, and we summarize and augment the results of investigations of the EM algorithm for such mixtures which have appeared in the literature. Finally, in Section 6, we discuss the performance of the algorithm in practice through qualitative comparisons with other algorithms and numerical studies in simple but important cases.

2. A Review of the Literature

The following is a skeletal survey of papers which are primarily directed toward some part of the mixture density estimation problem. No attempt has been made to include papers which are strictly concerned with applications of estimation procedures and results developed elsewhere. For additional references relating to mixture densities as well as more detailed summaries of the contents of many of the papers touched on below, we refer the reader to the recently published monograph by Everitt and Hand [45]. As a convenience, this survey has been divided somewhat arbitrarily by topics into four subsections. Not surprisingly, many papers are cited in more than one subsection.

2.1 The method of moments.

The first published investigation relating to the mixture density estimation problem appears to be that of Pearson [97]. In that paper, as in Example 1.1, the problem considered is the estimation of the parameters in a mixture of two univariate normal densities. The sample from which the estimates are obtained is assumed to be independent and to consist entirely of unlabeled observations on the mixture. (Since this is the sort of sample dealt with in the vast majority of work on the problem at hand, it is understood in this review that all samples are of this type unless otherwise indicated.) The approach suggested by Pearson for solving the problem is known as the method of

moments. The method of moments consists generally of equating some set of sample moments to their expected values and thereby obtaining a system of (generally nonlinear) equations for the parameters in the mixture density. To estimate the five independent parameters in a mixture of two univariate normal densities according to the procedure of [97], one begins with equations determined by the first five moments and, after considerable algebraic manipulation, ultimately arrives at expressions for estimates which depend on a suitably chosen root of a single ninth degree polynomial.

From the time of the appearance of Pearson's paper until the use of high speed electronic computers became widespread in the 1960's, only fairly simple mixture density estimation problems were studied, and the method of moments was usually the method of choice for their solution. During this period, most energy devoted to mixture problems was directed toward mixtures of normal densities, especially toward Pearson's case of two univariate normal densities. Indeed, most work on normal mixtures during this period was intended either to simplify the job of obtaining Pearson's estimates or to offer more accessible estimates in restricted cases. Charlier [24] described the implementation of Pearson's method as "an heroic task", and suggested a somewhat simpler method of solving the moment equations which involves a cubic and ratio of two other polynomials. Pearson and Lee [99] recommended using "incomplete" normal moment functions to obtain first approximations to the

roots of the nomic equation produced by the procedure of Pearson [97]. Charlier and Wicksell [25] further simplified the method of Pearson [97], suggested graphical methods for obtaining roots of the nomic, and studied estimates which can be obtained relatively easily under the assumption of known means, equal variances, or symmetry of the mixture density. Burrau [18] computed certain "half-invariant" functions of the moments, thereby obtaining new equations for the five unknown parameters; convenient methods for the solution of these equations are offered in the companion paper of Strömberg [119]. Gottschalk [51] exploited symmetry to obtain simple equations satisfied by the moment estimates for a symmetric mixture of two univariate normal densities. Graphical aids for obtaining Pearson's moment estimates were derived by Sittig [116], Wiechselberger [131], and Preston [104]. Cohen [31] suggested circumventing the solution of Pearson's nomic equation via an iteration which involves solving a cubic equation at each step. An independent sample from one component of the mixture was used by Dick and Bowden [42] to estimate one mean and one variance, thereby reducing to three the number of parameters to be estimated from an unlabeled sample on the mixture; their estimates were used as initial approximations in an iterative procedure for approximating maximum-likelihood estimates. Gridgeman [53] discussed moment estimates of the variances and the mixing proportion under the assumption of a common mean. Robertson and Fryer [113] and Fryer and Robertson [47] studied the statistical properties of the moment estimates and compared them to the multinomial maximum-

likelihood and minimum chi-square estimates obtained by grouping the sample observations. Assuming equal variances, Tan and Chang [121] compared the efficiency of the moment and maximum-likelihood estimates by computing the asymptotic variances of the estimates. The space of acceptable solutions of the moment equations was described by Bowman and Shenton [16]. Finally, Quandt and Ramsey [106] compared moment estimates with the estimates produced by their moment generating function method, of which we say more later.

Some work has been done extending Pearson's method of moments to more general mixtures of normal densities and to mixtures of other continuous densities. Pollard [103] obtained moment estimates for a mixture of three univariate normal densities by assuming symmetry and other simplifying features which reduce the number of unknown parameters to four. The problem of obtaining moment estimates for mixtures of multivariate normal densities was considered by Cooper [33]. Assuming equal mixing proportions for simplicity, he explored both the two-component case involving general component covariance matrices and the multiple-component case for spherically symmetric component densities. Day [36] investigated moment estimates for a mixture of two multivariate normal densities with a common covariance matrix. Gumbel [54] derived moment estimates for the means in a mixture of two exponential densities under the assumption that the mixing proportions are known. The results of [54] were extended by Rider [111] to

include estimates of unknown proportions as well as means. Later, Rider offered moment estimates for mixtures of Weibull distributions in [112].

Moment estimates for a variety of simple mixtures of discrete densities were derived more or less in parallel with moment estimates for mixtures of normal and continuous densities. Pearson [98] constructed moment estimates for a mixture of two binomial densities of common unknown power and for a mixture of two Poisson densities. Muench [90] published simpler estimates for a mixture of two binomial densities of known power; in [91], he sketched the extension of the results of [98] and [90] to mixtures of any number of Poisson densities or binomial densities of common known power. Later, the moment estimates for a mixture of two Poisson densities were independently re-derived by Schilling [115]. In the case of known mixing proportions, the moment estimates for a mixture of two Poisson densities were obtained independently by Gumbel [54] and Arley and Buch [3]. Further independent reconstruction and extension of earlier work was done by Rider [112] and Blischke [11]. In [112], moment estimates are derived for mixtures of two of either the Poisson, binomial, negative binomial, or (as mentioned above) Weibull densities. In a construction paralleling that of [112], moment estimates are given in [11] for a mixture of two binomial densities of common known power; in addition, properties of these estimates such as their limiting distributions and asymptotic relative efficiencies are considered. The results of Rider [112]

were simplified through the use of factorial rather than ordinary moments and extended to include certain alternative estimates and additional mixtures by Cohen [29]. Following the outline of Muench [91], Blischke [13] extended the results of [11] to give moment estimates for a mixture of any number of binomial densities of common known power. For additional information on moment estimation and many other topics of interest for mixtures of discrete distributions, we refer the reader to the extensive survey of Blischke [12].

Before leaving the method of moments, we mention the important problem of estimating the proportions alone in a mixture density under the assumption that the component densities, or at least some useful statistics associated with them, are known. Most general mixture density estimation procedures can be brought to bear on this problem, and the manner of applying these general procedures to this problem is usually independent of the particular forms of the densities in the mixture. In addition to the general estimation procedures, a number of special procedures have been developed for this problem; these are discussed in the third subsection of this review. The method of moments has the attractive property for this problem that the moment equations are linear in the mixture proportions. Moment estimates of proportions were discussed by Odell and Basu [92]. The sensitivity of moment estimates and other proportion estimates to changes in location of the component densities was studied by Tubba and Coberly [127].

2.2 The method of maximum likelihood.

With the arrival of increasingly powerful computers and increasingly sophisticated numerical methods during the 1960's, investigators began to turn from the method of moments to the method of maximum likelihood as the most widely preferred approach to mixture density estimation problems. To reiterate the working definition given in the introduction, we say that a maximum-likelihood estimate associated with a sample of observations is a choice of parameters which maximizes the probability density function of the sample, called in this context the likelihood function. In the next section, we define precisely the maximum-likelihood estimates of interest here and comment on their properties. In this subsection, we offer a very brief tour of the literature addressing maximum-likelihood estimation for mixture densities. Of course, more is said in the sequel about most of the work mentioned below.

Actually, maximum-likelihood estimates and their associated efficiency were often the subject of wishful thinking prior to the advent of computers, and some work was done then toward obtaining maximum-likelihood estimates for simple mixtures. Specifically, Baker [4] obtained maximum-likelihood estimates of the ratio of the proportions both in a mixture of two essentially arbitrary univariate densities for samples of sizes two and three and in a mixture of two univariate densities which are uniform over intervals for arbitrary sample sizes. Also, Rao [107] considered a mixture of two univariate normal densities with

equal variances and specified the likelihood equations, a system of four equations satisfied by the four unknown parameters at the maximum-likelihood estimate. He suggested solving the likelihood equations numerically with an iterative procedure known as the method of scoring, which we describe in Section 6. Finally, Mendenhall and Hader [89] obtained maximum-likelihood estimates of the parameters in a mixture of two exponential densities using a sample in which some of the observations are labeled. They reduced the problem of obtaining the estimates to that of solving a single nonlinear equation in one unknown; a numerical solution of this equation was found using Newton's method. Despite this early work, however, the problem of obtaining maximum-likelihood estimates was generally considered during this period to be completely intractable for computational reasons.

As computers became available to ease the burden of computation, maximum-likelihood estimation was proposed and studied in turn for a variety of increasingly complex mixture densities. As before, mixtures of normal densities were the subject of considerable attention. Hasselblad [64] treated maximum-likelihood estimation for mixtures of any number of univariate normal densities; his major results were later obtained independently by Behboodian [70]. Mixtures of two multivariate normal densities with a common unknown covariance matrix were addressed by Day [36]. The general case of a mixture of any number of multivariate normal densities was considered by Wolfe [132], and additional work on this case was done by Duda

and Hart [44] and Peters and Walker [101]. Tan and Chang [121] compared the moment and maximum-likelihood estimates for a mixture of two univariate normal densities with common variance by computing the asymptotic variances of the estimates; they found that maximum-likelihood estimates are much better, especially when the component densities are poorly separated. Hosmer [67] reported on a Monte Carlo study of maximum-likelihood estimates for a mixture of two univariate normal densities when the component densities are not well separated and the sample size is small; the results of his study suggest that the method of maximum-likelihood should be used with considerable caution in such cases.

Several interesting variations on the usual estimation problem for mixtures of normal densities have been addressed in the literature. Hosmer [68] compared the maximum-likelihood estimates for a mixture of two univariate normal densities obtained from three different types of samples, the first of which is the usual type consisting of only unlabeled observations and the second two of which consist of both labeled and unlabeled observations and are distinguished by whether or not the labeled observations contain information about the mixing proportions. (We elaborate on the nature of these samples and how they might arise in Section 2.) Earlier, Tan and Chang [120] considered a problem from genetics which is nearly identical to that considered by Hosmer [68] for partially labeled samples which contain no information about the mixing proportions. Also, Dick

and Bowden [42] independently addressed a special case of this problem in which maximum-likelihood estimates are obtained using a sample of labeled observations from one component population together with a sample of unlabeled observations on the mixture. Finally, a number of authors have investigated maximum-likelihood estimates for a "switching regression" model which is a certain type of estimation problem for mixtures of normal densities; see the papers of Quandt [105], Hosmer [69], Kiefer [77], and the comments by Hartley [63], Hosmer [70], and Kiefer [78] on the paper of Quandt and Ramsey [106]. A generalization of the model considered by these authors was touched on by Dennis [39].

Maximum-likelihood estimation has also been studied for a variety of unusual and general mixture density problems, some of which include but are not restricted to the usual normal mixture problem. Cohen [30] considered an unusual but simple mixture of two discrete densities, one of which has support at a single point; he focused in particular on the case in which the other density is a negative binomial density. Hasselblad [65] generalized his earlier results in [64] to include mixtures of any number of univariate densities from exponential families. He included a short study comparing maximum-likelihood estimates with the moment estimates of Blischke [13] for a mixture of two binomial distributions. Baum, Petrie, Soules, and Weiss [7] addressed a mixture estimation problem which is both unusual and in one respect more general than the problems considered in the sequel. In their problem, the a priori probabilities of sample

observations coming from the various component populations in the mixture are not independent from one observation to the next (that is, they are not simply the proportions of the component populations in the mixture) but rather are specified to follow a Markov chain. Their results are specifically applied to mixtures of univariate normal, gamma, binomial, and Poisson densities and to mixtures of general strictly log concave density functions which are identical except for unknown location and scale parameters. Peters and Coberly [100] and Peters and Walker [102] treated maximum-likelihood estimates of proportions and subsets of proportions for essentially arbitrary mixture densities. Maximum-likelihood estimates were included by Tubbs and Coberly [127] in their study of the sensitivity of various proportion estimators. Other maximum-likelihood estimation problems which are closely related to those considered here are the latent structure problems touched on by Wolfe [132] (see also Lazarsfeld and Henry [81]) and the problems concerning frequency tables derived by indirect observation addressed by Haberman [57], [58], [59]. Finally, although infinite mixture densities of the general form (1.2) are specifically excluded from consideration here, we mention a very interesting result of Laird [80] to the effect that under various assumptions, the maximum-likelihood estimate of a possibly infinite mixture density is actually a finite mixture density.

2.3 Other methods.

In addition to the method of moments and the method of maximum-likelihood, a variety of other methods have been proposed for estimating parameters in mixture densities. Some of these methods are general purpose methods. Others are (or were at the time of their derivation) intended for mixture problems the forms of which make (or made) them either ill-suited for the application of more widely used methods or particularly well-suited for the application of special purpose methods.

For mixtures of any number of univariate normal densities, Harding [61] and Cassie [19] suggested graphical procedures employing probability paper as an alternative to moment estimates, which were at that time practically unobtainable in all but the simplest cases. Later, Bhattacharya [10] prescribed other graphical methods as a particularly simple way of resolving a mixture density into normal components. These graphical procedures work best on mixture populations which are well-separated in the sense that each component has an associated region in which the presence of the other components can be ignored.

Also for general mixtures of univariate normal densities, Doetsch [43] exhibited a linear operator which reduces the variances of the component densities without changing their proportions or means and used this operator in a procedure which determines the component densities one at a time. Medgyessy [88]

(see also the review by Mallows [86]) extended the techniques of [43] to a large class of univariate mixture densities subject to the restriction that each component density have no more than two unknown parameters. Gregor [52] prescribed an algorithm for implementing the methods of Doetsch [43] and Medgyessy [88] on a mixture of univariate normal densities. Stanat [117] broadened the methods of [43] and [88] to study mixtures of multivariate normal and Bernoulli densities. In [114], Sammon considered a mixture density consisting of an unknown number of component densities which are identical except for translation by unknown location parameters; he derived techniques based on convolution for estimating both the number of components in the mixture and the location parameters.

A number of specialized procedures have been developed for application to the problem of estimating the proportions in a mixture under the assumption that something about the component densities is known. Choi and Bulgren [28] proposed an estimate determined by a least-squares criterion in the spirit of the minimum-distance method of Wolfowitz [133]. A variant of the method of [28] for which smaller bias and mean-square error were reported was offered by Macdonald [84]. A method termed the confusion matrix method was given by Odell and Chhikara [93] (see also the review of Odell and Bacu [92]). In this method, an estimate is obtained by subdividing R^n into disjoint regions R_1, \dots, R_m and then solving the equation $P\hat{\alpha} = e$, in which $\hat{\alpha}$ is the estimated vector of proportions, e is a vector whose i^{th}

component is the fraction of observations falling in R_i , and the "confusion matrix" P has ij^{th} entry

$$\int_{R_i} p_j(x|\rho_j) dx .$$

The confusion matrix method is a special case of a method of Macdonald [85], whose formulation of the problem as a least-squares problem allows for a singular or rectangular confusion matrix. Earlier, special cases of estimates of this type were considered by Boes [14], [15]. Guseman and Walton [55], [56] employed certain pattern recognition notions and techniques to obtain numerically tractable confusion matrix proportion estimates for mixtures of multivariate normal densities. James [73] studied several simple confusion matrix proportion estimates for a mixture of two univariate normal densities. Ganesalingam and McLachlan [49] compared the performance of confusion matrix proportion estimates with maximum-likelihood proportion estimates for a mixture of two multivariate normal densities. Finally, we mention that Walker [130] considered a mixture of two essentially arbitrary multivariate densities and, assuming only that the means of the component densities are known, suggested a simple procedure using linear maps which yields unbiased proportion estimates.

A stochastic approximation algorithm for estimating the parameters in a mixture of any number of univariate normal densities was offered by Young and Coraluppi [139]. In such an algorithm, one determines a sequence of recursively updated

estimates from a sequence of observations of indeterminate length considered on a one-at-a-time or few-at-a-time basis. Such an algorithm is likely to be appealing when a sample of desired size is either unavailable in toto at any one point in time or unwieldy because of its size. Stochastic approximation of mixture proportions alone was considered by Kazakos [76].

Quandt and Ramsey [106] derived a procedure called the moment generating function method and applied it to the problem of estimating the parameters in a mixture of two univariate normal densities and in a switching regression model. In brief, a moment generating function estimate is a choice of parameters which minimizes a certain sum of squares of differences between the theoretical and sample moment generating functions. In a comment by Kiefer [78], it is pointed out that the moment generating function method can be regarded as a natural generalization of the method of moments. Kiefer [78] further offers an appealing heuristic explanation of the apparent superiority of moment generating function estimates over moment estimates reported by Quandt and Ramsey [106]. In a comment by Hosmer [70], evidence is presented that moment generating function estimates may in fact perform better than maximum-likelihood estimates in the small-sample case. The moment generating function method appears to be a potentially valuable tool in mixture density estimation problems.

Minimum chi-square estimation is a general method of estimation which has been touched on by a number of authors in

connection with the mixture density estimation problem but which has not become the subject of much consideration in depth in this context. In minimum chi-square estimation, one subdivides R^n into cells R_1, \dots, R_k and seeks a choice of parameters which minimizes

$$\chi^2(\Phi) = \sum_{j=1}^k \frac{(N_j - E_j(\Phi))^2}{E_j(\Phi)}$$

or some similar criterion function. In this expression, N_j and $E_j(\Phi)$ are, respectively, the observed and expected numbers of observations in R_j for $j = 1, \dots, k$. For mixtures of normal densities, minimum chi-square estimates were mentioned by Hasselblad [64], Cohen [31], Day [36], and Fryer and Robertson [47]. Minimum chi-square estimates of proportions were reviewed by Odell and Basu [92] and included in the sensitivity study of Tubbs and Coberly [127]. Macdonald [85] remarked that his weighted least-squares approach to proportion estimation suggested a convenient iterative method for computing minimum chi-square estimates.

As a final note, we mention three methods which have been proposed for general mixture density estimation problems. Choi [27] discussed the extension to general mixture density estimation problems of the least-squares method of Choi and Bulgren [28] for estimating proportions. Deely and Kruse [37] suggested an estimation procedure which is in spirit like that of Choi and Bulgren [28] and Choi [27], except that a sup-norm

distance is used in place of the square integral norm. Deely and Kruse argued that their procedure is computationally feasible, but no concrete examples or computation results are given in [37]. Yakowitz [135], [136] outlined a very general "algorithm" for constructing consistent estimates of the parameters in mixture densities which are identifiable in the sense described in the fifth subsection of this review. The sense in which his "algorithm" is really an algorithm in the usually understood sense of the word is discussed in [136].

2.4 The EM algorithm.

At several points in the review above, we have alluded to computational difficulties associated with obtaining maximum-likelihood estimates. For mixture density problems, these difficulties arise because of the complex dependence of the likelihood function on the parameters to be estimated. The customary way of finding a maximum-likelihood estimate is first to determine a system of equations called the likelihood equations which are satisfied by the maximum-likelihood estimate and then to attempt to find the maximum-likelihood estimate by solving these likelihood equations. The likelihood equations are usually found by differentiating the logarithm of the likelihood function, setting the derivatives equal to zero, and perhaps performing some additional algebraic manipulations. For mixture density problems, the likelihood equations are almost certain to be nonlinear and beyond hope of solution by analytic means. Consequently, one must resort to seeking an approximate solution via some iterative procedure.

There are, of course, many general iterative procedures which are suitable for finding an approximate solution of the likelihood equations and which have been honed to a high degree of sophistication within the optimization community. We have in mind here principally Newton's method and various quasi-Newton methods which are variants of it. In fact, the method of scoring, which was mentioned above in connection with the work of Rao [107] and which we describe in detail in the sequel, falls

into the category of Newton-like methods and is one such method which is specifically formulated for solving likelihood equations.

Our main interest here, however, is in a special iterative method which is unrelated to Newton's method and which has been applied to a wide variety of mixture problems over the last fifteen or so years. Following the terminology of Dempster, Laird, and Rubin [38], we call this method the EM algorithm (E for "expectation" and M for "maximization"). As we mentioned in the introduction, it has been found in most instances to have the advantage of reliable global convergence, low cost per iteration, economy of storage, and ease of programming as well as a certain heuristic appeal; unfortunately its convergence can be maddeningly slow in simple problems which are often encountered in practice.

The EM algorithm has been derived and studied from at least two distinct viewpoints by a number of authors, many of them working independently. Hasselblad [64] obtained the EM algorithm for an arbitrary finite mixture of univariate normal densities and made empirical observations about its behavior. In an extension of [64], he further prescribed the algorithm for essentially arbitrary finite mixtures of univariate densities from exponential families in [65]. The EM algorithm of [64] for univariate normal mixtures was given again by Behboodan [8], while Day [36] and Wolfe [32] formulated it for, respectively, mixtures of two multivariate normal densities with common

covariance matrix and arbitrary finite mixtures of multivariate normal densities. All of these authors apparently obtained the EM algorithm independently, although Wolfe [132] referred to Hasselblad [64]. They all derived the algorithm by setting the partial derivatives of the log-likelihood function equal to zero, and after some algebraic manipulation, obtained equations which suggest the algorithm.

Following these early derivations, the EM algorithm was applied by Tan and Chang [120] to a mixture problem in genetics and used by Hosmer [67] in the Monte Carlo study of maximum-likelihood estimates referred to earlier. Duda and Hart [44] cited the EM algorithm for mixtures of multivariate normal densities and commented on its behavior in practice. Hosmer [68] extended the EM algorithm for mixtures of two univariate normal densities to include the partially labeled samples described briefly above. Hartley [63] prescribed the EM algorithm for a "switching regression" model. Peters and Walker [101] offered a local convergence analysis of the EM algorithm for mixtures of multi-variate normal densities and suggested modifications of the algorithm to accelerate convergence. Peters and Coberly [100] studied the EM algorithm for approximating maximum-likelihood estimates of the proportions in an essentially arbitrary mixture density and gave a local convergence analysis of the algorithm. Peters and Walker [102] extended the results of [100] to include subsets of mixture proportions and a local convergence analysis along the lines of [101].

All of the above investigators regarded the EM algorithm as arising naturally from the particular forms taken by the partial derivatives of the log-likelihood function. A quite different point of view toward the algorithm was put forth by Dempster, Laird, and Rubin [38]. They interpreted the mixture density estimation problem as an estimation problem involving incomplete data by regarding an unlabeled observation on the mixture as "missing" a label indicating its component population of origin. In doing so, they not only related the mixture density problem to a broader class of statistical problems but also showed that the EM algorithm for mixture density problems is really a specialization of a more general algorithm (also called the EM algorithm in [38]) for approximating maximum-likelihood estimates from incomplete data. As one sees in the sequel, this more general EM algorithm is defined in such a way that it has certain desirable theoretical properties by its very definition. Earlier, the EM algorithm was defined independently in a very similar manner by Baum et al [7] for very general mixture density estimation problems and by Haberman [57], [58], [59] for mixture-related problems involving frequency tables derived by indirect observation. Haberman also refers in [59] to versions of his algorithm developed by Ceppellini, Siniscalco, and Smith [20], Chen [26], and Goodman [50]. In addition, an interpretation of mixture problems as incomplete data problems was given in the brief discussion of mixtures by Orchard and Woodbury [94]. The desirable theoretical properties automatically enjoyed by the EM algorithm suggest in turn the

good global convergence behavior of the algorithm which has been observed in practice by many investigators. Theorems which essentially confirm this suggested behavior have been recently obtained by Redner [109], Boyles [17], and Wu [134] and are outlined in the sequel.

2.5 Identifiability and information.

To complete this review, we touch on two topics which have to do with the general well-posedness of estimation problems rather than with any particular method of estimation. The first topic, identifiability, addresses the theoretical question of whether it is possible to uniquely estimate a parameter from a sample, however large. The second topic, information, relates to the practical matter of how good one can reasonably hope for an estimate to be. A thorough survey of these topics is far beyond the scope of this review; we try to cover below those aspects of them which have a specific bearing on the sequel.

In general, a parametric family of probability density functions is said to be identifiable if distinct parameter values determine distinct members of the family. For families of mixture densities, this general definition requires a special interpretation. For the purposes of this paper, let us first say that a mixture density $p(x|\Phi)$ of the form (1.1) is economically represented if, for each pair of integers i and j between 1 and m , one has that $p_i(x|\rho_i) = p_j(x|\rho_j)$ for almost all $x \in R^n$ (relative to the underlying measure on R^n appropriate for $p(x|\Phi)$) only if either $i = j$ or one of α_i and α_j is zero. Then it suffices to say that a family of mixture densities of the form (1.1) is identifiable for $\Phi \in \Omega$ if for each pair $\Phi' = (\alpha_1^i, \dots, \alpha_m^i, \rho_1^i, \dots, \rho_m^i)$ and $\Phi'' = (\alpha_1^m, \dots, \alpha_m^m, \rho_1^m, \dots, \rho_m^m)$ in Ω determining economically represented densities $p(x|\Phi')$ and $p(x|\Phi'')$, one has that $p(x|\Phi') = p(x|\Phi'')$ for almost all $x \in R^n$

only if there is a permutation π of $(1, \dots, m)$ such that $\alpha_i^i = \alpha_{\pi(i)}^m$ and, if $\alpha_i^i \neq 0$, $\phi_i^i = \phi_{\pi(i)}^m$ for $i = 1, \dots, m$. For a more general definition suitable for possibly infinite mixture densities of the form (1.2), see, for example, Yakowitz and Spragins [137].

It is tacitly assumed here that all families of mixture densities under consideration are identifiable. One can easily determine the identifiability of specific mixture densities using, for example, the identifiability characterization theorem of Yakowitz and Spragins [137]. For more on identifiability of mixture densities, the reader is referred to the papers of Teicher [123], [124], [125], [126], Barndorff-Nielsen [5], Yakowitz and Spragins [137], and Yakowitz [135], [136] and to the book by Maritz [70].

The Fisher information matrix is given by

$$I(\Phi) = \int_{R^n} [\nabla_{\Phi} \log p(x|\Phi)] [\nabla_{\Phi} \log p(x|\Phi)]^T p(x|\Phi) d\mu, \quad (2.5.1)$$

provided that $p(x|\Phi)$ is such that this expression is well-defined. (In writing ∇_{Φ} , we suppose that one can conveniently redefine Φ as a vector $\Phi = (\xi_1, \dots, \xi_\nu)^T$ of unconstrained scalar parameters, and we take $\nabla_{\Phi} = (\frac{\partial}{\partial \xi_1}, \dots, \frac{\partial}{\partial \xi_\nu})^T$. Also, in (2.5.1), μ denotes the underlying measure on R^n appropriate for $p(x|\Phi)$.) The Fisher information matrix has general significance concerning the distribution of unbiased and asymptotically unbiased estimates. For the present purposes, the importance of the

- Fisher information matrix lies in its role in determining the asymptotic distribution of maximum-likelihood estimates (see Theorem 3.1 below).

A number of authors have considered the Fisher information matrix for finite mixture densities in a variety of contexts. We mention in particular several investigations in which the Fisher information matrix is of central interest. (There have been others in which the Fisher information matrix or some approximation of it has played a significant but less prominent role; see those of Mendenhall and Hader [89], Hasselblad [64], [65], Day [36], Wolfe [132], Dick and Bowden [42], Hosmer [67], James [73], and Ganesalingam and McLachlan [49].) Hill [66] exploited simple approximations obtained in limiting cases from a general power series expansion to investigate the Fisher information for estimating the proportion in a mixture of two normal or exponential densities. Behboodian [9] offered methods for computing the Fisher information matrix for the proportion, means, and variances in a mixture of two univariate normal densities; he also provided four-place tables from which approximate information matrices for a variety of parameter values can be easily obtained. In their comparison of moment and maximum-likelihood estimates, Tan and Chang [121] numerically evaluated the diagonal elements of the inverse of the Fisher information matrix at a variety of parameter values for a mixture of two univariate normal densities with a common variance. Using the Fisher information matrix, Chang [22] investigated the

effects of adding a second variable on the asymptotic distribution of the maximum-likelihood estimates of the proportion and parameters associated with the first variable in a mixture of two normal densities. Later, Chang [23] extended the methods of [22] to include mixtures of two normal densities on variables of arbitrary dimension. For a mixture of two univariate normal densities, Hosmer and Dick [71] considered Fisher information matrices determined by a number of sample types. They compared the asymptotic relative efficiencies of estimates from totally unlabeled samples, estimates from two types of partially labeled samples, and estimates from two types of completely labeled samples.

3. Maximum-likelihood

In this section, maximum-likelihood estimates for mixture densities are defined precisely, and their important properties are discussed. It is assumed that a parametric family of mixture densities of the form (1.1) is specified and that a particular $\Phi^* = (\alpha_1^*, \dots, \alpha_m^*, \rho_1^*, \dots, \rho_m^*) \in \Omega$ is the "true" parameter value to be estimated. As before, it is both natural and convenient to regard $p(x|\Phi)$ in (1.1) as modeling a statistical population which is a mixture of m component populations with associated component densities $\{p_i\}_{i=1, \dots, m}$ and mixing proportions $\{\alpha_i\}_{i=1, \dots, m}$.

In order to suggest to the reader the variety of samples which might arise in mixture problems as well as to provide a framework within which to discuss samples of interest in the sequel, we introduce samples of observations in R^n of four distinct types. All of the mixture density estimation problems which we have encountered in the literature involve samples which are expressible as one or a stochastically independent union of samples of these types, although the imaginative reader can probably think of samples for mixture problems which can not be so represented. The four types of samples and the notation which we associate with them are given as follows:

Type 1. Suppose that $\{x_k\}_{k=1, \dots, N}$ is an independent sample of N unlabeled observations on the mixture, i.e., a set of N observations on independent, identically distributed random variables with density $p(x|\Phi^*)$. Then

$S_1 = \{x_k\}_{k=1, \dots, N}$ is a sample of Type 1.

Type 2. Suppose that J_1, \dots, J_m are arbitrary non-negative integers and that for $i = 1, \dots, m$, $\{y_{ik}\}_{k=1, \dots, J_i}$ is an independent sample of observations on the i^{th} component population, i.e., a set of J_i observations on independent, identically distributed random variables with density $p_i(x|\rho_i^*)$. Then $S_2 = \bigcup_{i=1}^m \{y_{ik}\}_{k=1, \dots, J_i}$ is a sample of Type 2.

Type 3. Suppose that an independent sample of K unlabeled observations is drawn on the mixture, that these observations are subsequently labeled, and that for $i = 1, \dots, m$, a set $\{z_{ik}\}_{k=1, \dots, K_i}$ of them is associated with the i^{th} component population with $K = \sum_{i=1}^m K_i$. Then $S_3 = \bigcup_{i=1}^m \{z_{ik}\}_{k=1, \dots, K_i}$ is a sample of Type 3.

Type 4. Suppose that an independent sample of M unlabeled observations is drawn on the mixture, that the observations in the sample which fall in some set $E \subseteq R^n$ are subsequently labeled, and that for $i = 1, \dots, m$, a set $\{w_{ik}\}_{k=1, \dots, M_i}$ of them is thereby associated with the i^{th} component population while a set $\{w_{0k}\}_{k=1, \dots, M_0}$ remains unlabeled. Then $S_4 = \bigcup_{i=0}^m \{w_{ik}\}_{k=1, \dots, M_i}$ is a sample of Type 4.

A totally unlabeled sample S_1 of Type 1 is the sort of sample considered in almost all of the literature on mixture densities. Throughout most of the sequel, it is assumed as a convenience that samples under consideration are of this type. The major qualitative difference between completely labeled samples S_2 and S_3 of Types 2 and 3, respectively, is that the numbers K_1 contain information about the mixing proportions while the numbers J_1 do not. Thus if estimation of proportions is of interest, then a sample S_2 is useful only as a subset of a larger sample which includes samples of other types. For mixtures of two univariate densities, Hosmer [68] considered samples of the forms S_1 , $S_1 \cup S_2$, and $S_1 \cup S_3$. Previously Tan and Chang [120] considered a problem involving an application of mixtures in explaining genetic variation which is almost identical to that of [68] in which the sample is of the form $S_1 \cup S_2$. Also Dick and Bowden [42] used a sample of the form $S_1 \cup S_2$ in which $m = 2$ and $J_2 = 0$. Hosmer and Dick [71] evaluated the Fisher information matrix for a variety of samples of Types 1, 2, 3, and their unions.

A sample S_4 of Type 4 is likely to be associated with a mixture problem involving censored sampling. While the numbers M_1 contain information about the mixing proportions, as do the numbers K_1 of a sample S_3 of Type 3, they also contain information about the parameters of the component densities while the numbers K_1 do not. An interesting and informative example of how a sample of Type 4 might arise is the following, which is

in the area of life testing and is outlined by Mendenhall and Hader [89].

Example. In life testing, one is interested in testing "products" (systems, devices, etc.), recording failure times or causes, and hopefully thereby being better able to understand and improve the performance of the product. It often happens that products of a particular type fail as a result of two or more distinct causes. (An example of Acheson and McElwee [1] is quoted in [89] in which the causes of electronic tube failure are divided into gaseous defects, mechanical defects, and normal deterioration of the cathode.) It is therefore natural to regard collections of such products as mixture populations, the component populations of which correspond to the distinct causes of failure. The first objective of life testing in such cases is likely to be estimation of the proportions and other statistical parameters associated with the failure component populations.

Because of restrictions on time available for testing, life testing experiments must often be concluded after a predetermined length of time has elapsed or after a predetermined number of product units have failed, resulting in censored sampling. If the causes of failure of the failed products are determined in the course of such an experiment, then the (labeled) failed products together with those (unlabeled) products which did not fail constitute a sample of Type 4.

The likelihood function of a sample of observations is the probability density function of the random sample evaluated at

OF POOR QUALITY

the observations at hand. When maximum-likelihood estimates are of interest, it is usually convenient to deal with the logarithm of the likelihood function, called the log-likelihood function, rather than with the likelihood function itself. The following are the log-likelihood functions L_1 , L_2 , L_3 and L_4 of samples S_1 , S_2 , S_3 and S_4 of Types 1, 2, 3 and 4 respectively:

$$L_1(\Phi) = \sum_{k=1}^N \log p(x_k | \Phi) \quad (3.1)$$

$$L_2(\Phi) = \sum_{i=1}^m \sum_{k=1}^{J_i} \log p_i(y_{ik} | \rho_i) \quad (3.2)$$

$$L_3(\Phi) = \sum_{i=1}^m \sum_{k=1}^{K_i} \log[\alpha_i p_i(z_{ik} | \rho_i)] + \log \frac{K!}{K_1! \cdots K_m!} \quad (3.3)$$

$$L_4(\Phi) = \sum_{k=0}^{M_0} \log p(w_{0k} | \Phi) + \sum_{i=1}^m \sum_{k=1}^{M_i} \log[\alpha_i p_i(w_{ik} | \rho_i)] + \log \frac{M!}{M_0! \cdots M_m!} \quad (3.4)$$

Note that if a sample of observations is a union of independent samples of the types considered here, then the log-likelihood function of the sample is just the corresponding sum of log-likelihood functions defined above for the samples in the union.

If S is a sample of observations of the sort under consideration, then by a maximum-likelihood estimate of Φ^* , we mean any choice of Φ in Ω at which the log-likelihood

function of S , denoted by $L(\Phi)$, attains its largest local maximum in Ω . In defining a maximum-likelihood estimate in this way, we have taken into account two practical difficulties associated with maximum-likelihood estimation for mixture densities.

The first difficulty is that one cannot always in good conscience take Ω to be a set in which the log-likelihood function is bounded above, and so there are not always points in Ω at which L attains a global maximum over Ω . Perhaps the most notorious mixture problem for which L is not bounded above in Ω is that in which p is a mixture of normal densities and $S = S_1$, a sample of Type 1. It is easily seen in this case that if one of the mixture means coincides with a sample observation and if the corresponding variance tends to zero (or if the corresponding covariance matrix tends in certain ways to a singular matrix in the multivariate case), then the log-likelihood function increases without bound. For the normal mixture problem, an advantage of including labeled observations in a sample is that with probability one, this difficulty does not occur if the sample includes more than n labeled observations from each component population. This was observed in the univariate case by Hosmer [68].

The second difficulty is that mixture problems are very often such that the log-likelihood function attains its largest local maximum at several different choices of Φ . Indeed, if p_1 and p_j are of the same parametric family for some i and j

and if $S = S_1$, a sample of Type 1, then the value of $L(\Phi)$ will not change if the component pairs (α_i, ρ_i) and (α_j, ρ_j) are interchanged in Φ , i.e., if in effect there is "label switching" of the i^{th} and j^{th} component populations. The results reviewed below show that whether or not such "label switching" is a cause for concern depends on whether estimates of the particular component density parameters are of interest or whether only an approximation of the mixture density is desired. We remark that this "label switching" difficulty can certainly occur in mixtures which are identifiable (see Section 2.5).

In the remainder of this section, our interest is in the important general qualitative properties of maximum-likelihood estimates of mixture density parameters. For convenience, we restrict the discussion to the case which is most often addressed in the literature, namely that in which the sample S at hand is a sample S_1 of Type 1. We also assume that each component density p_i is differentiable with respect to ρ_i and make the nonessential assumption that the parameters ρ_i are unconstrained in Ω_1 and mutually independent variables. It is not difficult to modify the discussion below to obtain similar statements which are appropriate for other mixture density estimation problems of interest. For a discussion of the properties of maximum-likelihood estimates of constrained variables, see the paper of Aitchison and Silvey [2].

The traditional general approach to determining a maximum-likelihood estimate is first to arrive at a system of likelihood

equations satisfied by the maximum-likelihood estimate and then to try to obtain a maximum-likelihood estimate by solving the likelihood equations. Basically, the likelihood equations are found by considering the partial derivatives of the log-likelihood function with respect to the components of Φ . If $\hat{\Phi} = (\hat{\alpha}_1, \dots, \hat{\alpha}_m, \hat{\rho}_1, \dots, \hat{\rho}_m)$ is a maximum-likelihood estimate, then one has the likelihood equations

$$\nabla_{\rho_i} L(\hat{\Phi}) = 0, \quad i = 1, \dots, m \quad (3.5)$$

determined by the unconstrained parameters $\rho_i, i = 1, \dots, m$. (Our convention is that "∇" with a variable appearing as a subscript indicates the gradient of first partial derivatives with respect to the components of the variable.)

To obtain likelihood equations determined by the proportions, which are constrained to be non-negative and to sum to one, we follow Peters and Walker [102]. Setting $\hat{\alpha} = (\hat{\alpha}_1, \dots, \hat{\alpha}_m)^T$, one sees that

$$0 > \nabla_{\alpha} L(\hat{\Phi})^T (\alpha - \hat{\alpha}) \quad (3.6)$$

for all $\alpha = (\alpha_1, \dots, \alpha_m)^T$ such that $\sum_{i=1}^m \alpha_i = 1$ and $\alpha_i > 0, i = 1, \dots, m$. Now (3.6) holds for all α satisfying the given constraints if and only if

$$0 > \nabla_{\alpha} L(\hat{\Phi})^T (e_i - \hat{\alpha}), \quad i = 1, \dots, m,$$

with equality for those values of i for which $\hat{\alpha}_i > 0$. (Here, e_i is the vector the i^{th} component of which is one and the other components of which are zero.) It follows that (3.6) is equivalent to

$$1 > \frac{1}{N} \sum_{k=1}^N \frac{p_i(x_k | \hat{\rho}_i)}{p(x_k | \hat{\phi})}, \quad i = 1, \dots, m, \quad (3.7)$$

with equality for those values of i for which $\hat{\alpha}_i > 0$. Finally, multiplying each side of (3.7) by $\hat{\alpha}_i$ for $i = 1, \dots, m$ yields likelihood equations in the convenient form

$$\hat{\alpha}_i = \frac{1}{N} \sum_{k=1}^N \frac{\hat{\alpha}_i p_i(x_k | \hat{\rho}_i)}{p(x_k | \hat{\phi})}, \quad i = 1, \dots, m. \quad (3.8)$$

We remark that it is easily seen by considering the matrix of second partial derivatives of L with respect to $\alpha_1, \dots, \alpha_m$ that L is a concave function of $\alpha = (\alpha_1, \dots, \alpha_m)^T$ for any fixed set of values $\hat{\rho}_i \in \Omega_i$, $i = 1, \dots, m$. Thus, for any fixed $\hat{\rho}_i$, $i = 1, \dots, m$, (3.6) and, hence, (3.7) are sufficient as well as necessary for $\hat{\alpha}$ to maximize L over the set of all α satisfying the given constraints. On the other hand, the likelihood equations (3.8) are necessary but not sufficient conditions for $\hat{\alpha}$ to maximize L for fixed $\hat{\rho}_i$, $i = 1, \dots, m$. Indeed, $\hat{\alpha} = e_i$ satisfies (3.8) for $i = 1, \dots, m$. In fact, it follows from the concavity of L that there is a solution of (3.8) in each (closed) face of the simplex of points α satisfying the given constraints. In spite of perhaps suffering

from a surplus of solutions, the likelihood equations (3.8) nevertheless have a useful form which takes on additional significance later in the context of the EM algorithm.

The equations (3.5) and (3.8) together constitute a full set of likelihood equations which are necessary but not sufficient conditions for a maximum likelihood estimate. Of course, some irrelevant solutions of the likelihood equations can be avoided in practice by using one of a number of procedures for obtaining a numerical solution of them (among which is the EM algorithm) which in all but the most unfortunate circumstances will yield a local maximizer of the log-likelihood function (or a singularity near which it grows without bound) rather than some stationary point which is a local minimizer or a saddle point. Still, it is natural to ask at this point the extent to which solving the likelihood equations can be expected to produce a maximum-likelihood estimate and the extent to which a maximum-likelihood estimate can be expected to be a good approximation of Φ^* .

Two general theorems are offered below which give a fair summary of the results in the literature most pertinent to the question put forth above. As a convenience, we assume that $\alpha_i^* > 0$ for $i = 1, \dots, m$. For the purposes of the theorems and the discussion following them, this justifies writing, say,

$\alpha_m = 1 - \sum_{i=1}^{m-1} \alpha_i$ and considering the redefined, locally unconstrained variable $\Phi = (\alpha_1, \dots, \alpha_{m-1}, \rho_1, \dots, \rho_m)$ in the modified set

OF POOR QUALITY

$$\Omega = \{(\alpha_1, \dots, \alpha_{m-1}, \rho_1, \dots, \rho_m) : \sum_{i=1}^{m-1} \alpha_i < 1 \text{ and } \alpha_i > 0, \rho_i \in \Omega_i \text{ for } i = 1, \dots, m\} .$$

The likelihood equations (3.5) and (3.8) can now be written in the general unconstrained form

$$\nabla_{\phi} L(\hat{\phi}) = 0 , \tag{3.9}$$

which facilitates our presenting the theorems as general results which are not restricted to the mixture problem at hand or, for that matter, to mixture problems at all. In our discussion of the theorems, all statements regarding measure and integration are made with respect to the underlying measure on \mathbb{R}^n appropriate for $p(x|\phi)$, which we denote by μ .

The first theorem states roughly that under reasonable assumptions, there is a unique strongly consistent solution of the likelihood equations (3.9) and this solution at least locally maximizes the log-likelihood function and is asymptotically normally distributed. (Consistent in the usual sense means converging with probability approaching 1 to the true parameters as the sample size approaches infinity; strongly consistent means having the same limit with probability 1.) This theorem is a compendium of results generalizing the initial work of Cramér [35] concerning existence, consistency, and asymptotic normality of the maximum-likelihood estimate of a single scalar parameter. The conditions below, on which the theorem rests, were

essentially given by Chanda [21] as multi-dimensional generalizations of those of Cramér. With them, Chanda claimed that there exists a unique solution of the likelihood equations which is consistent in the usual sense (this fact was correctly proved by Tarone and Gruenhagen [122]) and established its asymptotic normal behavior. (See also the summary in Kiefer [77], the discussion in Zacks [71], and the related material for constrained maximum-likelihood estimates in Aitchison and Silvey [2].) Using these same conditions, Peters and Walker [101, Appendix A] showed that there is a unique strongly consistent solution of the likelihood equations and that it at least locally maximizes the log-likelihood function.

In stating the following conditions and in the discussion after the theorem, it is convenient to adopt temporarily the notation $\Phi = (\xi_1, \dots, \xi_\nu)$, where $\nu = (m - 1 + \sum_{i=1}^m n_i)$ and $\xi_i \in R^1$ for $i = 1, \dots, \nu$. Also, we remark that because the results of the theorem below implied by these conditions are strictly local in nature, there is no loss of generality in restricting Ω to be any neighborhood of Φ^* if such a restriction is necessary for the first condition to be met.

Condition 1. For all $\Phi \in \Omega$, for almost all $x \in R^n$, and for $i, j, k = 1, \dots, \nu$, the partial derivatives $\frac{\partial p}{\partial \xi_i}$, $\frac{\partial^2 p}{\partial \xi_i \partial \xi_j}$, and $\frac{\partial^3 p}{\partial \xi_i \partial \xi_j \partial \xi_k}$ exist and satisfy

$$\left| \frac{\partial p(x|\Phi)}{\partial \xi_i} \right| < f_i(x) , \quad \left| \frac{\partial^2 p(x|\Phi)}{\partial \xi_i \partial \xi_j} \right| < f_{ij}(x) , \quad \left| \frac{\partial^3 \log p(x|\Phi)}{\partial \xi_i \partial \xi_j \partial \xi_k} \right| < f_{ijk}(x) ,$$

where f_i and f_{ij} are integrable and f_{ijk} satisfies

$$\int_{R^n} f_{ijk}(x) p(x|\Phi^*) d\mu < \infty$$

Condition 2. The Fisher information matrix $I(\Phi)$ given by (2.5.1) is well-defined and positive definite at Φ^* .

Theorem 3.1. If Conditions 1 and 2 are satisfied and any sufficiently small neighborhood of Φ^* in Ω is given, then with probability 1, there is for sufficiently large N a unique solution Φ^N of the likelihood equations (3.9) in that neighborhood and this solution locally maximizes the log-likelihood function. Furthermore, $\sqrt{N}(\Phi^N - \Phi^*)$ is asymptotically normally distributed with mean zero and covariance matrix $I(\Phi^*)^{-1}$.

The second theorem is directed toward two questions left unresolved by the theorem above regarding Φ^N , the unique strongly consistent solution of the likelihood equations. The first question is whether Φ^N is really a maximum-likelihood estimate, i.e., a point at which the log-likelihood function attains its largest local maximum. The second is whether, even if the answer to the first question is "yes", there are maximum-likelihood estimates other than Φ^N which lead to limiting densities other than $p(x|\Phi^*)$. Given our assumption of identifiability of the family of mixture densities $p(x|\Phi)$,

$\phi \in \Omega$, one easily sees that the theorem below implies that if Ω' is any compact subset of Ω which contains ϕ^* in its interior, then with probability 1, ϕ^N is a maximum-likelihood estimate in Ω' for sufficiently large N . Furthermore, every other maximum-likelihood estimate in Ω' is obtained from ϕ^N by the "label switching" described earlier and, hence, leads to the same limiting density $p(x|\phi^*)$. Accordingly, we usually assume in the sequel that Conditions 1 through 4 are satisfied and refer to ϕ^N as the unique strongly consistent maximum-likelihood estimate. The theorem is a slightly restricted version of a general result of Redner [110] which extends earlier work by Wald [129] on the consistency of maximum-likelihood estimates. It should be remarked that the result of [110] rests on somewhat weaker assumptions than those made here and is specifically aimed at families of distributions which are not identifiable.

For $\phi \in \Omega$ and sufficiently small $r > 0$, let $N_r(\phi)$ denote the closed ball of radius r about ϕ in Ω and define

$$p(x|\phi, r) = \sup_{\phi' \in N_r(\phi)} p(x|\phi')$$

and

$$p^*(x|\phi, r) = \max(1, p(x|\phi, r)) .$$

Condition 3. For each $\phi \in \Omega$ and sufficiently small $r > 0$,

ORIGINAL PAGE IS
OF POOR QUALITY

$$\int_{R^n} \log p^*(x|\phi^*, r) p(x|\phi^*) d\mu < \infty .$$

Condition 4.

$$\int_{R^n} \log p(x|\phi^*) p(x|\phi^*) d\mu < \infty .$$

Theorem 3.2. Let Ω' be any compact subset of Ω which contains ϕ^* in its interior, and set

$$C = \{ \phi \in \Omega' : p(x|\phi) = p(x|\phi^*) \text{ almost everywhere} \} .$$

If Conditions 3 and 4 are satisfied and D is any closed subset of Ω' not intersecting C , then with probability 1,

$$\lim_{N \rightarrow \infty} \sup_{\phi \in D} \frac{\prod_{k=1}^N p(x_k|\phi)}{\prod_{k=1}^N p(x_k|\phi^*)} = 0 .$$

From a theoretical point of view, Theorems 3.1 and 3.2 are adequate for mixture density estimation problems in providing assurance of the existence of strongly consistent maximum-likelihood estimates, characterizing them as solutions of the likelihood equations, and prescribing their asymptotic behavior. In practice, however, one must still contend with certain potential mathematical, statistical, and even numerical difficulties associated with maximum-likelihood estimates. Some possible mathematical problems have been suggested above: The log-likelihood function may have many local and global maxima and

perhaps even singularities; furthermore, the likelihood equations are likely to have solutions which are not local maxima of the log-likelihood function. According to Theorem 3.1, the statistical soundness (as measured by bias and variance) of the strongly consistent maximum-likelihood estimate is determined, at least for large samples, by the Fisher information matrix $I(\Phi^*)$. As it happens, $I(\Phi^*)$ also plays a role in determining the numerical well-posedness of the problem of approximating the strongly consistent maximum-likelihood estimate for large samples.

To show how $I(\Phi^*)$ enters into the problem of numerically approximating Φ^N for large samples, we recall that the condition of a problem is reflected by the relative sensitivity of its solution to perturbations in the data associated with the problem. For an optimization problem, the condition is customarily measured by the condition number of the Hessian matrix of the function to be optimized evaluated at the solution. (For the definition and properties of the condition number of a matrix, see, for example, Stewart [118].) For the log-likelihood function at hand, the Hessian matrix, which we denote by $H(\Phi)$, is given by

$$H(\Phi) = \sum_{k=1}^N \nabla_{\Phi} \nabla_{\Phi}^T \log p(x_k | \Phi), \quad (3.10)$$

where $\nabla_{\Phi} \nabla_{\Phi}^T = \left(\frac{\partial^2}{\partial \xi_i \partial \xi_j} \right)$. If Conditions 1 and 2 above are satisfied, then it follows from the Strong Law of Large Numbers

(see Loève [82]) that with probability 1,

$$\lim_{N \rightarrow \infty} \frac{1}{N} H(\Phi^N) = -I(\Phi^*) . \quad (3.11)$$

Since $\frac{1}{N} H(\Phi^N)$ has the same condition number as $H(\Phi^N)$, (3.11) is the desired result.

To illustrate the potential severity of the statistical and numerical problems associated with maximum-likelihood estimates, we augment the material on the Fisher information matrix in the literature cited in Section 2.5 with Table 3.3 below, which lists approximate values of the condition number and the diagonal elements of the inverse of $I(\Phi^*)$ for a mixture of two univariate normal densities (see (1.3) and (1.4)) at a variety of choices of Φ^* . To prepare this table, we took $\Phi = (\alpha_1, \mu_1, \mu_2, \sigma_1^2, \sigma_2^2)$ and numerically evaluated $I(\Phi^*)$, its condition number, and its inverse for selected values of Φ^* using IMSL Library routines DCADRE, EIGRS, and LINV2P on a CDC7600.² The choices of Φ^* were obtained by taking $\alpha_1^* = .3$ and $\sigma_1^{2*} = \sigma_2^{2*} = 1$ and varying the mean separation $\mu_1^* - \mu_2^*$. In the table, the condition number of $I(\Phi^*)$ is denoted by κ , and the first through fifth diagonal elements of $I(\Phi^*)^{-1}$ are denoted by $I^{-1}(\alpha_1)$, $I^{-1}(\mu_1)$, $I^{-1}(\mu_2)$, $I^{-1}(\sigma_1^2)$, and $I^{-1}(\sigma_2^2)$, respectively.

2. We are grateful to the Mathematics and Statistics Division of the Lawrence Livermore National Laboratory for allowing us to use their computing facility in generating this table.

$\mu_1^* - \mu_2^*$	κ	$I^{-1}(\alpha_1)$	$I^{-1}(\mu_1)$	$I^{-1}(\mu_2)$	$I^{-1}(\sigma_1^2)$	$I^{-1}(\sigma_2^2)$
0.2	3.06×10^{10}	4.39×10^{10}	4.86×10^9	8.98×10^8	2.15×10^7	4.02×10^6
0.5	8.05×10^6	5.54×10^6	3.81×10^6	7.17×10^5	1.04×10^5	2.07×10^4
1.0	5.18×10^4	8.59×10^3	2.32×10^4	4.55×10^3	2.58×10^3	578.
1.5	4.80×10^3	237.	1.43×10^3	290.	383.	95.0
2.0	1.10×10^3	20.4	216.	45.8	115.	31.3
3.0	187.	.874	18.9	4.81	28.2	8.83
4.0	71.7	.267	5.72	1.95	13.4	4.71
6.0	35.7	.211	3.44	1.45	7.47	3.06

Table 3.3: Condition number and diagonal elements of the inverse of $I(\Phi^*)$ for a mixture of two univariate normal densities with $\alpha_1^* = .3$, $\sigma_1^{2*} = \sigma_2^{2*} = 1$.

ORIGINAL PART IS
OF POOR QUALITY

Table 3.3 reinforces one's intuitive understanding that for mixture density estimation problems, maximum-likelihood estimates are more appealing from both a statistical and a numerical standpoint if the component densities in the mixture are well separated than if they are poorly separated. Perhaps the most troublesome implication of Table 3.3 is that if the component densities are poorly separated, then impractically large sample sizes might be required in order to expect even moderately precise maximum-likelihood estimates. For example, Table 3.3 indicates that if one considers data from a mixture of two univariate normal densities with $\alpha_1^* = .3$, $\sigma_1^{2*} = \sigma_2^{2*} = 1$, and $\mu_1^* - \mu_2^* = 1$, then a sample size on the order of 10^6 is necessary to insure that the standard deviation of each component of the maximum-likelihood estimate is about 0.1 or less. Even if a sample of such horrendous size were available, the fact that evaluating the log-likelihood function and associated functions such as its derivatives involves summation over observations in the sample, considered together with the condition number of 5.18×10^4 for the information matrix, suggests that computing undertaken in seeking a maximum-likelihood estimate should be carried out with great care.

Similar observations regarding the asymptotic dependence of the accuracy of maximum-likelihood estimates on sample sizes and separation of the component populations have been made by a number of authors (Mendenhall and Hader [89], Hill [66], Hasselblad [64], [65], Day [36], Tan and Chang [121], Dick and

Bowden [42], Hosmer [67], [68], Hosmer and Dick [71]). Several of them (Mendenhall and Hader [89], Day [36], Hasselblad [65], Dick and Bowden [42], Hosmer [67]) also suggested that things are worse for small samples (less than a few hundred observations) than the asymptotic theory indicates. Hosmer [67] specifically addressed the small-sample, poor-separation case for a mixture of two univariate normals and concluded that in this case maximum-likelihood estimates "should be used with extreme caution or not at all." Dick and Bowden [42], Hosmer [68], and Hosmer and Dick [71] offered evidence which suggests that considerable improvement in the performance of maximum-likelihood estimates can result from including labeled observations in the samples by which the estimates are determined, particularly when the component densities are poorly separated. In fact, it is pointed out in [71] that most of the improvement occurs for small to moderate proportions of labeled observations in the sample.

In spite of the rather pessimistic comments above, maximum-likelihood estimates have fared well in comparisons with most other estimates for mixture density estimation problems. Day [36], Hasselblad [65], Tan and Chang [121], and Dick and Bowden [42] found maximum-likelihood estimates to be markedly superior to moment estimates in their investigations, especially in cases involving poorly separated component populations. (See also the comment by Hosmer [70] on the paper of Quandt and Ramsey [106].) Day [36] also remarked that minimum chi-square and Bayes estimates have less appeal than maximum-likelihood estimates,

OF POOR QUALITY

primarily because of the difficulty of obtaining them in most cases. James [73] and Ganesalingam and McLachlan [49] observed that their proportion estimates are less efficient than maximum-likelihood estimates; however, they also outlined circumstances in which their estimates might be preferred. On the other hand, as we remarked in Sec. 2.3, the moment generating function method of Quandt and Ramsey [106] provides estimates which may outperform maximum-likelihood estimates in the small-sample case (see the comment by Hosmer [70]). This method should be kept in mind as a promising alternative to the method of maximum likelihood.

4. The EM Algorithm.

We now derive the EM algorithm for general mixture density estimation problems and discuss its important general properties. As stated in the introduction, we feel that the EM algorithm for mixture density estimation problems is best regarded as a specialization of the general EM algorithm formalized by Dempster, Laird and Rubin [38] for obtaining maximum-likelihood estimates from incomplete data. Accordingly, we begin by reviewing the formulation of the general EM algorithm given in [38].

Suppose that one has a measure space \mathcal{Y} of "complete data" and a measurable map $\underline{y} \rightarrow \underline{x}(\underline{y})$ of \mathcal{Y} to a measure space \mathcal{X} of "incomplete data". Let $f(\underline{y}|\phi)$ be a member of a parametric family of probability density functions defined on \mathcal{Y} for $\phi \in \Omega$, and suppose that $g(\underline{x}|\phi)$ is a probability density function on \mathcal{X} induced by $f(\underline{y}|\phi)$. For a given $\underline{x} \in \mathcal{X}$, the purpose of the EM algorithm is to maximize the incomplete data log-likelihood $L(\phi) = \log g(\underline{x}|\phi)$ over $\phi \in \Omega$ by exploiting the relationship between $f(\underline{y}|\phi)$ and $g(\underline{x}|\phi)$. It is intended especially for applications in which the maximization of the complete data log-likelihood $\log f(\underline{y}|\phi)$ over $\phi \in \Omega$ is particularly easy.

For $\underline{x} \in \mathcal{X}$, set $\mathcal{Y}(\underline{x}) = \{\underline{y} \in \mathcal{Y} : \underline{x}(\underline{y}) = \underline{x}\}$. The conditional density $k(\underline{y}|\underline{x}, \phi)$ on $\mathcal{Y}(\underline{x})$ is given by $f(\underline{y}|\phi) = k(\underline{y}|\underline{x}, \phi)g(\underline{x}|\phi)$. For ϕ and ϕ' in Ω , one then has

$$L(\phi) = Q(\phi|\phi') - H(\phi|\phi') ,$$

where $Q(\phi|\phi') = E(\log f(\underline{y}|\phi) | \underline{x}, \phi')$ and $H(\phi|\phi') = E(\log k(\underline{y}|\underline{x}, \phi) | \underline{x}, \phi')$. The general EM algorithm of Dempster, Laird and Rubin [38] is the following: Given a current approximation ϕ^c of a maximizer of $L(\phi)$, obtain a next approximation ϕ^+ as follows:

1. E-step: Determine $Q(\phi|\phi^c)$.
2. M-step: Choose $\phi^+ \in \arg \max_{\phi \in \Omega} Q(\phi|\phi^c)$.

Here, $\arg \max_{\phi \in \Omega} Q(\phi|\phi^c)$ denotes the set of values $\phi \in \Omega$ which maximize $Q(\phi|\phi^c)$ over Ω . (Of course, this set must be nonempty for the M-step of the algorithm to be well-defined.) If this set is a singleton, then we denote its sole member in the same way and write $\phi^+ = \arg \max_{\phi \in \Omega} Q(\phi|\phi^c)$. Similar notation is used without further explanation in the sequel.

From this general description, it is not clear that the EM algorithm even deserves to be called an algorithm. However, as we indicated above, the EM algorithm is used most often in applications which permit the easy maximization of $\log f(\underline{y}|\phi)$ over $\phi \in \Omega$. In such applications, the M-step maximization of $Q(\phi|\phi^c)$ over $\phi \in \Omega$ is usually carried out with corresponding ease. In fact, as one sees in the sequel, the E-step and the M-step are usually combined into one very easily implemented step in most applications involving mixture density estimation problems. At any rate, the sense of the EM algorithm lies in the

fact that $L(\Phi^+) > L(\Phi^c)$. Indeed, the manner in which Φ^+ is determined guarantees that $Q(\Phi^+|\Phi^c) > Q(\Phi^c|\Phi^c)$; and it follows from Jensen's Inequality that $H(\Phi^+|\Phi^c) < H(\Phi^c|\Phi^c)$. (See Theorem 1 of Dempster, Laird and Rubin [38].) This fact implies that L is monotone increasing on any iteration sequence generated by the EM algorithm, which is the fundamental property of the algorithm underlying the convergence theorems given below.

To discuss the EM algorithm for mixture density estimation problems, we assume as in the preceding section that a parametric family of mixture densities of the form (1.1) is specified and that a particular $\Phi^* = (\alpha_1^*, \dots, \alpha_m^*, \rho_1^*, \dots, \rho_m^*)$ is the "true" parameter value to be estimated. In the usual way, we regard this family of densities as being associated with a statistical population which is a mixture of m component populations. The EM algorithm for a mixture density estimation problem associated with this family is derived by first interpreting the problem as one involving incomplete data and then obtaining the algorithm from its general formulation given above. The problem is interpreted as one involving incomplete data by regarding each unlabeled observation in the sample at hand as "missing" a label indicating its component population of origin.

It is instructive to consider the forms which the EM algorithm might take for mixture density estimation problems involving samples of the types introduced in the preceding section. We first illustrate in some detail the derivation of the function $Q(\Phi|\Phi')$ of the E-step of the algorithm, assuming

ORIGINAL PAGE IS
OF POOR QUALITY

for convenience that the sample at hand is a sample $S_1 = \{x_k\}_{k=1 \dots N}$ of Type 1 described in the preceding section. One can regard S_1 as a sample of incomplete data by considering each x_k to be the "known" part of an observation $Y_k = (x_k, i_k)$, where i_k is an integer between 1 and m indicating a component population of origin of x_k . For $\Phi = (\alpha_1, \dots, \alpha_m, \rho_1, \dots, \rho_m) \in \Omega$, the sample variables $\underline{x} = (x_1, \dots, x_N)$ and $\underline{y} = (y_1, \dots, y_N)$ have associated probability density functions $g(\underline{x}|\Phi) = \prod_{k=1}^N p(x_k|\Phi)$ and $f(\underline{y}|\Phi) = \prod_{k=1}^N \alpha_{i_k} p_{i_k}(x_k|\rho_{i_k})$, respectively. Then for $\Phi' = (\alpha'_1, \dots, \alpha'_m, \rho'_1, \dots, \rho'_m) \in \Omega$, the conditional density $k(\underline{y}|\underline{x}, \Phi')$ is given by

$$k(\underline{y}|\underline{x}, \Phi') = \prod_{k=1}^N \frac{\alpha'_{i_k} p_{i_k}(x_k|\rho'_{i_k})}{p(x_k|\Phi')},$$

and the function $Q(\Phi|\Phi')$, which we denote by $Q_1(\Phi|\Phi')$, is determined to be

$$\begin{aligned} Q_1(\Phi|\Phi') &= \sum_{i_1=1}^m \dots \sum_{i_N=1}^m \sum_{k=1}^N \log \alpha_{i_k} p_{i_k}(x_k|\rho_{i_k}) - \sum_{k=1}^N \frac{\alpha'_{i_k} p_{i_k}(x_k|\rho'_{i_k})}{p(x_k|\Phi')} \\ &= \sum_{i=1}^m \sum_{k=1}^N \log \alpha_i p_i(x_k|\rho_i) - \sum_{k=1}^N \frac{\alpha'_{i_k} p_{i_k}(x_k|\rho'_{i_k})}{p(x_k|\Phi')} \end{aligned} \quad (4.1)$$

OF POOR QUALITY

$$= \sum_{i=1}^m \left[\sum_{k=1}^N \frac{\alpha_i p_i(x_k | \rho_i)}{p(x_k | \Phi')} \right] \log \alpha_i + \sum_{i=1}^m \sum_{k=1}^N \log p_i(x_k | \rho_i) \frac{\alpha_i p_i(x_k | \rho_i)}{p(x_k | \Phi')}$$

For samples $S_2 = \bigcup_{i=1}^m \{y_{ik}\}_{k=1, \dots, J_i}$, $S_3 = \bigcup_{i=1}^m \{z_{ik}\}_{k=1, \dots, K_i}$, and $S_4 = \bigcup_{i=0}^m \{w_{ik}\}_{k=1, \dots, M_i}$ of Types 2, 3 and 4, one determines in a similar manner the respective functions $Q_2(\Phi | \Phi')$, $Q_3(\Phi | \Phi')$, and $Q_4(\Phi | \Phi')$ for the E-step of the EM algorithm to be

$$Q_2(\Phi | \Phi') = \sum_{i=1}^m \sum_{k=1}^{J_i} \log p_i(y_{ik} | \rho_i), \quad (4.2)$$

$$Q_3(\Phi | \Phi') = \sum_{i=1}^m K_i \log \alpha_i + \sum_{i=1}^m \sum_{k=1}^{K_i} \log p_i(z_{ik} | \rho_i), \quad (4.3)$$

$$Q_4(\Phi | \Phi') = \sum_{i=1}^m [M_i + \sum_{k=1}^{M_0} \frac{\alpha_i p_i(w_{0k} | \rho_i)}{p(w_{0k} | \Phi')}] \log \alpha_i + \quad (4.4)$$

$$+ \sum_{i=1}^m \left[\sum_{k=1}^{M_i} \log p_i(w_{ik} | \rho_i) + \sum_{k=1}^{M_0} \log p_i(w_{0k} | \rho_i) \frac{\alpha_i p_i(w_{0k} | \rho_i)}{p(w_{0k} | \Phi')} \right]$$

for $\Phi = (\alpha_1, \dots, \alpha_m, \rho_1, \dots, \rho_m)$ and $\Phi' = (\alpha'_1, \dots, \alpha'_m, \rho'_1, \dots, \rho'_m)$ in Ω . We note that $Q_2(\Phi | \Phi')$ and $Q_3(\Phi | \Phi')$ are just $L_2(\Phi)$ and (except for an additive constant) $L_3(\Phi)$ given by (3.2) and (3.3), respectively; and one might well wonder why they are of interest in this context. By way of explanation, we observe that if a sample of interest is a stochastically independent union of smaller samples, then the function for the E-step of the EM

algorithm which is appropriate for this sample is just the sum of the functions which are appropriate for the smaller samples. Thus, for example, if $S = S_1 \cup S_2 \cup S_3$ is a union of independent samples of Types 1, 2, and 3, then the function for the E-step appropriate for S is $Q(\phi|\phi') = Q_1(\phi|\phi') + Q_2(\phi|\phi') + Q_3(\phi|\phi')$, where $Q_1(\phi|\phi')$, $Q_2(\phi|\phi')$, and $Q_3(\phi|\phi')$ are given by (4.1), (4.2), and (4.3), respectively.

Having determined an appropriate function $Q(\phi|\phi')$ for the E-step of the EM algorithm as one or a sum of the functions $Q_i(\phi|\phi')$ defined above, one is likely to find that the maximization problem of the M-step has a number of attractive features. It is clear from (4.1), (4.2), (4.3), and (4.4) that this maximization problem separates into two maximization problems, the first involving the proportions $\alpha_1, \dots, \alpha_m$ alone and the second involving only the remaining parameters ρ_1, \dots, ρ_m . Since $\log \alpha_1, \dots, \log \alpha_m$ appear linearly in each function $Q_i(\phi|\phi')$ for $i \neq 2$, the first maximization problem has a unique solution if the sample is not strictly of Type 2; and this solution is easily and explicitly determined regardless of the functional forms of the component densities $p_i(x|\rho_i)$. If ρ_1, \dots, ρ_m are mutually independent variables, then the second maximization problem separates further into m component problems, each of which involves only one of the parameters ρ_i . Both these component problems and the maximization problem for the proportions alone have the appealing property that they can be regarded as "weighted" maximum-likelihood estimation problems

involving sums of logarithms weighted by posterior probabilities that sample observations belong to appropriate component populations, given the current approximate maximum-likelihood estimate of Φ^* .

To illustrate these remarks, we consider a sample $S_1 = (x_k)_{k=1, \dots, N}$ of Type 1 and assume that ρ_1, \dots, ρ_m are mutually independent variables. If $\Phi^C = (\alpha_1^C, \dots, \alpha_m^C, \rho_1^C, \dots, \rho_m^C)$ is a current approximate maximizer of the log-likelihood function $L_1(\Phi)$ given by (3.1), then one easily verifies that the next approximate maximizer $\Phi^+ = (\alpha_1^+, \dots, \alpha_m^+, \rho_1^+, \dots, \rho_m^+)$ prescribed by the M-step of the EM algorithm satisfies

$$\alpha_i^+ = \frac{1}{N} \sum_{k=1}^N \frac{\alpha_i^C p_i(x_k | \rho_i^C)}{p(x_k | \Phi^C)} \quad (4.5)$$

$$\rho_i^+ \in \arg \max_{\rho_i \in \Omega_i} \sum_{k=1}^N \log p_i(x_k | \rho_i) \frac{\alpha_i^C p_i(x_k | \rho_i^C)}{p(x_k | \Phi^C)} \quad (4.6)$$

for $i = 1, \dots, m$. Note that, as promised, each α_i^+ is uniquely and explicitly determined and each α_i^+ and ρ_i^+ is obtained as the solution of a weighted maximum-likelihood estimation problem involving a sum of logarithms multiplied by weights $\frac{\alpha_i^C p_i(x_k | \rho_i^C)}{p(x_k | \Phi^C)}$, each of which is just the posterior probability that x_k originated in the i^{th} component population, given the current approximate maximum-likelihood estimate Φ^C .

In addition to prescribing each α_j^+ and ρ_1^+ as the solution of a heuristically appealing weighted maximum-likelihood estimation problem, there are other attractions to (4.5) and (4.6). For example, (4.5) insures that the next approximate proportions α_i^+ inherit from the current approximate proportions α_i^c the property of being non-negative and summing to 1. Furthermore, although there is no guarantee that the maximization problems (4.6) will have nice properties in general, it happens that each ρ_1^+ is usually easily (even uniquely and explicitly) determined by (4.6) in most applications of interest, especially in those applications in which each component density $p_i(x|\rho_1)$ is one of the common parametric densities for which ordinary (labeled-sample) maximum-likelihood estimates of ρ_1 are uniquely and explicitly determined. As an illustration, consider the case in which some $p_i(x|\rho_1)$ is a multivariate normal density, i.e., $p_i(x|\rho_1)$ and ρ_1 are given by

$$p_i(x|\rho_1) = \frac{1}{(2\pi)^{n/2} (\det \Sigma_1)^{1/2}} e^{-1/2(x-\mu_1)^T \Sigma_1^{-1} (x-\mu_1)}, \quad \rho_1 = (\mu_1, \Sigma_1), \quad (4.7)$$

where $\mu_1 \in R^n$ and Σ_1 is a positive-definite symmetric $n \times n$ matrix. For a given $\rho_1^c = (\mu_1^c, \Sigma_1^c)$, the unique solution $\rho_1^+ = (\mu_1^+, \Sigma_1^+)$ of (4.6) is given by

$$\mu_1^+ = \left\{ \sum_{k=1}^N x_k \frac{\alpha_i^c p_i(x_k|\rho_1^c)}{p(x_k|\phi^c)} \right\} / \left\{ \sum_{k=1}^N \frac{\alpha_i^c p_i(x_k|\rho_1^c)}{p(x_k|\phi^c)} \right\} \quad (4.8)$$

$$\Sigma_i^+ = \left\{ \sum_{k=1}^N (x_k - \mu_i^+) (x_k - \mu_i^+) \frac{\alpha_i^c p_i(x_k | \phi_i^c)}{p(x_k | \Phi^c)} \right\} / \left\{ \sum_{k=1}^N \frac{\alpha_i^c p_i(x_k | \phi_i^c)}{p(x_k | \Phi^c)} \right\} \quad (4.9)$$

(The factors α_i^c have been left in the numerators and denominators of these expressions for aesthetic reasons only.) Note that Σ_i^+ is positive-definite symmetric with probability 1 if $N > n$.

What convergence properties hold for a sequence of iterates generated by applying the EM algorithm to a mixture density estimation problem? If nothing in particular is said about the parametric family of interest, then the properties which can be specified are essentially those obtained by specializing the convergence results associated with the EM algorithm for general incomplete data problems. The convergence results below are formulated so that they are valid for the EM algorithm in general. Points relating to these results which are of particular interest in the mixture density context are made in remarks following the theorems.

The first theorem is a global convergence result for sequences generated by the EM algorithm. It essentially summarizes the results of Wu [134] for the general EM algorithm and of Redner [109] for the more specialized case of the EM algorithm applied to a mixture of densities from exponential families. Similar but weaker results have been formulated for the general EM algorithm by Boyles [17]. Statements (i), (ii), and (iii) of the theorem are valid for any sequence and are

OF POOR QUALITY

stated here as a convenience because of their usefulness in applications. Statements (iv), (v), and (vi) are based on the fact reviewed earlier and reiterated in the statement of the theorem that the log-likelihood function increases monotonically on a sequence generated by the EM algorithm. Through the use of this fact, the theorem can be related to general results in optimization theory such as the convergence theorems of Zangwill [141; pages 91, 128, and 232] concerning point-to-set maps which increase an objective function. One such general result was used explicitly by Wu [134] in his study of the EM algorithm.

Theorem 4.1: Suppose that for some $\phi^{(0)} \in \Omega$, $\{\phi^{(j)}\}_{j=0,1,2,\dots}$ is a sequence in Ω generated by the EM algorithm, i.e., a sequence in Ω satisfying

$$\phi^{(j+1)} \in \arg \max_{\phi \in \Omega} Q(\phi | \phi^{(j)}), \quad j = 0, 1, 2, \dots,$$

where $Q(\phi | \phi')$ is the function determined in the E-step of the EM algorithm. Then the log-likelihood function $L(\phi)$ increases monotonically on $\{\phi^{(j)}\}_{j=0,1,2,\dots}$ to a (possibly infinite) limit L^* . Furthermore, denoting the set of limit points of $\{\phi^{(j)}\}_{j=0,1,2,\dots}$ in Ω by \mathcal{L} , one has the following:

- (i) \mathcal{L} is a closed set in Ω .
- (ii) If $\{\phi^{(j)}\}_{j=0,1,2,\dots}$ is contained in a compact subset of Ω , then \mathcal{L} is compact.
- (iii) If $\{\phi^{(j)}\}_{j=0,1,2,\dots}$ is contained in a compact subset of Ω and $\lim_{j \rightarrow \infty} \|\phi^{(j+1)} - \phi^{(j)}\| = 0$ for a norm $\|\cdot\|$ on Ω , then \mathcal{L}

OF POOR QUALITY

is connected as well as compact.

(iv) If $L(\Phi)$ is continuous in Ω and $\mathcal{L} \neq \emptyset$, then L^* is finite and $L(\hat{\Phi}) = L^*$ for $\hat{\Phi} \in \mathcal{L}$.

(v) If $Q(\Phi|\Phi')$ and $H(\Phi|\Phi') = Q(\Phi|\Phi') - L(\Phi)$ are continuous in Φ and Φ' in Ω , then each $\hat{\Phi} \in \mathcal{L}$ satisfies $\hat{\Phi} \in \arg \max_{\Phi \in \Omega} Q(\Phi|\hat{\Phi})$.

(vi) If $Q(\Phi|\Phi')$ and $H(\Phi|\Phi')$ are continuous in Φ and Φ' in Ω and differentiable in Φ at $\Phi = \Phi' = \hat{\Phi} \in \mathcal{L}$, then $L(\Phi)$ is differentiable at $\Phi = \hat{\Phi}$ and the likelihood equations $\nabla_{\Phi} L(\Phi) = 0$ are satisfied by $\Phi = \hat{\Phi}$.

Proof: The monotonicity of $L(\Phi)$ on $\{\Phi^{(j)}\}_{j=0,1,2,\dots}$ has already been established; the existence of a (possibly infinite) limit L^* follows. Statement (i) holds since closedness is a general property of sets of limit points. To obtain (ii), note that if $\{\Phi^{(j)}\}_{j=0,1,2,\dots}$ is contained in a compact subset of Ω , then \mathcal{L} is a closed subset of this compact subset and, hence, is compact. To prove (iii), suppose that $\{\Phi^{(j)}\}_{j=0,1,2,\dots}$ is contained in a compact subset of Ω , that $\lim_{j \rightarrow \infty} \|\Phi^{(j+1)} - \Phi^{(j)}\| = 0$, and that \mathcal{L} is not connected. Since \mathcal{L} is compact, there is a minimal distance between distinct components of \mathcal{L} ; and the fact that $\lim_{j \rightarrow \infty} \|\Phi^{(j+1)} - \Phi^{(j)}\| = 0$ implies that there is an infinite subsequence of $\{\Phi^{(j)}\}_{j=0,1,2,\dots}$ whose members are bounded away from \mathcal{L} . This subsequence lies in a compact set, and so it has limit points. Since these limit points cannot be in \mathcal{L} , one has a

contradiction.

Statement (iv) follows immediately from the monotonicity of $L(\Phi)$ on $\{\Phi^{(j)}\}_{j=0,1,2,\dots}$. To prove (v), suppose that $Q(\Phi|\Phi')$ and $H(\Phi|\Phi')$ are continuous in Φ and Φ' in Ω and that one can find some $\hat{\Phi} \in \mathcal{L}$ and $\Phi \in \Omega$ for which $Q(\Phi|\hat{\Phi}) > Q(\hat{\Phi}|\hat{\Phi})$. Then for every j ,

$$\begin{aligned} L(\Phi^{(j+1)}) &= Q(\Phi^{(j+1)}|\Phi^{(j)}) - H(\Phi^{(j+1)}|\Phi^{(j)}) \\ &> Q(\Phi|\Phi^{(j)}) - H(\Phi^{(j)}|\Phi^{(j)}) \end{aligned}$$

by the M -step determination of $\Phi^{(j+1)}$ and Jensen's inequality. Since $Q(\Phi|\Phi')$ and $H(\Phi|\Phi')$ are continuous, it follows by taking limits along a subsequence converging to $\hat{\Phi}$ that

$$\begin{aligned} L^* &> Q(\Phi|\hat{\Phi}) - H(\hat{\Phi}|\hat{\Phi}) \\ &> Q(\hat{\Phi}|\hat{\Phi}) - H(\hat{\Phi}|\hat{\Phi}) - L(\hat{\Phi}) - L^* , \end{aligned}$$

which is a contradiction. To establish (vi), suppose that $Q(\Phi|\Phi')$ and $H(\Phi|\Phi')$ are continuous in Φ and Φ' in Ω and differentiable in Φ at $\Phi = \Phi' = \hat{\Phi} \in \mathcal{L}$. Then $L(\Phi) = Q(\Phi|\hat{\Phi}) - H(\Phi|\hat{\Phi})$ is differentiable at $\Phi = \hat{\Phi}$; and, since $\hat{\Phi} \in \arg \max_{\Phi \in \Omega} Q(\Phi|\hat{\Phi})$ by (v) and $\hat{\Phi} \in \arg \max_{\Phi \in \Omega} H(\Phi|\hat{\Phi})$ by Jensen's inequality, one has $\nabla_{\Phi} L(\hat{\Phi}) = 0$. This completes the proof.

Statement (iii) of Theorem 4.1 has precedent in such results as Theorem 28.1 of Ostrowski [96]. It is usually satisfied in

practice, especially in the mixture density context. Indeed, it often happens that each ϕ^+ is uniquely determined by the EM algorithm as a function of ϕ^c which is continuous in Ω . For example, one sees that $\alpha_1^+, \dots, \alpha_m^+$ are determined in this way by (4.5) whenever each $p_i(x|\rho_i)$ depends continuously on ρ_i . In addition, each ρ_i is likely to be determined in this way by (4.6) whenever each $p_i(x|\rho_i)$ is one of the common parametric densities for which ordinary maximum-likelihood estimates are determined as a continuous function of ρ_i ; see, for example, (4.8) and (4.9). If ϕ^+ is determined in this way from ϕ^c and if the conditions of (v) are also satisfied, then each $\hat{\phi} \in \mathcal{L}$ is a fixed point of a continuous function. It follows that if in addition $\{\phi^{(j)}\}_{j=0,1,2,\dots}$ is contained in a compact subset of Ω , then the elements of a "tail" sequence $\{\phi^{(j)}\}_{j=J,J+1,\dots}$ can all be made to lie arbitrarily close to the compact set \mathcal{L} by taking J sufficiently large and, hence, $\lim_{j \rightarrow \infty} \|\phi^{(j+1)} - \phi^{(j)}\| = 0$ by the uniform continuity near \mathcal{L} of the function determining $\phi^{(j+1)}$ from $\phi^{(j)}$.

It is useful to expand a little on the interpretation of statement (vi) in the mixture density context. Assuming that each $p_i(x|\rho_i)$ is differentiable with respect to ρ_i , that the parameters ρ_i are unconstrained in Ω_i and mutually independent, and, for convenience, that the sample of interest is of Type 1, one can reasonably interpret the likelihood equations $\nabla_{\phi} L(\phi) = 0$ in the sense of (3.9) at a point $\hat{\phi} = (\hat{\alpha}_1, \dots, \hat{\alpha}_m, \hat{\rho}_1, \dots, \hat{\rho}_m) \in \mathcal{L}$ which is such that each $\hat{\alpha}_i$ is

positive. Now it is certainly possible for some $\hat{\alpha}_1$ to be zero for $\hat{\phi} \in \mathcal{L}$, in which case (3.9) might not be valid. Fortunately, (3.5) and (3.8) provide a better interpretation than (3.9) of the likelihood equations in the mixture density context which is valid whether each $\hat{\alpha}_1$ is positive or not. Indeed, if the conditions of (v) hold, then it follows from (4.5) that the equations (3.8) are satisfied on \mathcal{L} . Thus in the mixture density context under the present assumptions, (vi) should be replaced with the following:

(vi)' If $Q(\phi|\phi')$ and $H(\phi|\phi')$ are continuous in ϕ and ϕ' in Ω and differentiable in ρ_1, \dots, ρ_m at $\phi = \phi' = \hat{\phi} \in \mathcal{L}$, then $L(\phi)$ is differentiable in ρ_1, \dots, ρ_m at $\phi = \hat{\phi}$ and the likelihood equations (3.5) and (3.8) are satisfied by $\hat{\phi}$.

To illustrate the application of Theorem 4.1, we consider the problem of estimating the proportions in a mixture under the assumption that each component density $p_i(x|\rho_i)$ is known (and denoted for the present purposes by $p_i(x)$ for simplicity). The theorem below is a global convergence result for the EM algorithm applied to this problem. For convenience in presenting the theorem, it is assumed that the sample at hand is a sample $S_1 = \{x_k\}_{k=1, \dots, N}$ of Type 1. Similar results hold for other cases in which the sample at hand is one or a union of the types considered in the preceding section. For this problem, one has simply $\phi = (\alpha_1, \dots, \alpha_m)$; and it is, of course, always understood that $\sum_{i=1}^m \alpha_i = 1$ and $\alpha_i > 0$, $i = 1, \dots, m$, for all such ϕ .

We remark that the condition of the theorem on the matrix of second derivatives of $L_1(\Phi)$ is quite reasonable. This matrix is always defined and negative semi-definite whenever $p(x_k|\Phi) \neq 0$ for $k = 1, \dots, N$; and if $p_1(x), \dots, p_m(x)$ are linearly independent non-vanishing functions on the support of the underlying measure on R^n appropriate for p , then with probability 1 it is defined and negative definite for all Φ whenever N is sufficiently large.

Theorem 4.2: Suppose that the matrix of second derivatives of $L_1(\Phi)$ is defined and negative definite for all Φ . Then there is a unique maximum-likelihood estimate; and for any $\phi^{(0)} = (\alpha_1^{(0)}, \dots, \alpha_m^{(0)})$ with $\alpha_i^{(0)} > 0$ for $i = 1, \dots, m$, the sequence $\{\phi^{(j)} = (\alpha_1^{(j)}, \dots, \alpha_m^{(j)})\}_{j=0,1,2,\dots}$ generated by the EM algorithm, i.e., determined inductively by

$$\alpha_i^{(j+1)} = \frac{1}{N} \sum_{k=1}^N \frac{\alpha_i^{(j)} p_i(x_k)}{p(x_k|\phi^{(j)})}, \quad i = 1, \dots, m,$$

converges to the maximum-likelihood estimate.

Proof: It follows from Theorem 4.1 and the subsequent remarks that the set of limit points of $\{\phi^{(j)}\}_{j=0,1,2,\dots}$ is a compact, connected subset of the simplex of proportion vectors Φ on which the likelihood equations (3.8) are satisfied. Since the matrix of second derivatives of $L_1(\Phi)$ is negative definite, $L_1(\Phi)$ is strictly concave. It follows that there is a unique

OF POOR QUALITY

maximum-likelihood estimate and, furthermore, that the likelihood equations (3.8) have at most one solution on the interior of each face of the proportion simplex. Consequently, each component of the set of solutions of the likelihood equations consists of a single point; and $\{\phi^{(j)}\}_{j=0,1,2,\dots}$ must converge to one such point. But if $\{\phi^{(j)}\}_{j=0,1,2,\dots}$ is convergent, then its limit must be the maximum-likelihood estimate by Theorem 2 of Peters and Coberly [100] or Theorem 1 of Peters and Walker [102].

Despite the usefulness of Theorem 4.1 in characterizing the set of limit points of an iteration sequence generated by the EM algorithm, it leaves unanswered the questions of whether such a sequence converges at all and, if it does, whether it converges to a maximum-likelihood estimate. In an attempt to provide reasonable sufficient conditions under which the answer to these questions is "yes", we offer the local convergence theorem below.

Theorem 4.3: Suppose that Conditions 1 through 4 of Section 3 are satisfied in Ω , and let Ω' be a compact subset of Ω which contains ϕ^* in its interior and which is such that $p(x|\phi) = p(x|\phi^*)$ almost everywhere in x for $\phi \in \Omega'$ only if $\phi = \phi^*$. Suppose further that with probability 1, the function $Q(\phi|\phi')$ of the E-step of the EM algorithm is continuous in ϕ and ϕ' in Ω' and both $Q(\phi|\phi')$ and the log-likelihood function $L(\phi)$ are differentiable in ϕ for ϕ and ϕ' in Ω' whenever N is sufficiently large. Finally, for $\phi^{(0)}$ in Ω' , denote by $\{\phi^{(j)}\}_{j=0,1,2,\dots}$ a sequence generated by the EM

ORIGINAL
OF POOR QUALITY.

algorithm in Ω' , i.e., a sequence in Ω' satisfying

$$\phi^{(j+1)} \in \arg \max_{\phi \in \Omega'} Q(\phi | \phi^{(j)}), \quad j = 0, 1, 2, \dots$$

Then with probability 1, whenever N is sufficiently large, the unique strongly consistent maximum-likelihood estimate ϕ^N is well-defined in Ω' and $\phi^N = \lim_{j \rightarrow \infty} \phi^{(j)}$ whenever $\phi^{(0)}$ is sufficiently near ϕ^N .

Proof: It follows from Theorems 3.1 and 3.2 that with probability 1, N can be taken sufficiently large that the unique strongly consistent maximum-likelihood estimate ϕ^N is well-defined, lies in the interior of Ω' , and is the unique maximizer of $L(\phi)$ in Ω' . Also with probability 1, we can assume that N is sufficiently large that $Q(\phi | \phi')$ is continuous in ϕ and ϕ' in Ω' and $Q(\phi | \phi')$ and $L(\phi)$ are differentiable in ϕ for ϕ and ϕ' in Ω' . Since $L(\phi)$ is continuous, one can find a neighborhood Ω'' of ϕ^N of the form

$$\Omega'' = \{\phi \in \Omega' : L(\phi) > L(\phi^N) - \epsilon\}$$

for some $\epsilon > 0$ which lies in the interior of Ω' and which is such that ϕ^N is the only solution of the likelihood equations contained in it. If $\phi^{(0)}$ lies in Ω'' , then $\{\phi^{(j)}\}_{j=0,1,2,\dots}$ must also lie in Ω'' since $L(\phi)$ is monotone increasing on $\{\phi^{(j)}\}_{j=0,1,2,\dots}$. It follows that each limit point of $\{\phi^{(j)}\}_{j=0,1,2,\dots}$ lies in Ω'' and, by statement (v1) of Theorem

4.1, also satisfies the likelihood equations. Since ϕ^N is the only solution of the likelihood equations in Ω^n , one concludes that $\phi^N = \lim_{j \rightarrow \infty} \phi^{(j)}$.

As in the case of Theorem 4.1, Theorem 4.3 is stated so that it is valid for the EM algorithm in general. It should be noted, however, that Theorem 4.3 makes heavy use of Theorems 3.1 and 3.2 as well as Theorem 4.1; and so for mixture density estimation problems, it pertains as it stands, strictly speaking, to the case to which Theorems 3.1 and 3.2 apply, namely that in which the sample at hand is of Type 1 and $L(\Phi) = L_1(\Phi)$ given by (3.1) and $Q(\Phi|\Phi') = Q_1(\Phi|\Phi')$ given by (4.1). Of course, Theorems 3.1 and 3.2 and, therefore, Theorem 4.3 can be modified to treat mixture density estimation problems involving samples of other types.

5. The EM Algorithm for Mixtures of Densities from Exponential Families

Almost all mixture density estimation problems which have been studied in the literature involve mixture densities whose component densities are members of exponential families. As it happens, the EM algorithm is especially easy to implement on problems involving densities of this type. Indeed, in an application of the EM algorithm to such a problem, each successive approximate maximum-likelihood estimate ϕ^+ is uniquely and explicitly determined from its predecessor ϕ^c , almost always in a continuous manner. Furthermore, a sequence of iterates produced by the EM algorithm on such a problem is likely to have relatively nice convergence properties.

In this section, we first determine the special form which the EM algorithm takes for mixtures of densities from exponential families. We then look into the desirable properties of the algorithm and sequences generated by it which are apparent from this form. Finally, we discuss several specific examples of the EM algorithm for component densities from exponential families which are commonly of interest.

A very brief discussion of exponential families of densities is in order. For an elaboration on the topics touched on here, the reader is referred to the book of Barndorff Nielsen [6]. A parametric family of densities $q(x|\theta)$, $\theta \in \tilde{\Omega} \subseteq R^k$, on R^n is said to be an exponential family if its members have the form

$$q(x|\theta) = a(\theta)^{-1} b(x) e^{\theta^T t(x)}, \quad x \in R^n, \quad (5.1)$$

where $b: R^n \rightarrow R^1$, $t: R^n \rightarrow R^r$, and $a(\theta)$ is given by

$$a(\theta) = \int_{R^n} b(x) e^{\theta^T t(x)} d\mu$$

for an appropriate underlying measure μ on R^n . It is, of course, assumed that $b(x) > 0$ for all $x \in R^n$ and that $a(\theta) < \infty$ for $\theta \in \tilde{\eta}$. Note that every member of an exponential family has the same support in R^n , namely that of the function $b(x)$.

The representation (5.1) of the members of an exponential family, in which the parameter θ appears linearly in the argument of the exponential function, is called the "natural" parametrization; and θ is called the "natural" parameter. If the set $\tilde{\eta}$ is open and convex and if the component functions of $t(x)$ together with the function which is identically 1 on R^n are linearly independent functions on the intersection of the supports of $b(x)$ and μ , then there is another parametrization of the members of the family, called the "expectation" or "mean value" parametrization, in terms of the "expectation" parameter

$$\rho = E(t(X)|\theta) = \int_{R^n} t(x) q(x|\theta) d\mu.$$

Indeed, under these conditions on $\tilde{\eta}$ and $t(x)$, one can show that

$$[E(t(X)|\theta') - E(t(X)|\theta)]^T(\theta' - \theta) > 0$$

whenever $\theta' \neq \theta$; and it follows that the assignment $\theta \rightarrow \rho = E(t(X)|\theta)$ is one-to-one and onto from $\tilde{\Omega}$ to an open set $\Omega \subseteq R^k$. In fact, the correspondence $\theta \leftrightarrow \rho = E(t(X)|\theta)$ is a both-ways continuously differentiable mapping between $\tilde{\Omega}$ and Ω . (See Barndorff-Nielsen [6; p. 121].) So under these conditions on $\tilde{\Omega}$ and $t(x)$, one can represent the members of the family as

$$p(x|\rho) = q(x|\theta(\rho)) = a(\rho)^{-1} b(x) e^{\theta(\rho)^T t(x)}, \quad x \in R^n, \quad (5.2)$$

for $\rho \in \Omega$, where $\theta(\rho)$ satisfies $\rho = E(t(X)|\theta(\rho))$ and $a(\theta(\rho))$ is written as $a(\rho)$ for convenience. Note that $p(x|\rho)$ is continuously differentiable in ρ , since $q(x|\theta)$ is continuously differentiable in θ and $\theta(\rho)$ is continuously differentiable in ρ .

Now suppose that a parametric family of mixture densities of the form (1.1) is given, with $\Phi^* = (\alpha_1^*, \dots, \alpha_m^*, \rho_1^*, \dots, \rho_m^*)$ the "true" parameter value to be estimated; and suppose that each component density $p_i(x|\rho_i)$ is a member of an exponential family. Specifically, we assume that each $p_i(x|\rho_i)$ has the "expectation" parametrization for $\rho_i \in \Omega_i \subseteq R^{n_i}$ given by

$$p_i(x|\rho_i) = a_i(\rho_i)^{-1} b_i(x) e^{\theta_i(\rho_i)^T t_i(x)}, \quad x \in R^n,$$

where $b_i : R^n \rightarrow R^1$, $t_i : R^n \rightarrow R^{n_i}$, $a_i(\rho_i)$ is given by

$$a_i(\rho_i) = \int_{R^n} b_i(x) e^{\theta_i(\rho_i)^T t_i(x)} d\mu,$$

and $\theta_i : \Omega_i \rightarrow R^{n_i}$. Here, μ is a measure on R^n appropriate for the mixture density $p(x|\Phi)$; and it is understood that $b_i(x) > 0$ for $x \in R^n$ and that $a_i(\rho_i) < \infty$ for $\rho_i \in \Omega_i$. It is also assumed that the component functions of $t_i(x)$ together with the function which is identically 1 on R^n are linearly independent on the intersection of the supports of $b_i(x)$ and μ and that the assignment $\rho_i \rightarrow \theta_i(\rho_i)$ maps Ω_i to a convex open set $\tilde{\Omega}_i \subseteq R^{n_i}$ in a one-to-one, onto way so that $\theta_i(\rho_i)$ is the unique solution in R^{n_i} of $\rho_i = E(t(X)|\theta_i)$ for $\rho_i \in \Omega_i$. These assumptions allow us to make use of the "natural" parametrization of the family to which $p_i(x|\rho_i)$ belongs using the "natural" parameter $\theta_i = \theta_i(\rho_i)$.

To investigate the special form and properties of the EM algorithm for the given family of mixture densities, we assume that ρ_1, \dots, ρ_m are mutually independent variables and consider for convenience a sample $S_1 = \{x_k\}_{k=1, \dots, N}$ of Type 1. (A discussion similar to the following is valid mutatis mutandis for samples of other types.) If $\Phi^C = (\alpha_1^C, \dots, \alpha_m^C, \rho_1^C, \dots, \rho_m^C)$ is a current approximate maximizer of the log-likelihood function $L_1(\Phi)$ given by (3.1), then the next approximate maximizer $\Phi^+ = (\alpha_1^+, \dots, \alpha_m^+, \rho_1^+, \dots, \rho_m^+)$ prescribed by the M-step of the EM algorithm satisfies (4.5) and (4.6). For $i = 1, \dots, m$, what ρ_i^+ satisfy (4.6)? If one replaces each $p_i(x_k|\rho_i)$ in the sum in

(4.6) by its expression in the "natural" parameter θ_1 , differentiates with respect to θ_1 , equates the sum of derivatives to zero, and finally restores the "expectation" parametrization, then one sees that the unique ρ_1^+ which satisfies (4.6) is given explicitly by

$$\rho_1^+ = \left(\sum_{k=1}^N t_1(x_k) \frac{\alpha_1^c p_1(x_k | \rho_1^c)}{p(x_k | \Phi^c)} \right) / \left(\sum_{k=1}^N \frac{\alpha_1^c p_1(x_k | \rho_1^c)}{p(x_k | \Phi^c)} \right). \quad (5.3)$$

(As in the case of (4.8) and (4.9), the factors α_1^c are left in the numerator and denominator for aesthetic reasons only.)

Not only are (4.5) and (5.3) easily evaluated and heuristically appealing formulas for determining Φ^+ from Φ^c , they also provide the key to a global convergence analysis of iteration sequences generated by the EM algorithm in the case at hand which goes beyond Theorem 4.1. Theorem 5.1 below summarizes such an analysis. In order to make the theorem complete and self-contained, some of the general conclusions of Theorem 4.1 are repeated; its statement.

Theorem 5.1: Suppose that $\{\Phi^{(j)} = (\alpha_1^{(j)}, \dots, \alpha_m^{(j)}, \rho_1^{(j)}, \dots, \rho_m^{(j)})\}_{j=0,1,2,\dots}$ is a sequence in Ω generated by the EM iteration (4.5) and (5.3). Then $L_1(\Phi)$ increases monotonically on $\{\Phi^{(j)}\}_{j=0,1,2,\dots}$ to a (possibly infinite) limit L^* . Furthermore, for each i , $\{\rho_i^{(j)}\}_{j=1,2,\dots}$ is contained in the convex hull of $\{t_1(x_k)\}_{k=1,\dots,N}$. Consequently, the set $\bar{\mathcal{L}}$ of all limit

points of $\{\phi^{(j)}\}_{j=0,1,2,\dots}$ is compact; and the likelihood equations (3.5) and (3.8) are satisfied on $\mathcal{L} = \bar{\mathcal{L}} \cap \Omega$. If $\mathcal{L} \neq \mathcal{Q}$, then L^* is finite and each $\hat{\phi} \in \mathcal{L}$ satisfies $L(\hat{\phi}) = L^*$ and is a fixed-point of the EM iteration. Finally, if $\mathcal{L} = \bar{\mathcal{L}} \subseteq \Omega$, then \mathcal{L} is connected as well as compact.

Remark: If the convex hull of $\{t_i(x_k)\}_{k=1,\dots,N}$ is contained in Ω_i for each i , then $\mathcal{Q} = \bar{\mathcal{L}} \subseteq \Omega$ and all of the conditional conclusions of Theorem 5.1 hold. The convex hull of $\{t_i(x_k)\}_{k=1,\dots,N}$ is indeed contained in Ω_i for each i in many (but not all) applications. (See the examples at the end of this section.)

Proof: One sees from (5.3) that for each i , ρ_i^+ is always a convex combination of the values $\{t_i(x_k)\}_{k=1,\dots,N}$, and it follows that $\{\rho_i^{(j)}\}_{j=0,1,2,\dots}$ is contained in the convex hull of $\{t_i(x_k)\}_{k=1,\dots,N}$ for each i . Since these convex hulls are compact sets, one concludes that $\bar{\mathcal{L}}$ is compact.

Now each density $p_1(x|\rho_i)$ is continuously differentiable in ρ_i on Ω_i , and so it is clear from (3.1) and (4.1) that $L_1(\Phi)$ and $Q_1(\Phi|\Phi')$ are continuous in Φ and Φ' and differentiable in ρ_1, \dots, ρ_m in Ω . Furthermore, it is apparent from (4.5) and (5.3) that Φ^+ depends continuously on Φ^c ; and one sees from the discussion following Theorem 4.1 that $\lim_{j \rightarrow \infty} \|\Phi^{(j+1)} - \Phi^{(j)}\| = 0$ if $\mathcal{L} = \bar{\mathcal{L}} \subseteq \Omega$. In light of these points, one verifies the remaining conclusions of Theorem 5.1 via

a straightforward application of Theorem 4.1 (including statement (vi)'); and the proof is complete.

One can also exploit (4.5) and (5.3) to obtain a local convergence result which goes beyond Theorem 4.3 for mixture density estimation problems of the type now under consideration. Theorem 5.2 and its proof below provide not only a stronger local convergence statement than Theorem 4.3 for a sequence of iterates produced by the EM algorithm but also a means of both quantifying the speed of convergence of the sequence and gaining insight into properties of the mixture density which affect the speed of convergence. This theorem is essentially the generalization of Redner [108] of the local convergence results of Peters and Walker [101] for mixtures of multivariate normal densities, and its proof closely parallels the proofs of those results.

Theorem 5.2: Suppose that the Fisher information matrix $I(\Phi)$ given by (2.5.1) is positive-definite at Φ^* and that $\Phi^* = (\alpha_1^*, \dots, \alpha_m^*, \rho_1^*, \dots, \rho_m^*)$ is such that $\alpha_i^* > 0$ for $i = 1, \dots, m$. For $\Phi^{(0)}$ in Ω , denote by $\{\Phi^{(j)}\}_{j=0,1,2,\dots}$ the sequence in Ω generated by the EM iteration (4.5) and (5.3). Then with probability 1, whenever N is sufficiently large, the unique strongly consistent solution $\Phi^N = (\alpha_1^N, \dots, \alpha_m^N, \rho_1^N, \dots, \rho_m^N)$ of the likelihood equations is well-defined and there is a certain norm $\|\cdot\|$ on Ω in which $\{\Phi^{(j)}\}_{j=0,1,2,\dots}$ converges linearly to Φ^N whenever $\Phi^{(0)}$ is sufficiently near Φ^N , i.e., there is a constant λ , $0 < \lambda < 1$, for which

$$\|\phi^{(j+1)} - \phi^N\| < \lambda \|\phi^{(j)} - \phi^N\|, \quad j = 0, 1, 2, \dots \quad (5.4)$$

whenever $\phi^{(0)}$ is sufficiently near ϕ^N .

Proof: One sees not only that Condition 2 of Section 3 is satisfied but also, by restricting Ω to be a small neighborhood of ϕ^* if necessary, that Condition 1 holds as well for the family of mixture densities under consideration. It follows from Theorem 3.1 that with probability 1, ϕ^N is well-defined whenever N is sufficiently large and converges to ϕ^* as N approaches infinity. It must be shown that with probability 1, whenever N is sufficiently large, there is a norm $\|\cdot\|$ on Ω and a constant λ , $0 < \lambda < 1$, such that (5.4) holds whenever $\phi^{(0)}$ is sufficiently near ϕ^N . Toward this end, we observe that the EM iteration of interest is actually a functional iteration $\phi^+ = G(\phi^C)$, where $G(\phi)$ is the function defined in the obvious way by (4.5) and (5.3). Note that $G(\phi)$ is continuously differentiable in Ω and that any $\hat{\phi}$ which satisfies the likelihood equations (3.5) and (3.8) (and $\hat{\phi} = \phi^N$ in particular) is a fixed point of $G(\phi)$, i.e., $\hat{\phi} = G(\hat{\phi})$. Consequently one can write

$$\begin{aligned} \phi^+ - \phi^N &= G(\phi^C) - G(\phi^N) \\ &= G'(\phi^N)(\phi^C - \phi^N) + o(\|\phi^C - \phi^N\|^2) \end{aligned} \quad (5.5)$$

for any ϕ^C in Ω near ϕ^N and any norm $\|\cdot\|$ on Ω , where

$G'(\Phi^N)$ denotes the Frechet derivative of $G(\Phi)$ evaluated at Φ^N . (For questions concerning Frechet derivatives, see, for example, Luenberger [83].) We will complete the proof by showing that with probability 1, $G'(\Phi^N)$ converges as N approaches infinity to an operator which has operator norm less than 1 with respect to a certain norm on Ω .

For convenience, we introduce the following notation for $i = 1, \dots, m$:

$$\beta_i(x) = p_i(x|\rho_i^N)/p(x|\Phi^N),$$

$$\sigma_i = \int_{R^n} [t_i(x) - \rho_i^N][t_i(x) - \rho_i^N]^T p_i(x|\rho_i^N) d\mu,$$

$$\gamma_i(x) = \sigma_i^{-1}[t_i(x) - \rho_i^N].$$

Regarding an element $\Phi \in \Omega$ as an $(m + \sum_{i=1}^m n_i)$ -vector in the natural way, one can show via a very tedious calculation that $G'(\Phi^N)$ has the $(m + \sum_{i=1}^m n_i) \times (m + \sum_{i=1}^m n_i)$ matrix representation

$$G'(\Phi^N) = \text{diag}(1, \dots, 1, \frac{1}{N} \sum_{k=1}^N \beta_1(x_k) \sigma_1 \gamma_1(x_k) \gamma_1(x_k)^T, \dots,$$

$$\frac{1}{N} \sum_{k=1}^N \beta_m(x_k) \sigma_m \gamma_m(x_k) \gamma_m(x_k)^T)$$

$$= \text{diag}(\alpha_1^N, \dots, \alpha_m^N, \alpha_1^{N-1} \sigma_1, \dots, \alpha_m^{N-1} \sigma_m) \left(\frac{1}{N} \sum_{k=1}^N v(x_k) v(x_k)^T \right),$$

ORIGINAL PAGE IS
OF POOR QUALITY

where

$$V(x) = (\beta_1(x), \dots, \beta_m(x), \alpha_1 \beta_1(x) \gamma_1(x)^T, \dots, \alpha_m \beta_m(x) \gamma_m(x)^T)^T.$$

(Since Φ^N converges to Φ^* with probability 1, we can assume that with probability 1, each α_i^N is non-zero whenever N is sufficiently large.) It follows from the Strong Law of Large Numbers (see Loève [82]) that with probability 1, $G'(\Phi^N)$ converges to $E[G'(\Phi^*)] = I - QR$, where

$$Q = \text{diag}(\alpha_1^*, \dots, \alpha_m^*, \alpha_1^{*-1} \sigma_1, \dots, \alpha_m^{*-1} \sigma_m)$$

and

$$R = \int_{R^n} V(x)V(x)^T p(x|\Phi^*) d\mu.$$

It is understood that in these expressions defining Q and R , Φ^N and its components have been replaced by Φ^* and its components.

It remains to be shown that there is a norm $\|\cdot\|$ on Ω with respect to which $E[G'(\Phi^*)]$ has operator norm less than 1. Now Q and R are positive-definite symmetric operators with respect to the Euclidean inner product, and so QR is a positive-definite symmetric operator with respect to the inner product $\langle \cdot, \cdot \rangle$ defined by $\langle U, W \rangle = U^T Q^{-1} W$ for $(m + \sum_{i=1}^m n_i)$ -vectors U and W . Consequently, to prove the theorem, it suffices to show that the operator norm of QR with respect to the norm defined by $\langle \cdot, \cdot \rangle$ is less than 1.

Since QR is positive-definite symmetric with respect to $\langle \cdot, \cdot \rangle$, we need only show that $\langle U, QRU \rangle < \langle U, U \rangle$ for an arbitrary $(m + \sum_{i=1}^m n_i)$ -vector $U = (\delta_1, \dots, \delta_m, \psi_1^T, \dots, \psi_m^T)^T$. One has

$$\begin{aligned} \langle U, QRU \rangle &= U^T R U \\ &= \int_{R^n} \left\{ \sum_{i=1}^m \delta_i \beta_i(x) + \sum_{i=1}^m \psi_i^T [\alpha_i^* \beta_i(x) \gamma_i(x)] \right\}^2 p(x | \phi^*) d\mu \\ &= \int_{R^n} \left\{ \sum_{i=1}^m [\delta_i \alpha_i^{*-1} + \psi_i^T \gamma_i(x)] \alpha_i^* \beta_i(x) \right\}^2 p(x | \phi^*) d\mu \\ &< \int_{R^n} \sum_{i=1}^m [\delta_i \alpha_i^{*-1} + \psi_i^T \gamma_i(x)]^2 \alpha_i^* p_i(x | \rho_i^*) d\mu . \end{aligned}$$

The inequality is a consequence of the following corollary of Schwarz's inequality: If $\eta_i > 0$ for $i = 1, \dots, m$ and if $\sum_{i=1}^m \eta_i = 1$, then $(\sum_{i=1}^m \xi_i \eta_i)^2 < \sum_{i=1}^m \xi_i^2 \eta_i$ for all $(\xi_i)_{i=1, \dots, m}$. Since

$$\int_{R^n} \gamma_i(x) p_i(x | \rho_i^*) d\mu = 0 ,$$

one continues to obtain

$$\begin{aligned} \langle U, QRU \rangle &< \int_{R^n} \sum_{i=1}^m [\delta_i^2 \alpha_i^{*-2} + \psi_i^T \gamma_i(x) \gamma_i(x)^T \psi_i] \alpha_i^* p_i(x | \rho_i^*) d\mu \\ &= \langle U, U \rangle . \end{aligned}$$

This completes the proof.

It is instructive to explore the consequences of Theorem 5.2 and the developments in its proof. One sees from the proof of Theorem 5.2 that with probability 1, for sufficiently large N and $\phi^{(0)}$ sufficiently near ϕ^N , an inequality (5.4) holds in which $\|\cdot\|$ is the norm determined by the inner product $\langle \cdot, \cdot \rangle$ defined in the proof and λ is arbitrarily close to the operator norm of $E[G'(\phi^*)] = I - QR$ determined by $\|\cdot\|$. Since QR is positive-definite symmetric with respect to $\langle \cdot, \cdot \rangle$, this operator norm is just $\rho(I-QR)$, the spectral radius or largest absolute value of an eigenvalue of $I - QR$. Thus with probability 1, one can obtain a quantitative estimate of the speed of convergence to ϕ^N for large N of a sequence generated by the EM iteration (4.5) and (5.3) by taking $\lambda \approx \rho(I-QR)$ in (5.4).

What properties of the mixture density influence the speed of convergence to ϕ^N of an EM iteration sequence for large N ? Careful inspection shows that if the component populations in the mixture are "well separated" in the sense that

$$\frac{p_i(x|\phi_i^*)}{p(x|\phi^*)} \frac{p_j(x|\phi_j^*)}{p(x|\phi^*)} \approx 0 \text{ for } x \in R^n, \text{ whenever } i \neq j,$$

then $QR \approx I$. It follows that $\rho(I-QR) \approx 0$, and an EM iteration sequence which converges to ϕ^N exhibits rapid linear convergence. On the other hand, if the component populations in the mixture are "poorly separated" in the sense that, say, the

i^{th} and j^{th} component populations are such that

$$\frac{p_i(x|\phi_i^*)}{p(x|\phi^*)} \approx \frac{p_j(x|\phi_j^*)}{p(x|\phi^*)} \text{ for } x \in R^N,$$

then R is nearly singular. One concludes that $\rho(I-QR) \approx 1$ in this case and that slow linear convergence of an EM iteration sequence to ϕ^N can be expected.

In the interest of obtaining iteration sequences which converge more rapidly than EM iteration sequences, Peters and Walker [101], [102] and Redner [108] considered iterative methods which proceed at each iteration in the EM direction with a step whose length is controlled by a parameter ϵ . In the present context, these methods take the form

$$\phi^+ = F_\epsilon(\phi^C) = (1-\epsilon)\phi^C + \epsilon G(\phi^C), \quad (5.6)$$

where $G(\phi)$ is the EM iteration function defined by (4.5) and (5.3). The idea is to optimize the speed of convergence to ϕ^N of an iteration sequence generated by such a method for large N by choosing ϵ to minimize the spectral radius of $E[F'_\epsilon(\phi^*)] = I - \epsilon QR$. As in [101], [102] and [108], one can easily show that the optimal choice of ϵ is always greater than one, lies near one if the component populations in the mixture are "well-separated" in the above sense, and cannot be much smaller than two if the component populations are "poorly separated" in the above sense. The extent to which the speed of convergence of an iteration sequence can be enhanced by making

OF POOR QUALITY

the optimal choice of ϵ in (5.6) is determined by the length of the subinterval of $(0,1]$ in which the spectrum of QR lies. (Greater improvements in convergence speed are realized from the optimal choice of ϵ when this subinterval is relatively narrow.) The applications of iterative procedures of the form (5.6) are at present incompletely explored and might well bear further investigation.

We conclude this section by briefly reviewing some of the special forms which the EM iteration takes when a particular component density $p_1(x|\rho_1)$ is a member of one of the common exponential families. We also comment on some convergence properties of sequences $\{\rho_1^{(j)}\}_{j=0,1,2,\dots}$ generated by the EM algorithm in the examples considered. Hopefully, our comments will prove helpful in determining convergence properties of EM iteration sequences $\{\phi^{(j)}\}_{j=0,1,2,\dots}$ through the use of Theorem 5.1 or other means when all component densities are from one or more of these example families.

Example 1: Poisson density. In this example, $n = 1$ and a natural choice of Ω_1 is $\Omega_1 = \{\rho_1 \in \mathbb{R}^1: 0 < \rho_1 < \infty\}$. For $\rho_1 \in \Omega_1$, one has

$$p_1(x|\rho_1) = \frac{1}{x!} e^{-\rho_1} \rho_1^x, \quad x = 0, 1, 2, \dots;$$

and the EM iteration (5.3) for a sample of Type 1 becomes

$$\rho_i^+ = \left(\sum_{k=1}^N x_k \frac{\alpha_i^c p_i(x_k | \rho_i^c)}{p(x_k | \Phi^c)} \right) / \left(\sum_{k=1}^N \frac{\alpha_i^c p_i(x_k | \rho_i^c)}{p(x_k | \Phi^c)} \right).$$

Note that ρ_i^+ is always contained in the convex hull of $(x_k)_{k=1, \dots, N}$, which is a compact subset of Ω_i . Therefore, the set of limit points of an EM iteration sequence $(\rho_i^{(j)})_{j=0, 1, 2, \dots}$ is a nonempty compact subset of Ω_i .

Example 2: Binomial density. Here, $n = 1$ and one naturally chooses Ω_i to be the open set $\{\rho_i \in \mathbb{R}^1: 0 < \rho_i < 1\}$. For $\rho_i \in \Omega_i$, $p_i(x | \rho_i)$ is given by

$$p_i(x | \rho_i) = \binom{\nu_i}{x} \rho_i^x (1 - \rho_i)^{\nu_i - x}, \quad x = 0, 1, \dots, \nu_i,$$

for a prescribed integer ν_i . In this case, the EM iteration (5.3) for a sample of Type 1 becomes

$$\rho_i^+ = \left(\frac{1}{\nu_i} \sum_{k=1}^N x_k \frac{\alpha_i^c p_i(x_k | \rho_i^c)}{p(x_k | \Phi^c)} \right) / \left(\sum_{k=1}^N \frac{\alpha_i^c p_i(x_k | \rho_i^c)}{p(x_k | \Phi^c)} \right).$$

Since $p_i(x | \rho_i)$ is non-zero only if $x = 0, 1, \dots, \nu_i$ one sees from this expression that the set of limit points of an EM iteration sequence $(\rho_i^{(j)})_{j=0, 1, 2, \dots}$ is a nonempty compact subset of $\bar{\Omega}_i = \{\rho_i \in \mathbb{R}^1: 0 < \rho_i < 1\}$.

Example 3: Exponential density. Again, $n = 1$ and one takes $\Omega_i = \{\rho_i \in \mathbb{R}^1: 0 < \rho_i < \infty\}$. For $\rho_i \in \Omega_i$, one has

$$p_i(x|\rho_i) = \frac{1}{\rho_i} e^{-x/\rho_i}, \quad 0 < x < \infty.$$

The EM iteration (5.3) for a sample of Type 1 now becomes

$$\rho_i^+ = \left(\frac{\sum_{k=1}^N x_k \frac{\alpha_i^c p_i(x_k|\rho_i^c)}{p(x_k|\Phi^c)}}{\sum_{k=1}^N \frac{\alpha_i^c p_i(x_k|\rho_i^c)}{p(x_k|\Phi^c)}} \right),$$

and one sees that the set of limit points of an EM iteration sequence $\{\rho_i^{(j)}\}_{j=0,1,2,\dots}$ is a nonempty compact subset of Ω_i .

Example 4: Multivariate normal density. In this example, n is an arbitrary positive integer; and ρ_i is most conveniently represented as $\rho_i = (\mu_i, \Sigma_i)$, where $\mu_i \in R^n$ and Σ_i is a positive-definite symmetric $n \times n$ matrix. (Of course, this representation of ρ_i is not the usual representation of the "expectation" parameter.) Then Ω_i is the set of all such ρ_i , and $p_i(x|\rho_i)$ is given by (4.7). For a sample of Type 1, the EM iteration (5.3) becomes that of (4.8) and (4.9).

One can see from (4.9) that each Σ_i^+ is in the convex hull of $\{(x_k - \mu_i^+)(x_k - \mu_i^+)^T\}_{k=1,\dots,N}$, a set of rank-one matrices which, of course, are not positive definite. Thus there is no guarantee that a sequence of matrices $\{\Sigma_i^{(j)}\}_{j=0,1,2,\dots}$ produced by the EM iteration will remain bounded from below. Indeed, it has been observed in practice that sequences of iterates produced by the EM algorithm for a mixture of multivariate normal densities do occasionally converge to "singular solutions" (cf.

Duda and Hart [44]), i.e., points on the boundary of Ω_1 with associated singular matrices.

It was observed by Hosmer [68] that if enough labeled observations are included in a sample on a mixture of normal densities, then with probability 1, the log-likelihood function attains its maximum value at a point at which the covariance matrices are positive definite. Similarly, consideration of samples with a sufficiently large number of labeled observations alleviates with probability 1 the problem of an EM iteration sequence having "singular solutions" as limit points. For example, if one considers a sample $S = S_1 \cup S_3$ which is a stochastically independent union of a sample $S_1 = \{x_k\}_{k=1, \dots, N}$ of Type 1 and a sample $S_3 = \bigcup_{i=1}^m \{z_{ik}\}_{k=1, \dots, K_i}$ of Type 3, then the EM iteration becomes

$$\alpha_i^+ = \frac{1}{N+K} \left(\sum_{k=1}^N \frac{\alpha_i^c p_i(x_k | \rho_i^c)}{p(x_k | \Phi^c)} + K_i \right)$$

$$\mu_i^+ = \left\{ \sum_{k=1}^N x_k \frac{\alpha_i^c p_i(x_k | \rho_i^c)}{p(x_k | \Phi^c)} + \sum_{k=1}^{K_i} z_{ik} \right\} / \left(\sum_{k=1}^N \frac{\alpha_i^c p_i(x_k | \rho_i^c)}{p(x_k | \Phi^c)} + K_i \right)$$

$$\Sigma_i^+ = \left(\sum_{k=1}^N (x_k - \mu_i^+) (x_k - \mu_i^+)^T \frac{\alpha_i^c p_i(x_k | \rho_i^c)}{p(x_k | \Phi^c)} \right)$$

$$+ \sum_{k=1}^{K_i} (z_{ik} - \mu_i^+) (z_{ik} - \mu_i^+)^T \left/ \left(\sum_{k=1}^N \frac{\alpha_i^c p_i(x_k | \rho_i^c)}{p(x_k | \Phi^c)} + K_i \right) \right.,$$

OF POOR QUALITY

where $K = \sum_{i=1}^m K_i$. One sees from the expression for Σ_i^+ that Σ_i^+ is bounded below by

$$\left(\sum_{k=1}^{K_i} (z_{ik} - \mu_i^+) (z_{ik} - \mu_i^+)^T \right) / \left(\sum_{k=1}^N \frac{\alpha_i^c p_i(x_k | \phi_i^c)}{p(x_k | \Phi^c)} + K_i \right),$$

which is in turn bounded below by

$$\frac{1}{N+K_i} \left(\sum_{k=1}^{K_i} (z_{ik} - \bar{z}) (z_{ik} - \bar{z})^T \right),$$

where

$$\bar{z} = \frac{1}{K_i} \sum_{k=1}^{K_i} z_{ik}.$$

Now this last matrix is positive-definite with probability 1 whenever $K_i > n$. Consequently, if $K_i > n$, then with probability 1, the elements of a sequence $\{\Sigma_i^{(j)}\}_{j=0,1,2,\dots}$ produced by the EM algorithm are bounded below by a positive definite matrix; hence, such a sequence cannot have singular matrices as limit points.

6. Performance of the EM Algorithm

In this concluding section, we review and summarize features of the EM algorithm having to do with its effectiveness in practice on mixture density estimation problems. As always, it is understood that a parametric family of mixture densities of the form (1.1) is of interest and that a particular $\Phi^* = (\alpha_1^*, \dots, \alpha_m^*, \rho_1^*, \dots, \rho_m^*)$ is the "true" parameter value to be estimated.

In order to provide some perspective, we begin by offering a brief description of the most basic forms of several alternative methods for numerically approximating maximum-likelihood estimates. In describing these methods, it is assumed for convenience that the sample at hand is a sample $S_1 = \{x_k\}_{k=1, \dots, N}$ of Type 1 described in Section 3 and that one can write Φ as a vector $\Phi = (\xi_1, \dots, \xi_\nu)^T$ of unconstrained scalar parameters at points of interest in Ω . Each of the methods to be described seeks a maximum-likelihood estimate by attempting to determine a point $\hat{\Phi}$ such that

$$\nabla_{\Phi} L_1(\hat{\Phi}) = 0, \quad (6.1)$$

where $L_1(\Phi)$ is the log-likelihood function given by (3.1). The features of the methods which concern us here are their speed of convergence, the computation and storage required for their implementation, and the extent to which their basic forms need to be modified in order to make them effective and trustworthy in

practice.

OF POOR QUALITY.

The first of the alternative methods to be described is Newton's method. It is the method on which the other methods reviewed here are modeled, and it is given as follows: Given a current approximation ϕ^c of a solution of (6.1), determine a next approximation ϕ^+ by

$$\phi^+ = \phi^c - H(\phi^c)^{-1} \nabla_{\phi} L_1(\phi^c) . \quad (6.2)$$

The function $H(\phi)$ in (6.2) is the Hessian matrix of $L_1(\phi)$ given by (3.10).

Under reasonable assumptions on $L_1(\phi)$, one can show that a sequence of iterates $\{\phi^{(j)}\}_{j=0,1,2,\dots}$ produced by Newton's method enjoys quadratic local convergence to a solution $\hat{\phi}$ of (6.1) (see, for example, Ortega and Rheinboldt [95]). This is to say that given a norm $\|\cdot\|$ on Ω , there is a constant β such that if $\phi^{(0)}$ is sufficiently near $\hat{\phi}$, then an inequality

$$\|\phi^{(j+1)} - \hat{\phi}\| \leq \beta \|\phi^{(j)} - \hat{\phi}\|^2 \quad (6.3)$$

holds for $j = 0, 1, 2, \dots$. Quadratic convergence is ultimately very fast, and it is regarded as the major strength of Newton's method. Unfortunately, there are aspects of Newton's method which are associated with potentially severe problems in some applications. For one thing, Newton's method requires at each iteration the computation of the $\nu \times \nu$ Hessian matrix and the solution of a system of ν linear equations (at a cost of $O(\nu^3)$)

arithmetic operations in general) with this Hessian as the coefficient matrix; thus the computation required for an iteration of Newton's method is likely to become expensive very rapidly as m, n , and N grow large. (It should also be mentioned that one must allow for the storage of the Hessian or some set of factors of it.) For another thing, Newton's method in its basic form (6.2) requires for some problems an impractically accurate initial approximate solution $\phi^{(0)}$ in order for a sequence of iterates $\{\phi^{(j)}\}_{j=0,1,2,\dots}$ to converge to a solution of (5.1). Consequently, in order to be regarded as an algorithm which is safe and effective on applications of interest, the basic form (6.2) is likely to require augmentation with some procedure for enhancing the global convergence behavior of sequences of iterates produced by it. Such a procedure should be designed to insure that a sequence of iterates not only converges but also does not converge to a solution of (6.1) which is not a local maximum of $L_1(\Phi)$.

A broad class of methods which are based on Newton's method are quasi-Newton methods of the general form

$$\phi^+ = \phi^c - B^{-1} \nabla_{\phi} L_1(\phi^c), \quad (6.4)$$

in which B is regarded as an approximation of $H(\phi^c)$. Methods of the form (6.4) which are particularly successful are those in which the approximation $B \approx H(\phi^c)$ is maintained by doing a secant update of B at each iteration (see Dennis and Moré [40] or Dennis and Schnabel [41]). In the applications of interest

here, such updates are typically realized as rank one or (more likely) rank two changes in B . Methods employing such updates have the advantages over Newton's method of not requiring the evaluation of the Hessian matrix at each iteration and of being implementable in ways which require only $O(v^2)$ arithmetic operations to solve the system of v linear equations at each iteration. The price paid for these advantages is that the full quadratic convergence of Newton's method is lost; rather, under reasonable assumptions on $L_1(\Phi)$, a sequence of iterates $(\phi^{(j)})_{j=0,1,2,\dots}$ produced by one of these methods can only be shown to exhibit local superlinear convergence to a solution $\hat{\phi}$ of (6.1), i.e., one can only show that if a norm $\|\cdot\|$ on Ω is given and if $\phi^{(0)}$ is sufficiently near $\hat{\phi}$ (and an initial approximate Hessian $B^{(0)}$ is sufficiently near $H(\hat{\phi})$), then there exists a sequence $(\beta_j)_{j=0,1,2,\dots}$ which converges to zero and is such that

$$\|\phi^{(j+1)} - \hat{\phi}\| < \beta_j \|\phi^{(j)} - \hat{\phi}\|$$

for $j = 0, 1, 2, \dots$. Like Newton's method, methods of the general form (6.4), including those employing secant updates, are likely to require augmentation with safeguards to enhance global convergence properties and to insure that iterates do not converge to solutions of (6.1) which are not local maxima of $L_1(\Phi)$.

Finally, we describe a particular method of the form (6.4) which is specifically formulated for solving likelihood

equations. This is the method of scoring, mentioned earlier in connection with the work of Rao [107] and reviewed in a general setting by Kale [74], [75]. (Kale [74], [75] also discusses modifications of Newton's method and the method of scoring in which the Hessian matrix or an approximation of it is held fixed for some number of iterations in the hope of reducing overall computational effort.) In the method of scoring, one ideally chooses B in (6.4) to be

$$B = -NI(\Phi^C) , \quad (6.5)$$

where $I(\Phi)$ is the Fisher information matrix given by (2.5.1). Since the computation of $I(\Phi^C)$ is likely to be prohibitively expensive for most mixture density problems, a more appealing choice of B than (6.5) might be the sample approximation

$$B = - \sum_{k=1}^N [\nabla_{\Phi} \log p(x_k | \Phi^C)] [\nabla_{\Phi} \log p(x_k | \Phi^C)]^T . \quad (6.6)$$

The choice (6.6) can be justified in the following manner: The Hessian $H(\Phi)$ is given by

$$H(\Phi) = - \sum_{k=1}^N [\nabla_{\Phi} \log p(x_k | \Phi)] [\nabla_{\Phi} \log p(x_k | \Phi)]^T + \sum_{k=1}^N \frac{1}{p(x_k | \Phi)} \nabla_{\Phi} \nabla_{\Phi}^T p(x_k | \Phi) . \quad (6.7)$$

Now the second sum in (6.7) has zero expectation at $\Phi = \Phi^*$; furthermore, since the terms $\nabla_{\Phi} \log p(x_k | \Phi)$ must be computed in

OF POOR QUALITY

order to obtain $\nabla_{\Phi} L_1(\Phi)$, the first sum in (6.7) is available at the cost of only $O(Nv^2)$ arithmetic operations while determining the second sum is likely to involve a great deal more expense. Thus (6.6) is a choice of B which is readily available at relatively low cost and which is likely to constitute a major part of $H(\Phi^C)$ when N is large and Φ^C is near Φ^* . It is clear from this discussion that the method of scoring with B given by (6.6) is an analogue for general maximum-likelihood estimation of the Gauss-Newton method for nonlinear least-squares problems (see Ortega and Rheinboldt [95]). If the computation of $I(\Phi^C)$ is not too expensive, then the choice of B given by (6.5) can be justified in much the same way.

The method of scoring in its basic form requires $O(Nv^2)$ arithmetic operations to evaluate B given by (6.6) and $O(v^3)$ arithmetic operations to solve the system of v linear equations implicit in (6.4). Since these $O(Nv^2)$ arithmetic operations are likely to be considerably less expensive than the evaluation of the full Hessian given by (6.7), the cost of computation per iteration of the method of scoring lies between that of a quasi-Newton method employing a low-rank secant update and that of Newton's method. Under reasonable assumptions on $L_1(\Phi)$, one can show that with probability 1, if a solution $\hat{\Phi}$ of (6.1) is sufficiently near Φ^* and if N is sufficiently large, then a sequence of iterates $\{\Phi^{(j)}\}_{j=0,1,2,\dots}$ generated by the method of scoring with B given by either (6.5) or (6.6) exhibits local linear convergence to $\hat{\Phi}$, i.e., there is a norm $\|\cdot\|$ on Ω and

a constant λ , $0 < \lambda < 1$, for which

$$\|\phi^{(j+1)} - \hat{\phi}\| < \lambda \|\phi^{(j)} - \hat{\phi}\|, \quad j = 0, 1, 2, \dots \quad (6.8)$$

whenever $\phi^{(0)}$ is sufficiently near $\hat{\phi}$. If $\hat{\phi}$ is very near ϕ^* and if N is very large, then this convergence should be fast, i.e., (6.8) should hold for a small constant λ . Like Newton's method and all methods of the general form (6.4), the method of scoring is likely to require augmentation with global convergence safeguards in order to be considered trustworthy and effective.

Having reviewed the above alternative methods, we return now to the EM algorithm and summarize its attractive features. Its most appealing general property is that it produces sequences of iterates on which the log-likelihood function increases monotonically. This monotonicity is the basis of the general convergence theorems of Section 4, and these theorems reinforce a large body of empirical evidence to the effect that the EM algorithm does not require augmentation with elaborate safeguards such as those necessary for Newton's method and quasi-Newton methods in order to produce iteration sequences with good global convergence characteristics.

More can be said about the EM algorithm for mixtures of densities from exponential families under the assumption that ρ_1, \dots, ρ_m are mutually independent variables. One sees from (4.5) and (5.3) and similar expressions for samples of types other than Type 1 that it is unlikely that any other algorithm

would be nearly as easy to encode on a computer or would require as little storage. In view of (4.5) and (5.3), it also seems that any constraints on Φ are likely to be satisfied, or at least nearly satisfied for large samples. For example, it is clear from (4.9) that each Σ_i^+ generated by the EM algorithm for a mixture of multivariate normal densities is symmetric and, with probability 1, positive-definite whenever $N > n$. Certainly the mixing proportions generated by (4.5) are always non-negative and sum to 1. It is also apparent from (4.5) and (5.3) that the computational cost of each iteration of the EM algorithm is low compared to that of the alternative methods reviewed above. In the case of a mixture of multivariate normal densities, for example, the EM algorithm requires $O(mn^2N)$ arithmetic operations per iteration, compared to at least $[O_1(m^2n^4N) + O_2(m^3n^6)]$ for Newton's method and the method of scoring and $[O_1(mn^2N) + O_2(m^2n^4)]$ for a quasi-Newton method employing a low-rank secant update (All of these methods require the same number of exponential function evaluations per iteration.) Arithmetic per iteration for the three latter methods can, of course, be reduced by retaining a fixed approximate Hessian for some number of iterations at the risk of increasing the total number of iterations.

In spite of these attractive features, the EM algorithm can encounter problems in practice. The source of the most serious practical problems associated with the algorithm is the speed of convergence of sequences of iterates generated by it, which can

often be annoyingly or even hopelessly slow. In the case of mixtures of densities from exponential families, Theorem 5.2 suggests that one can expect the convergence of EM iteration sequences to be linear, as opposed to the (very fast) quadratic convergence associated with Newton's method, the (fast) superlinear convergence associated with a quasi-Newton method employing a low-rank secant update, and the (perhaps fast) linear convergence of the method of scoring. The discussion following Theorem 5.2 suggests further that the speed of this linear convergence depends in a certain sense on the separation of the component populations in the mixture. To demonstrate the speed of this linear convergence and its dependence on the separation of the component populations, we again consider the example of a mixture of two univariate normal densities (see (1.3) and (1.4)).

Table 6.1 below summarizes the results of a numerical experiment involving a mixture of two univariate normal densities for the choices of Φ^* appearing in Table 3.3. (These choices were obtained as before by taking $\alpha_1^* = .3$, $\sigma_1^{2*} = \sigma_2^{2*} = 1$, and varying the mean separation $\mu_1^* - \mu_2^*$. For convenience, we took $\mu_2^* = -\mu_1^*$.) In this experiment, a Type 1 sample of 1000 observations on the mixture was generated for each choice of Φ^* ; and a sequence of iterates was produced by the EM algorithm (see (4.5), (4.8), and (4.9)) from starting values $\alpha_1^{(0)} = \alpha_2^{(0)} = .5$, $\mu_1^{(0)} = 1.5\mu_1^*$, $\mu_2^{(0)} = 1.5\mu_2^*$, and $\sigma_1^{2(0)} = \sigma_2^{2(0)} = .5$. An accurate determination of the limit of the sequence was made in each case, and observations were made of

the iteration numbers at which various degrees of accuracy were first obtained. These iteration numbers are recorded in Table 6.1 beneath the corresponding degrees of accuracy; in the table, "E" denotes the largest absolute value of the components of the difference between the indicated iterate and the limit. In addition, the spectral radius of the derivative of the EM iteration function at the limit was calculated in each case (cf. Theorem 5.2 and the following discussion). These spectral radii, appearing in the column headed by " ρ " in Table 6.1, provide quantitative estimates of the factors by which errors are reduced from one iteration to the next in each case. Finally, to give an idea of the point in an iteration sequence at which numerical error first begins to affect the theoretical performance of the algorithm, we observed in each case the iteration numbers at which loss of monotonicity of the log-likelihood function first occurred; these iteration numbers appear in Table 6.1 in the column headed by "LM".

In preparing Table 6.1, all computing was done in double precision on an IBM 3032.³ Eigenvalues were calculated with EISPACK subroutines TRED1 and TQL1, and normally distributed data was obtained by transforming uniformly distributed data generated by the subroutine URAND of Forsythe, Malcolm, and Moler [46] based on suggestions of Knuth [79].

3. We are grateful to the Mathematics and Statistics Department of the University of New Mexico for providing the computing support for the generation of this table.

ORIGINAL PAGE IS
OF POOR QUALITY

$\mu_1^* - \mu_2^*$	$E < 10^{-1}$	$E < 10^{-2}$	$E < 10^{-3}$	$E < 10^{-4}$	$E < 10^{-5}$	$E < 10^{-6}$	$E < 10^{-7}$	$E < 10^{-8}$	LM	ρ
0.2	2078	2334	2528	2717	2906	3095	3283	3472	3056	.9879
0.5	710	852	985	1117	1249	1381	1513	1643	1361	.9827
1.0	349	442	526	610	693	777	861	949	779	.9728
1.5	280	414	537	660	783	906	1028	1151	887	.9814
2.0	126	281	432	582	732	883	1033	1183	846	.9849
3.0	2	31	62	93	124	155	185	216	173	.9280
4.0	1	6	16	25	35	44	54	63	55	.7864
6.0	1	1	2	3	4	5	7	8	8	.2143

Table 6.1: Results of applying the EM algorithm to a problem involving a Type 1 sample on a mixture of two univariate normal densities with $\alpha_1^* = .3$, $\sigma_1^{2*} = \sigma_2^{2*} = 1$.

A number of comments about the contents of Table 6.1 are in order. First, it is clear from the table that an exorbitantly large number of EM iterations may be required to obtain a very accurate numerical approximation of the maximum-likelihood estimate if the sample is from a mixture of poorly separated component populations. However, in such a case, one sees from Table 3.3 that the variance of the estimate is likely to be such that it may be pointless to seek very much accuracy in a

numerical approximation. Second, we remark on the pleasing consistency between the computed values of the spectral radius of the derivative of the EM iteration function and the differences between the iteration numbers needed to obtain varying degrees of accuracy. What we have in mind is the following: If the errors among the members of a linearly convergent sequence are reduced more or less by a factor of ρ , $0 < \rho < 1$, from one iteration to the next, then the number of iterations Δk necessary to obtain an additional decimal digit of accuracy is given approximately by $\Delta k \approx \log 10 / \log \rho$. This relationship between Δk and ρ is borne out very well in Table 6.1. This fact strongly suggests that after a number of EM iterations have been made, the errors in the iterates lie almost entirely in the eigenspace corresponding to the dominant eigenvalue of the derivative of the EM iteration function. We take this as evidence that one might very profitably apply simple relaxation-type acceleration procedures such as those of Peters and Walker [101], [102] and Redner [108] to sequences of iterates generated by the EM algorithm.

Third, in all of the cases listed in Table 6.1 except one, we observed that over 95 percent of the change in the log-likelihood function between the starting point and the limit of the EM iteration sequence was realized after only five iterations, regardless of the number of iterations ultimately required to approximate the limit very closely. (The exceptional case is that in which $\mu_1^* - \mu_1^* = 1.0$; in that case, about 83

percent of the change in the log-likelihood function was observed after five iterations.) This suggests to us that even when the component populations in a mixture are poorly separated, the EM algorithm can be expected to produce in a very small number of iterations parameter values such that the mixture density determined by them reflects the sample data very well. Fourth, it is evident from Table 6.1 that elements of an EM iteration sequence continue to make steady progress toward the limit even after numerical error has begun to interfere with the theoretical properties of the algorithm.

Fifth, the apparently anomalous decrease in ρ occurring when $\mu_1^* - \mu_2^*$ decreases from 2.0 to 1.0 happened concurrently with the iteration sequence limit of the proportion of the first population in the mixture becoming very small. (Such very small limit proportions continued to be observed in the cases $\mu_1^* - \mu_2^* = 0.5, 0.2$.) We do not know whether this decrease in the limit proportion of the first population indicates a sudden movement of the maximum-likelihood estimate as $\mu_1^* - \mu_2^*$ drops below 2.0 or whether the iteration sequence limit is something other than the maximum-likelihood estimate in the cases in which $\mu_1^* - \mu_2^*$ is less than 2.0. Finally, we also conducted more than 60 trials similar to those reported on in Table 6.1 except with samples of 200 rather than 1000 generated observations on the mixture. The results were comparable to those given in Table 6.1. It should be mentioned, however, that the EM iteration sequences obtained using samples of 200 observations did

occasionally converge to "singular solutions," i.e., limits associated with zero component variances. Convergence to such "singular solutions" did not occur among the relatively small number of trials involving samples of 1000 observations.

At present, the EM algorithm is being widely applied not only to mixture density estimation problems but also to a wide variety of other problems as well. We would like to conclude this survey with a little speculation about the future of the algorithm. It seems likely that the EM algorithm in its basic form will find a secure niche as an algorithm useful in situations in which some resources are limited. For example, the limited time which an experimenter can afford to spend writing programs coupled with a lack of available library software for safely and efficiently implementing competing methods could make the simplicity and reliability of the EM algorithm very appealing. The EM algorithm might be very well suited for use on small computers for which limitations on program and data storage are more stringent than limitations on computing time.

Although meaningful comparison tests have not yet been made, it seems doubtful to us that the unadorned EM algorithm can be competitive as a general tool with well-designed general optimization algorithms such as those implemented in good currently-available software library routines. Our doubt is based on the intolerably slow convergence of sequences of iterates generated by the EM algorithm in some applications. On the other hand, it is entirely possible that the EM algorithm

could be modified to incorporate procedures for accelerating convergence and that such modification would enhance its competitiveness. It is also possible that an effective hybrid algorithm might be constructed which first takes advantage of the good global convergence properties of the EM algorithm by using it initially and then exploits the rapid local convergence of Newton's method or one of its variants by switching to such a method later. Our feeling is that time might well be spent on research addressing these possibilities.

REFERENCES

1. M.A. Acheson and E.M. McElwee, "Concerning the reliability of electron tubes," *The Sylvania Technologist* 4 (1951).
2. J. Aitchison and S.D. Silvey, "Maximum-likelihood estimation of parameters subject to restraints," *Ann. Math. Statist.* 3 (1958), 813-828.
3. N. Arley and K.R. Buch, Introduction to the Theory of Probability and Statistics, John Wiley and Sons, Inc., New York (1950).
4. G.A. Baker, "Maximum-likelihood estimation of the ratio of the two components of non-homogeneous populations," *Tohoku Math. Jour.* 47 (1940), 304-308.
5. O. Barndorff-Nielsen, "Identifiability of mixtures of exponential families," *J. Math. Anal. Appl.* 12 (1965), 115-121.
6. O. Barndorff-Nielsen, Information and Exponential Families in Statistical Theory, John Wiley and Sons, Ltd., New York (1978).
7. L.E. Baum, T. Petrie, G. Soules, and N. Weiss, "A maximization technique occurring in the statistical analysis of probabilistic functions of Markov chains," *Ann. Math. Statist.* 41 (1970), 164-171.
8. J. Behboodian, "On a mixture of normal distributions," *Biometrika* 57, Part 1 (1970), 215-217.
9. J. Behboodian, "Information matrix for a mixture of two normal distributions," *J. Statist. Comput. Simul.* 1 (1972), 295-314.
10. C.T. Bhattacharya, "A simple method of resolution of a distribution into Gaussian components," *Biometrics* 23 (1967), 115-137.
11. W. R. Blischke, "Moment estimators for the parameters of a mixture of two binomial distributions," *Ann. Math. Statist.* 33 (1962), 444-454.
12. W.R. Blischke, "Mixtures of discrete distributions," *Proceedings of the International Symposium on Classical and Contagious Discrete Distributions*, Pergamon Press (1963), 351-372.
13. W.R. Blischke, "Estimating the parameters of mixtures of binomial distributions," *Amer. Statist. Assoc.* 59 (1964),

510-528.

14. D. Boes, "On the estimation of mixing distributions," *Ann. Math. Statist.* 37 (1966), 177-188.
15. D.C. Boes, "Minimax unbiased estimator of mixing distribution for finite mixtures," *Sankhyā A*, 29 (1967), 417-420.
16. K.O. Bowman and L.R. Shenton, "Space of solutions for a normal mixture," *Biometrika* 60 (1973), 629-636.
17. R.A. Boyles, "Convergence results for the EM algorithm," *Tech. Rep. No. 13*, University of California, Davis (June, 1980).
18. C. Burrau, "The half-invariants of the sum of two typical laws of errors, with an application to the problem of dissecting a frequency curve into components," *Skand. Aktuarietidskrift* 17 (1934), 1-6.
19. R. M. Cassie, "Some uses of probability paper in the analysis of size frequency distributions," *Austral. Jour. Marine and Freshwater Research* 5 (1954), 513-523.
20. R. Ceppellini, S. Siniscalco, and C.A.B. Smith, "The estimation of gene frequencies in a random-mating population," *Ann. Hum. Genetics* 20 (1955), 97-115.
21. K.C. Chanda, "A note on the consistency and maxima of the roots of the likelihood equations," *Biometrika* 41 (1954), 56-61.
22. W.C. Chang, "The effects of adding a variable in dissecting a mixture of two normal populations with a common covariance matrix," *Biometrika* 63 (1976), 676-678.
23. W.C. Chang, "Confidence interval estimation and transformation of data in a mixture of two multivariate normal distributions with any given large dimension," *Technometrics* 21 (1979), 351-355.
24. C.V.L. Charlier, "Researches into the theory of probability," *Acta. Univ. Lund. (Noue Folge. Abt. 2)* 1 (1906), 33-38.
25. C.V.L. Charlier and S.D. Wicksell, "On the dissection of frequency functions," *Arkiv. for Matematik Astronomi Och Fysik*, Bd. 18, No. 6 (1924), Stockholm.
26. T. Chen, "Mixed-up frequencies in contingency tables," Ph.D. dissertation, Univ. of Chicago (1972).

27. K. Choi, "Estimators for the parameters of a finite mixture of distributions," *Ann. Inst. Statist. Math.* 21 (1969), 107-116.
28. K. Choi and W.B. Bulgren, "An estimation procedure for mixtures of distributions," *J. Royal Statist. Soc., Ser. B*, 30 (1968), 444-460.
29. A.C. Cohen, Jr., "Estimation in mixtures of discrete distributions," *Proceedings of the International Symposium on Classical and Contagious Discrete Distributions*, Pergamon Press (1963), 351-372.
30. A.C. Cohen, "A note on certain discrete mixed distributions," *Biometrics* 22 (1966), 566-571.
31. A.C. Cohen, "Estimation in mixtures of two normal distributions," *Technometrics* 9 (1967), 15-28.
32. *Communications in Statistics, Special Issue on Remote Sensing*, *Comm. Statist. - Theor. Meth.* A5 (12)(1976).
33. P. W. Cooper, "Some topics on nonsupervised adaptive detection for multivariate normal distributions," in Computer and Information Sciences II, J.T. Tou, Ed., Academic Press (1967), 143-146.
34. D.R. Cox, "The analysis of exponentially distributed lifetimes with two types of failure," *J. Royal Statist. Soc.* 21 B (1959), 411-421.
35. H. Cramér, Mathematical Methods of Statistics, Princeton University Press, Princeton, N.J. (1946).
36. N.E. Day, "Estimating the components of a mixture of normal distributions," *Biometrika* 56 (1969), 463-474.
37. J.J. Deely and R.L. Kruse, "Construction of sequences estimating the mixing distribution," *Ann. Math. Statist.* 39 (1968), 286-288.
38. A. P. Dempster, N.M. Laird, and D.B. Rubin, "Maximum-likelihood from incomplete data via the EM algorithm," *J. Royal Statist. Soc., Ser. B (methodological)* 39 (1977), 1-38.
39. J.E. Dennis, Jr., "Algorithms for nonlinear fitting," *Proceedings of the NATO Advanced Research Symposium*, Cambridge University, Cambridge, England (July, 1981).
40. J.E. Dennis, Jr., and J.J. Moré, "Quasi-Newton methods, motivation and theory," *SIAM Rev.* 19 (1977), 46-89.

41. J.E. Dennis, Jr. and R.B. Schnabel, "Least-change secant updates for quasi-Newton methods," *SIAM Rev.* 21 (1979), 443-459.
42. N.P. Dick and D.C. Bowden, "Maximum-likelihood estimation for mixtures of two normal distributions," *Biometrics* 29 (1973), 781-791.
43. G. Doetsch, "Zerlegung einer Funktion in Gausche Fehlerkurven und zeitliche Zuruckverfolgung eines Temperaturzustandes," *Mathematische Zeitschrift* 41 (1936), 283-318.
44. R.O. Duda and P.E. Hart, Pattern Classification and Scene Analysis, John Wiley and Sons, Inc., New York (1973).
45. B.S. Everitt and D.J. Hand, Finite Mixtre Distributions, Chapman and Hall Ltd., London (1981).
46. G.E. Forsythe, M.A. Malcolm, C.B. Moler, Computer Methods for Mathemaical Computations, Prentice Hall, Inc., Englewood Cliffs, N.J. (1977).
47. J.G. Fryer and C.A. Robertson, "A comparison of some methods for estimating mixed normal distributions," *Biometrika* 59 (1972), 639-648.
48. K. Fukunaga, Introduction to Statistical Pattern Recognition, Academic Press, New York (1972).
49. S. Ganesalingam and G.J. McLachlan, "Some efficiency results for the estimation of the mixing proportion in a mixture of two normal distributions," *Biometrics* 37 (1981), 23-33.
50. L.A. Goodman, "The analysis of systems of qualitative variables when some of the variables are unobservable: Part I—a modified latent structure approach," *Amer. J. Sociol.* 79 (1974), 1179-1259.
51. V. H. Gottschalk, "Symmetric bimodal frequency curves," *J. Franklin Inst.* 245 (1948), 245-252.
52. J. Gregor, "An algorithm for the decomposition of a distribution into Gaussian components," *Biometrics* 25 (1969), 79-93.
53. N.T. Gridgeman, "A comparison of two methods of analysis of mixtures of normal distributions," *Technometrics* 12 (1970), 823-833.
54. E.J. Gumbel, "La dissection d'une repartition," *Annales de l'Universite de Lyon*, (3), A (1939), 39-51.

55. L.F. Guseman, Jr., and Jay R. Walton, "An application of linear feature selection to estimation of proportions," *Commun. Statist. -Theor. Meth. A6* (1977), 611-617.
56. L.F. Guseman, Jr., and Jay R. Walton, "Methods for estimating proportions of convex combinations of normals using linear feature selection," *Commun. Statist. - Theor. Meth. A7* (1978), 1439-1450.
57. S. J. Haberman, "Log-linear models for frequency tables derived by indirect observations: Maximum-likelihood equations," *Ann. Statist. 2* (1974), 911-924.
58. S. J. Haberman, "Iterative scaling procedures for log-linear models for frequency tables derived by indirect observation," *Proc. Amer. Statist. Assoc. (Statist. Comp. Sect. 1975)*, 1976, 45-50.
59. S.J. Haberman, "Product models for frequency tables involving indirect observation," *Ann. Stat. 5* (1977), 1124-1147.
60. A. Hald, "The compound hypergeometric distribution and a system of single sampling inspection plans based on prior distributions and costs," *Technometrics 2* (1960), 275-340.
61. J.P. Harding, "The use of probability paper for the graphical analysis of polynomial frequency distributions," *J. of the Marine Biological Assoc. 28* (1949), 141-153.
62. J.A. Hartigan, Clustering Algorithms, John Wiley and Sons, New York (1975).
63. M.J. Hartley, comment on [105], *J. Amer. Statist. Assoc. 73* (1978), 738-741.
64. V. Hasselblad, "Estimation of parameters for a mixture of normal distributions," *Technometrics 8* (1966), 431-444.
65. V. Hasselblad, "Estimation of finite mixtures of distributions from the exponential family," *J. Amer. Statist. Assoc. 64* (1969), 1459-1471.
66. B.M. Hill, "Information for estimating the proportions in mixtures of exponential and normal distributions," *J. Amer. Statist. Assoc. 58* (1963), 918-932.
67. D.W. Hosmer, Jr., "On MLE of the parameters of a mixture of two normal distributions when the sample size is small," *Commun. Statist. 1* (1973), 217-227.
68. D.W. Hosmer, Jr., "A comparison of iterative maximum-likelihood estimates of the parameters of a mixture of two

- normal distributions under three different types of sample," *Biometrics* 29 (1973), 761-770.
69. D. W. Hosmer, Jr., "Maximum-likelihood estimates of the parameters of a mixture of two regression lines," *Commun. Statist.* 3 (1974), 995-1006.
 70. D. W. Hosmer, Jr., comment on [105], *J. Amer. Statist. Assoc.* 73 (1978), 741-744.
 71. D.W. Hosmer, Jr., and N.P. Dick, "Information and mixtures of two normal distributions," *J. Statist. Comput. Simul.* 6 (1977), 137-148.
 72. V.S. Huzurbazar, "The likelihood equation, consistency and the maxima of the likelihood function," *Annals of Eugenics, Lond.* 14 (1948), 185-200.
 73. I.R. James, "Estimation of the mixing proportion in a mixture of two normal distributions from simple, rapid measurements," *Biometrics* 34 (1978), 265-275.
 74. B. K. Kale, "On the solution of the likelihood equation by iteration processes," *Biometrika* 48 (1961), 452-456.
 75. B.K. Kale, "On the solution of likelihood equations by iteration processes. The multiparametric case," *Biometrika* 49 (1962), 479-486.
 76. D. Kazakos, "Recursive estimation of prior probabilities using a mixture," *IEEE Trans. Inform. Theory* IT-23 (1977), 203-211.
 77. N.M. Kiefer, "Discrete parameter variation: efficient estimation of a switching regression model," *Econometrica* 46 (1978), 427-434.
 78. N.M. Kiefer, comment on [105], *J. Amer. Statist. Assoc.* 73 (1978), 744-745.
 79. D.E. Knuth, "Seminumerical algorithms," The Art of Computer Programming, Vol. 2, Addison-Wesley, Reading, Mass. (1969).
 80. N. Laird, "Nonparametric maximum-likelihood estimation of a mixing distribution," *J. Amer. Statist. Assoc.* 73 (1978), 805-811.
 81. P. F. Lazarsfeld and N.W. Henry, "Latent Structure Analysis," Houghton Mifflin Company, Boston (1968).
 82. M. Loève, Probability Theory, Van Nostrand, New York (1963).
 83. D. G. Luenberger, Optimization by Vector Space Methods, John

- Wiley and Sons, Inc., New York (1969).
84. P.D. M. Macdonald, "Comment on 'an estimation procedure for mixtures of distribution' by Choi and Bulgren," J. Royal Statist. Soc., Ser. B, 33 (1971), 326-329.
 85. P.D. M. Macdonald, "Estimation of finite distribution mixtures," in Applied Statistics, R.P. Gupta, Ed., North Holland Publishing Co. (1975).
 86. C. L. Mallows, review of [88], Biometrics 18 (1962), 617.
 87. J. S. Maritz, Empirical Bayes Methods, Methuen and Co., London (1970).
 88. P. Medgyessy, Decomposition of Superpositions of Distribution Functions, Budapest: Publishing House of the Hungarian Academy of Sciences (1961).
 89. W. Mendenhall and R. J. Hader, "Estimation of parameters of mixed exponentially distributed failure time distributions from censored life test data," Biometrika 45 (1958), 504-520.
 90. H. Muench, "Probability distribution of protection test results," J. Amer. Statist. Assoc. 31 (1936), 677-689.
 91. H. Muench, "Discrete frequency distributions arising from mixtures of several single probability values," J. Amer. Statist. Assoc. 33 (1938), 390-398.
 92. P.L. Odell and J.P. Basu, "Concerning several methods for estimating crop acreages using remote sensing data," Commun. Statist.-Theor. Meth. A5 (12), 1091-1114 (1976).
 93. P.L. Odell and R. Chhikara, "Estimation of a large area crop acreage inventory using remote sensing technology," Annual Report: Statistical Theory and Methodology for Remote Sensing Data Analysis, The University of Texas at Dallas, NASA/JSC-09703 (1975).
 94. T. Orchard and M.A. Woodbury, "A missing information principle: theory and applications," Proceedings of the Sixth Berkeley Symposium on Mathematical Statistics and Probability 1 (1972), 697-715.
 95. J. M. Ortega and W.C. Rheinboldt, Iterative Solution of Nonlinear Equations in Several Variables, Academic Press, New York (1970).
 96. A.M. Ostrowski, Solutions of Equations and Systems of Equations, Academic Press, New York (1966).

97. K. Pearson, "Contributions to the mathematical theory of evolution," *Phil. Trans. Royal Soc.* 185A (1894), 71-110.
98. K. Pearson, "On certain types of compound frequency distributions in which the components can be individually described by binomial series," *Biometrika* 11 (1915-17), 139-144.
99. K. Pearson and A. Lee, "On the generalized probable error in multiple normal correlation," *Biometrika* 6 (1908-09), 59-68.
100. B.C. Peters, Jr., and W.A. Coberly, "The numerical evaluation of the maximum-likelihood estimate of mixture proportions," *Commun. Statist.-Theor. Meth.*, A5 (1976), 1127-1135.
101. B.C. Peters, Jr., and H.F. Walker, "An iterative procedure for obtaining maximum-likelihood estimates of the parameters for a mixture of normal distributions," *SIAM J. Appl. Math.* 35 (1978), 362-378.
102. B.C. Peters, Jr., and H.F. Walker, "The numerical evaluation of the maximum-likelihood estimate of a subset of mixture proportions," *SIAM J. Appl. Math.* 35 (1978), 447-452.
103. H.S. Pollard, "On the relative stability of the median and the arithmetic mean, with particular reference to certain frequency distributions which can be dissected into normal distributions," *Ann. Math. Statist.* 5 (1934), 227-262.
104. E.J. Preston, "A graphical method for the analysis of statistical distributions into two normal components," *Biometrika* 40 (1953), 460-464.
105. R.E. Quandt, "A new approach to estimating switching regressions," *J. Amer. Statist. Assoc.* 67 (1972), 306-310.
106. R.E. Quandt and J.B. Ramsey, "Estimating mixtures of normal distributions and switching regressions," *J. Amer. Statist. Assoc.* 73 (1978), 730-738.
107. C.R. Rao, "The utilization of multiple measurements in problems of biological classification," *J. Royal Statist. Soc., Ser. B*, 10 (1948).
108. R.A. Redner, "An iterative procedure for obtaining maximum likelihood estimates in a mixture model," contract Rep. No. SR-T1-04081, NASA Contract NAS9-14689, Texas A&M University (Sept., 1980).
109. R.A. Redner, "Maximum-likelihood estimation for mixture models," NASA Tech. Memorandum (to appear).

110. R.A. Redner, "Note on the consistency of the maximum-likelihood estimate for nonidentifiable distributions," *Ann. Statist.* 9 (1981), 225-228.
111. P.R. Rider, "The method of moments applied to a mixture of two exponential distributions," *Ann. Math. Statist.* 32 (1961), 143-147.
112. P.R. Rider, "Estimating the parameters of mixed Poisson, binomial, and Weibull distributions by the method of moments," *Bulletin of the International Statistical Institute* 39, Part 2 (1962), 225-232.
113. C.A. Robertson and J.G. Fryer, "The bias and accuracy of moment estimators," *Biometrika* 57, Part 1 (1970), 57-65.
114. J.W. Sammon, Jr., "An adaptive technique for multiple signal detection and identification," in Pattern Recognition, L.N. Kanal, Ed., Thompson Book Co. (1968), 409-439.
115. W. Schilling, "A frequency distribution represented as the sum of two Poisson distributions," *J. Amer. Statist. Assoc.* 42 (1947), 407-424.
116. J. Sittig, "Superpositie van twee frequentieverdelingen," *Statistica* 2 (1948), 206-227.
117. D. F. Stanat, "Unsupervised learning of mixtures of probability functions," in Pattern Recognition, L.N. Kanal, Ed., Thompson Book Co. (1968), 357-389.
118. G.W. Stewart, Introduction to Matrix Computations, Academic Press, New York (1973).
119. B. Strömberg, "Tables and diagrams for dissecting a frequency curve into components by the half-invariant method," *Skand. Aktuarietidskrift* 17 (1934), 7-54.
120. W.Y. Tan and W.C. Chang, "Convolution approach to genetic analysis of quantitative characters of self-fertilized population," *Biometrics* 28 (1972), 1073-1090.
121. W.Y. Tan and W.C. Chang, "Some comparisons of the method of moments and the method of maximum-likelihood in estimating parameters of a mixture of two normal densities," *J. Amer. Statist. Assoc.* 67 (1972), 702-708.
122. R.D. Tarone and G. Gruenlage, "A note on the uniqueness of roots of the likelihood equations for vector-valued parameters," *J. Amer. Statist. Assoc.* 70 (1975), 903-904.
123. H. Teicher, "On the mixture of distributions," *Ann. Math.*

- Statist. 31 (1960), 55-73.
124. H. Teicher, "Identifiability of mixtures," Ann. Math. Statist. 32 (1961), 244-248.
 125. H. Teicher, "Identifiability of finite mixtures," Ann. Math. Statist. 34 (1963), 1265-1269.
 126. H. Teicher, "Identifiability of mixtures of product measures," Ann. Math. Statist. 38 (1967), 1300-1302.
 127. J.D. Tubbs and W.A. Coberly, "An empirical sensitivity study of mixture proportion estimators," Commun. Statist.-Theor. Meth. A5 (12), 1115-1125 (1976).
 128. J. Van Ryzin, Ed., Classification and Clustering, proceedings of an advanced seminar conducted by the Mathematics Research Center, The University of Wisconsin, Madison (May, 1976), Academic Press, New York (1977).
 129. A. Wald, "Note on the consistency of the maximum-likelihood estimate," Ann. Math. Statist. 20 (1949), 595-600.
 130. H.F. Walker, "Estimating the proportions of two populations in a mixture using linear maps," Commun. Statist.-Theor. Meth. A9 (1980), 837-849.
 131. K. Weichselberger, "Über ein graphisches Verfahren zur Trennung von Mischverteilungen and zur Identifikation kupierter Normalverteilungen bei grossem Stichprobenumfang," Metrika 4 (1951), 178-229.
 132. J.H. Wolfe, "Pattern clustering by multivariate mixture analysis," Multivariate Behavioral Research 5 (1970), 329-350.
 133. J. Wolfowitz, "The minimum distance method," Ann. Math. Statist. 28 (1957), 75-88.
 134. C.-P. Wu, "On the convergence of the EM algorithm," Dept of Statistics Tech. Rep. No. 642, University of Wisconsin, Madison, Wisc. (May, 1981).
 135. S. J. Yakowitz, "A consistent estimator for the identification of finite mixtures," Ann. Math. Stat. 40 (1969), 1728-1735.
 136. S.J. Yakowitz, "Unsupervised learning and the identification of finite mixtures," IEEE Trans. Info. Theory IT-16 (1970), 330-338.
 137. S.J. Yakowitz and J.D. Spragins, "On the identifiability of finite mixtures," Ann. Math. Stat. 39 (1968), 209-214.

138. T.Y. Young and T.W. Calvert, Classification, Estimation, and Pattern Recognition, American Elsevier Publishing Co., New York (1973).
139. T.Y. Young and G. Coraluppi, "Stochastic estimation of a mixture of normal density functions using an information criterion," IEEE Trans. Inform. Theory IT-16 (1970), 258-263.
140. S. Zacks, The Theory of Statistical Inference, John Wiley and Sons, Inc., New York (1971).
141. W. Zangwill, Nonlinear Programming: A Unified Approach, Prentice-Hall, Englewood Cliffs, N.J. (1969).

N83

15787

UNCLAS

7
L 1483 15787

ORIGINAL PAGE IS
OF POOR QUALITY

REGRESSION METHODS FOR SPATIAL DATA

by

S. J. Yakowitz

and

F. Szidarovszky

Please Mail Correspondence to:

Dr. Sidney Yakowitz
Department of Systems &
Industrial Engineering
University of Arizona
Tucson, Arizona 85721

OF POOR QUALITY

ABSTRACT

"Kriging" is the name of a parametric regression method used by hydrologists and mining engineers, among others. Features of the kriging approach are that it also provides an error estimate and that it can conveniently be employed also to estimate the integral of the regression function.

In the present work, we describe the kriging method and explore some of its statistical characteristics. Also, some extensions of the Watson method are made and theory so that it too displays the kriging features. Theoretical and computational comparisons of the kriging and Watson approaches are offered.

Regression Methods for Spatial Data*

by S. Yakowitz and F. Szidarovszky
Systems & Industrial Engineering Department
University of Arizona

1. Background and Scope of this Study

Specialists in hydrology, mining, petroleum engineering, and other geoscience-based subjects have recently exhibited considerable interest and enthusiasm for a methodology known as "kriging". To name only a few recent (mostly water-resource oriented) works, we mention in this regard, Bakr et al (1978), Chirlin and Dagan (1980), David (1977), Delhomme (1978, 1979), Dendrou and Houstis (1978), Gambolati and Giampero (1979), Gambolati and Volpi (1979), Gelhar et al (1979), Huijbregts (1978), Journel (1974, 1977), Journel and Huijbregts (1978), and Villeneuve et al (1979). The name "Kriging" derives, according to Journel (1977), from Krige (1951), where the basic idea was first outlined. Matheron (1963) should be credited with its early dissemination. In the present section, we will carefully examine the statistical problems which the kriging method is intended to solve, and in Section II, we will reveal the popular kriging algorithms themselves and derive their properties. It turns out that there are certain unsatisfactory aspects to the current kriging techniques, and yet prior to the present study, they appear to be the only methods appropriate for the problems in their domain. However, methods of nonparametric regression are certainly somewhat relevant. In Section III, we have provided an extension of nonparametric regression theory to increase its relevance to kriging problems.

*Some of the results of this work have been announced in Szidarovszky and Yakowitz (1981), and also in a presentation at the SIAM Conference on Deep Mineral Exploration, Tucson, October, 1981.

Let $f(x)$ and $n(x)$ be uncorrelated real-valued functions defined on a domain X in R^m . Suppose $\{(x_i, y_i)\}_{i=1}^N$ is a sequence of "noisy" function pairs; that is, suppose

$$y_i = f(x_i) + n(x_i), \quad 1 \leq i \leq N. \quad (1.1)$$

The interpretation is that $f(x)$ is a function whose values are to be estimated, and $n(x)$ represents a noise if a measurement is taken at position x . We discuss below two problems which are central to the kriging literature:

Problem 1: Let $x^* \in X$ be specified. It may or may not be among the sample pairs. On the basis of the sample pairs $\{(x_i, y_i)\}_{i=1}^N$,

- (a) Provide an estimate $f_N(x^*)$ of $f(x^*)$, and
- (b) Provide an estimate of the expected squared error

$$E[(f_N(x^*) - f(x^*))^2 | x_1, \dots, x_N]. \quad (1.2)$$

Remarks. The goal of part (a) coincides with the objectives of "nonparametric regression" methods, but to our knowledge, investigators in this latter area have not concerned themselves with task (b). Because practitioners desire to estimate piezometric head in oil and water aquifers or the grade of an ore body as a function of position, the dimension m of the domain X is often 2 or 3.

Problem 2. Let $\{(x_i, y_i)\}_{i=1}^N$ be as above and let D be a subregion of domain X .

- (a) Estimate the integral $\int_D f(x) dx$, and
- (b) Provide a formula for the (sample-dependent) expected square error of this estimate.

Remark. An application motivating Problem 2 is that of estimating the total weight of metal which can be extracted from the ore body occupying volume D, given imperfect assay estimates of the grade at distinct locations.

Problems 1 and 2 seem to have their roots in the forestry and geostatistics literature. In fact, it seems that "geostatistics" is almost synonymous with kriging. We have no doubt that problems

1 and 2 are important and interesting. In this connection, in a review of a geostatistics book, Watson (1977) has written, "The time is certainly ripe for a more serious attack on the estimation of the earths' resources,..."

2. Introduction to Kriging Methodology

In the kriging approach, it is presumed that $f(x)$ and $n(x)$ in (1.1) are realizations of stochastic processes uncorrelated from one another with finite second moments. It is further assumed that $f(x)$ is a realization of an intrinsic random function (IRF); that is, for some functions $\{\phi_j(x)\}_{j=1}^J$ known to the user and perhaps unknown constants a_1, \dots, a_J , for all x, h such that $x, x+h \in X$,

$$E[f(x+h)-f(x)] = \sum_{j=1}^J a_j (\phi_j(x+h) - \phi_j(x)) \quad (2.1)$$

and, independently of x with "var" signifying "variance",

$$1/2 \text{ var } [f(x+h) - f(x)] = \gamma(h). \quad (2.2)$$

The constants $\{a_j; 1 \leq j \leq J\}$ and the function $\gamma(h)$ are quantities which must be inferred from the data $\{(x_i, y_i)\}_{i=1}^N$. In what follows, it is presumed always that $J \leq N$. The function $\gamma(h)$ is called the variogram. Even in the case in which the mean $E[f(x)]$ is known to be constant in x (i.e., $J=1, \phi_1 = 1$), the hypothesis of

"intrinsic random function" is weaker than second-order stationarity. For example, Brownian motion is an intrinsic random function, but it is well-known to be a nonstationary process.

The kriging method is composed of two activities, (i) Inferring the variogram from the data, and (ii) Assuming that the inferred variogram is indeed exact, providing a best linear unbiased estimator and associated error variance, as required by Problem 1 or Problem 2.

Activity (ii) is a standard least-squared problem, and is consequently by far the best understood of the two facets of kriging. There are some inconsistencies in the fundamental definitions and results in the kriging literature. For example, the definitions of "intrinsic random function" given by David (1977) and Matheron (1971) do not coincide. The discussions of noise and the "nugget effect" have likewise been inconsistent. The equations for kriging in the presence of noise as given by Rendu (1980), for example, agrees with our calculations, but differs from formulas offered by other authors (e.g., Journé (1978)). In view of these inconsistencies, we have elected to derive the "universal kriging" equations for prediction with known variogram from first principles.

ORIGINAL PAGE IS
OF POOR QUALITY

Linear estimation from known variograms

To begin with, suppose the noise term in (1.1) is zero. Let us assume that the variogram $\gamma(h)$ and the mean function components $\{\phi_i(x)\}$, of the expectation (2.1) are given. The assumption that one of these functions, say ϕ_1 , is 1, seems to be a universal and perhaps unavoidable assumption which we also will adopt. To begin with, let us discuss the solution of Problem 1. The objective is to choose the parameters $\{\lambda_i\}_{i=1}^N$ so that the linear estimator

$$f_N(x^*) = \lambda_1 Y_1 + \lambda_2 Y_2 + \dots + \lambda_N Y_N \quad (2.3)$$

minimizes

$$E\{[f(x^*) - f_N(x^*)]^2\}, \quad (2.4a)$$

subject to

$$E[f_N(x^*)] = E[f(x^*)]. \quad (2.4b)$$

ON FOUR POINTS

In view of the assumed form (2.1) of the mean value function, a sufficient (but not necessary) condition for the unbiasedness equation (2.4b) to hold is that

$$\sum_{i=1}^N \lambda_i \phi_j(x_i) = \phi_j(x^*), \quad 1 \leq j \leq J. \quad (2.5)$$

Equation (2.5) with $\phi_1 = 1$, implies that

$$\sum_{i=1}^N \lambda_i = 1. \quad (2.6)$$

Use this fact, with the unbiasedness of the estimator $f_N(x^*)$ of $f(x^*) = f^*$, to conclude that, with "cov" signifying "covariance",

$$\begin{aligned} E\left[\left(f^* - \sum_{i=1}^N \lambda_i y_i\right)^2\right] &= \text{var}\left(f^* - \sum_{i=1}^N \lambda_i y_i\right) \\ &= \text{var}\left(\sum \lambda_i (f^* - y_i)\right) \\ &= \sum_i \sum_j \lambda_i \lambda_j \text{cov}[(f^* - y_i), (f^* - y_j)]. \end{aligned} \quad (2.7)$$

Now observe that

$$\begin{aligned} \text{cov}[(f^* - y_i), (f^* - y_j)] &= 1/2 [-\text{var}((f^* - y_i) - (f^* - y_j)) \\ &\quad + \text{var}(f^* - y_i) + \text{var}(f^* - y_j)] \quad (2.8) \\ &= -\gamma(x_i - x_j) + \gamma(x^* - x_i) + \gamma(x^* - x_j). \end{aligned}$$

One makes these substitutions into (2.7) and after some algebra, sees that the Lagrange multiplier technique for minimizing $E[(f(x^*) - f_N(x^*))^2]$ subject to (2.5) yields

$$\sum_{k=1}^N \lambda_k \gamma(x_1 - x_k) = \gamma(x_1 - x^*) + \sum_{j=1}^J \mu_j \phi_j(x_1), \quad 1 \leq 1 \leq N \quad (2.9a)$$

$$\sum_{i=1}^N \lambda_i \phi_j(x_i) = \phi_j(x^*), \quad 1 \leq j \leq J. \quad (2.9b)$$

OF POOR QUALITY

The variables μ_j are the Lagrange multipliers. Journal (1977) calls the above linear equation the universal kriging system.

From substitution according to (2.9) into (2.7), one concludes that the mean squared prediction error is given by

$$E[(f^* - f_N(x^*))^2] = \sum_{i=1}^N \lambda_i \gamma(x^* - x_i) - \sum_{j=1}^J \mu_j \phi(x^*) \quad (2.10)$$

If the noise term $n(x)$ in (1.1) has zero mean, one accounts for its presence by noting that, because it is presumed uncorrelated from the f -process,

$$\begin{aligned} \text{cov}((f^* - y_i), (f^* - y)) &= \text{cov}((f^* - f_i - n_i), (f^* - f_j - n_j)) \\ &= \text{cov}((f^* - f_i), (f^* - f_j)) + \text{cov}(n_i, n_j). \end{aligned}$$

In the above equation, we have, of course, intended that n_i signify $n(x_i)$. As a result of the above, one readily sees that in the presence of noise (2.9a) should be replaced by (2.9'a):

$$\begin{aligned} \sum_{k=1}^N \lambda_k (\gamma(x_i - x_k) - 2\text{cov}(n(x_i), n(x_k))) &= \gamma(x_i - x^*) \\ &+ \sum_{j=1}^J \mu_j \phi_j(x_i), \quad 1 \leq i \leq N. \end{aligned} \quad (2.9'a)$$

Let us now investigate the modifications necessary for solution of Problem 2 described in Section 1. Assume $\int_D dx = 1$. In this case, we ^{replace} the objective (2.4) by the task of minimizing $\int_D f(x) dx$.

$$E\left[\left(\int_D f(x) dx - \sum \lambda_1 y_1\right)^2\right] \quad (2.11a)$$

subject to

$$E[\sum \lambda_1 y_1] = E\left[\int_D f(x) dx\right]. \quad (2.11b)$$

The preceding kriging analysis leads, in the integral estimation case, to the following universal kriging system:

$$\sum_{k=1}^N \lambda_k \gamma(x_i - x_k) = \int_D \gamma(x_i - x) dx + \sum_{j=1}^J \mu_j \phi_j(x_i), \quad 1 \leq i \leq N, \quad (2.12)$$

$$\sum_{i=1}^N \lambda_i \phi_j(x_i) = \int_D \phi_j(x) dx, \quad 1 \leq j \leq J.$$

The expected squared error of the integral estimate is given by

$$E\left[\left(\int_D f(x) dx - \sum_{i=1}^N \lambda_i Y_i\right)^2\right] = \sum_{i=1}^N \lambda_i \int_D \gamma(x_i - x) dx - \sum_{j=1}^J \mu_j \int_D \phi_j(x) dx - \int_D \int_D \gamma(x - x') dx dx' \quad (2.13)$$

Inference of the variogram

The task of inferring a covariance function or power spectral density from data is known by experienced statisticians to be somewhat delicate, and one which furthermore requires a considerable quantity of data. The subtleties of the covariance inference problem translate directly to the task of inferring a variogram from data.

There are some very real difficulties with variogram estimation in the published kriging applications. To avoid effects of "non-stationarity", practitioners tend to have a single variogram apply only to a relatively small region X of domain points of $f(x)$. Moreover they have not developed procedures to ascertain whether the intrinsic random function hypothesis is tenable for their applications. A particular difficulty is that in the bounded domain case, ergodic theorems are inapplicable to the task of demonstrating consistency. To our knowledge, with the exception of certain extreme cases such as white noise, no methods for inferring the covariance function from sample pairs $\{(x_i, f(x_i))\}$, $f(\cdot)$ a fixed sample function,

ORIGINAL PAGES
OF POOR QUALITY

are known to be consistent.

We now concern ourselves with outlining the present practice with regard to variogram inference. The recommended procedure is to choose a ^{parametric} family of variograms from the five or six popular families mentioned in the literature, and then to select the variogram from the chosen family which agrees best, in some sense, with the covariance function constructed from the data $\{(x_i, y_i)\}_{i=1}^N$. We list in Table 2.1 some of the prominent variogram families.

Monomial	$\gamma_\theta(h) = \omega h ^a$
Spherical	$\gamma_\theta(h) = \begin{cases} \omega \left[\frac{3}{2} \frac{ h }{a} - \frac{1}{2} \left(\frac{ h }{a} \right)^3 \right] & h \leq a \\ \omega, & h > a \end{cases}$
Exponential	$\gamma_\theta(h) = \omega [1 - \exp(- h /a)]$
Gaussian	$\gamma_\theta(h) = \omega [1 - \exp(- h ^2/a^2)]$
where $\theta = (a, \omega)$	

Table 2.1

A Listing of Popular
Variogram Families

There seems to be no consensus in the literature on methodology for the selection of a parametric family from Table 2.1 on the basis of an observed sample $\{(x_i, y_i)\}_{i=1}^n$. Some heuristic approaches are proposed by David (1977). Concerning the task of selection of the member $\gamma_\theta(h)$ the foremost criteria seem to be (1) least squares,

ORIGINAL PAGE IS
OF POOR QUALITY

and a geometric procedure (David (1977)).
(ii) cross validation, / In the least squares approach, one selects the parameter θ^* so as to minimize

$$I_1(\theta) = \sum_v (\gamma_n(h_v) - \gamma_\theta(h_v))^2,$$

the index v running over some finite collection of arguments h_v and $\gamma_n(h)$ being some sample approximation to the variogram, such as

$$h_n(h) = 1/2N(h) \sum_{j=1}^{N(h)} (y_j - y_j(h))^2$$

where $j(h)$ is an index selected so that $x_j - x_{j(h)} = h$ and $N(h)$ is the number of such points selected. "drift" is

If/ though to be present (that is, if $\phi_j, j > 1$, in (2.1) is not zero), then this approach entails some serious conceptual difficulties.

Matheron (1971, Chapter 4) has addressed these difficulties.

The cross-validation approach to parameter selection is as follows.

Let $P(x_j, \theta)$ be the universal kriging estimate of $f(x_j)$ on the basis of the sample points $\{(x_i, y_i)\}_{i \neq j}$ and parametric variogram $\gamma_\theta(h)$.

One then chooses θ^* to minimize the squared error of the predicted values, which is

$$I_2(\theta) = \sum_{j=1}^n (y_j - P(\theta, x_j))^2.$$

Practitioners insist, quite rightly, that one should not select a variogram entirely algorithmically, but with attention also to past experience with similar geological data.

ORIGINAL PAGE IS
OF POOR QUALITY

Convergence and Consistency

With the exception of studies by Matheron (1971,1973), the literature of kriging tends to be practical and pragmatic. Major issues of consistency and convergence rates have not been addressed. In the developments to follow, we attempt to obtain initial results in these areas.

As has been noted earlier, there is no consistent variogram estimator based on observations $\{(x_i, \bar{f}(x_i))\}_{i=1}^N$ for x_i in a bounded domain X and \bar{f} a fixed sample of an intrinsic random function f . In short, the variogram cannot be consistently inferred, even if it is known to be a member of a given family such as listed in Table 2.1. On the other hand, as we will later demonstrate, under certain circumstances, the kriging estimate will converge, with increasing number of samples, to the correct value, even when the variogram is not correct. An interpretation of these remarks is that the kriging method can be effective for estimating values on the basis of noisy samples, but that the associated error estimate need not be consistent. This interpretation is borne out by our simulation studies. The fact that the estimate of the squared error need not become more accurate with increasing data is significant because kriging practitioners and their clients place great value on the error estimation feature.

Let us begin our analysis of convergence of kriging estimate under the simplest of conditions by assuming that

- i) The observations are noiseless ($n(x_1)=0$)
- ii) $\gamma(0) = 0$, and γ is continuous in a neighborhood of the origin.
- iii) There is no "drift"; that is, $J=1$, and $\phi_1 = 1$.
- iv) The "true" variogram is known.

Theorem 2.1. Let X be the domain of the intrinsic random function $f(x)$ and assume the conditions above are in force. If the infinite sequence $\{x_i\}$ is dense in X , then for any $x^* \in X$ and for $f_N(x^*)$ as in (2.3),

$$E[(f(x^*) - f_N(x^*))^2] \rightarrow 0 \text{ as } N \rightarrow \infty. \quad (2.14)$$

Proof. In view of assumption (iii), for every i , $\gamma_i = f(x_i)$ is itself an unbiased linear estimator of $f(x^*)$, and so for $N \geq 1$.

$$E[(f(x^*) - f_N(x_i))^2] \leq E[(f(x^*) - f(x_i))^2] = 2\gamma(x^* - x_i).$$

Let $x^*(N)$ denote the member of $\{x_i\}_{i=1}^N$ which is closest to x^* . By the assumption that $\{x_i\}$ is dense, $x^*(N) \rightarrow x^*$ as $N \rightarrow \infty$, and therefore

$$E[(f(x^*) - f_N(x^*))^2] \leq E[(f(x^*) - f(x^*(N)))^2] = 2\gamma(x^* - x^*(N)). \quad (2.15)$$

The proposition follows by observing that, in light of property (ii), $\gamma(x^* - x^*(N))$ must converge to 0. The bound given by (2.15) may be of some practical interest in itself. □

The Brownian motion process affords an example of a situation in which the best estimate is not consistent unless x^* is an accumulation point of the sample points $\{x_i\}$. For Brownian motion is Markov, and the best estimate of $f(x^*)$ will depend only on the points $(x_a, f(x_a))$ and $(x_b, f(x_b))$, where x_a is the largest domain sample less than x^* and x_b the smallest sample greater than x^* .

There are many common situations in which the hypothesis that $\{x_i\}$ is dense in X will be satisfied. One important case is that in which the x_i 's are selected independently from X according to a measure that assigns positive probability to every open set (such as when the probability density function exists and is positive).

Corollary. Assume that the hypotheses of Proposition 1 are in force and additionally that X is open, and has finite Lebesgue measure $\gamma(h)$ has a continuous second derivative and the samples $\{x_i\}$ are/identically and independently distributed on X with pdf bounded away from 0 / in a neighborhood of x^* . Then for some fixed constant C and all N ,

$$E[(f(x^*) - f_N(x^*))^2] < C/(N^{(2/m)}), \quad (2.16)$$


m being the dimension of the space containing X .

Proof. Since $\gamma(h)$ is an even function, its first derivative or gradient must be 0, and we have

$$\begin{aligned} \gamma(x^* - x^*(N)) &= (1/2) (x^* - x(N))^T \gamma^{(2)}(G) (x^* - x(N)) \\ &+ o(\|x^* - x^*(N)\|^2) \end{aligned} \quad (2.17)$$

It is known (e.g., Yakowitz et al (1978), p. 1299) that under the independent, uniformly distributed sample case, for all points $x^* \in X$ and some constant C_1 ,

$$E[\|x^* - x^*(N)\|^2] < C_1 / (N^{(2/m)}), \quad N=1,2,\dots \quad (2.18)$$

From the argument in that reference, one can conclude that (2.18) holds whenever the pdf is bounded away from 0 in a neighborhood of x^* . The Corollary now follows from (2.17) and (2.18). 

From our experience in groundwater analysis, where the domain points correspond to well locations, the hypotheses of the corollary are of some use. On the other hand, for some ore sampling strategies, it may be more reasonable to assume that the x_i 's form a grid of similar-sized rectangles. For such regular patterns, one may conclude that (2.18) is true without expectations, and hence the conclusions of the Corollary remains valid.

We will now discuss convergence of the kriging estimate when accounting for drift. Assume that x^* is a limit point of $\{x_i\}$. Assume furthermore that for some subsequence x_{n_1}, \dots, x_{n_J} the matrix $\underline{\Delta} \triangleq \{\phi_i(x_{n_j})\}_{i,j=1}^J$ is nonsingular. (Otherwise, there is no hope of being able to obtain estimates satisfying (2.5) for arbitrary $\phi_i(x^*)$ values.) For $N > n_J$, define the linear estimate

$$\tilde{f}_N(x^*) = (1 - \alpha_N) f(x^*(N)) + \alpha_N \sum_{i=1}^J \lambda_N^i f(x_{n_i}), \quad (2.19)$$

where $x^*(N)$ is, as before, the nearest neighbor (among the first N samples) to x^* , and

$$\alpha_N \triangleq ||x^* - x^*(N)||. \quad (2.20)$$

In order to assure that the constraint condition (2.5) holds, we set $\underline{\phi}(x) = (\phi_1(x), \dots, \phi_J(x))^T$ and determine $\underline{\lambda}_N = (\lambda_N^1, \dots, \lambda_N^J)^T$ by

$$\alpha_N \underline{\phi} \underline{\lambda}_N = \underline{\phi}(x^*) - (1 - \alpha_N) \underline{\phi}(x^*(N)). \quad (2.21)$$

The consistency of the estimate $\tilde{f}_N(x^*)$ will follow if only we can show that the sequence $\{\underline{\lambda}_N\}$ remains in a bounded region. Toward that end, note that after taking a Taylor's series expansion of $\underline{\phi}(x^*) - \underline{\phi}(x^*(N))$ and dividing by α_N , we may rewrite (2.21) as

$$\begin{aligned} \underline{\phi} \underline{\lambda}_N &= \underline{\phi}(x^*) - (1/\alpha_N) \nabla \underline{\phi}(x^*(N) - x^*) \\ &\quad + 1/\alpha_N O(||x^* - x^*(N)||), \end{aligned} \quad (2.22)$$

where the matrix

$$\underline{\Phi} \triangleq \begin{pmatrix} \nabla \phi_1(x^*) \\ \vdots \\ \nabla \phi_J(x^*) \end{pmatrix}. \quad (2.23)$$

ORIGINAL PAGE IS
OF POOR QUALITY

From (2.22), we see that $\underline{\lambda}_N$ remains bounded when α_N is chosen according to (2.20). In fact,

$$\|\underline{\lambda}_N\| \leq \|\underline{\phi}^{-1}\| [\|\nabla \underline{\phi}\| + \|\underline{\phi}(x^*)\|] + o(1).$$

We have demonstrated above that the constrained linear predictor $f_N(x^*)$ converges to $f(x^*)$. By an earlier argument, this, in turn, implies that $f_N(x^*)$, the best linear unbiased estimator must likewise converge.

The interested reader can apply the convergence analysis of the preceding discussion to achieve convergence bounds/using structures for estimation with drift by the preceding in the proof of Corollary.

Our attention now turns to the case that noise $n(x_i)$ is present in the observation law (1.1). For simplicity, assume that $J=1$, and $\phi_1=1$. If $n(\cdot)$ is a continuous function, then apparently consistent identification of $f(x^*)$ is not possible since local samples cannot distinguish between the effects of signal and noise. However, the linear estimate provided by the universal kriging equations is an appropriate procedure and in fact coincides with what is known to communication engineers as a "smoothing filter". If $\{n(x_i)\}$ are independent variables, then, as we now demonstrate, under some circumstances, consistent estimation of $f(x^*)$ is possible. Toward verifying this assertion, as in earlier arguments, we find a linear estimator whose properties are understood, and then appeal to the fact that since the kriging estimate is optimal in the least squares sense, it must be at least as good as the estimator under consideration.

For the particular task at hand of verifying consistency in the presence of independent noise, it is sufficient to call attention to the fact that Stone (1977) has discussed a general class of nonparametric regression (NPR) formulas of the form

$$\hat{f}_N(x^*) = \sum_{i=1}^N y_i w_{i,N}(x^*; x_1, \dots, x_N).$$

The weights $w_{i,N}$ can be taken to add to 1 (i.e., $\sum_{i=1}^N w_{i,N} = 1$), so the unbiasedness condition (2.5), with $J=1$ holds.

His results imply that if x^* and x_i are i.i.d. observations, and if $f(\cdot)$ is measurable, and provided the weight functions $w_{i,N}(\cdot)$ satisfy certain natural properties, then $\hat{f}_N(x^*) \rightarrow f(x^*)$ in the mean.

Toward applying Stone's results to the issue of consistency of kriging estimates in the noisy observation case, let $\bar{f}(\cdot)$ denote a realization of the intrinsic random function $f(\cdot)$. Then if $y_i = \bar{f}(x_i) + n(x_i)$, the sequence $\{(x_i, y_i)\}$ constitutes i.i.d. observations and the hypotheses of Stone's convergence result holds, provided $\bar{f}(\cdot)$ is so much as measurable and a few technical assumptions of little practical concern hold. So we conclude that

$$E[(\hat{f}_N(x^*) - \bar{f}(x^*))^2] \rightarrow 0, \text{ as } N \rightarrow \infty.$$

It may be concluded that if the noise measurements and the sample functions \bar{f} are uniformly bounded, then convergence occurs without the conditioning/ alternatively, without the boundedness assumption, one can assert that convergence in the mean is assured outside an f -set of any positive measure. From results in the next section, it may be seen that if one is willing to assume that the sample functions are twice-continuously differentiable, then convergence in the mean is on the entire f -space without the set qualification. Convergence in the mean of the linear estimate \hat{f}_N implies, of course, mean convergence of the kriging estimate.

For certain specific NPR estimates, rates of convergence are known (e.g., Fisher and Yakowitz (1976), Parthasarathy and Bhattacharya (1961), Sacks and Spiegelman (1980), Schuster and Yakowitz (1979)). The strongest results related to convergence of point NPR estimates known to us are that of Schuster (1972) for one-dimensional x_i 's, and for m -dimensional x_i 's, the result to be demonstrated in the next section, that for $m_N(x^*)$ the Watson NPR estimate for $f(x^*)$, that with some provisos to be specified in Section 3,

$$E[(m_N(x^*) - f(x^*))^2] = O(n^{-(1/(m/4 + 1))}). \quad (2.24)$$

This convergence rate will be seen to be optimal, in a certain sense.

In evaluating the convergence statements concerning kriging up to this point, it should be emphasized that they are valid only if $f(\cdot)$ really is an intrinsic random function and the variogram and drift functions are known perfectly.

Our next discussion of kriging convergence is directed at Problem 2 of Section 1, i.e., the integral estimation problem. For Problem 2, as has been observed earlier, one must modify the universal kriging equation development by replacing f^* in (2.7) by $\int_D f(x) dx$. The effect of this substitution is that $\gamma(x_i - x^*)$ and $\phi_j(x^*)$ are replaced by $\int_D \gamma(x_i - x) dx$ and $\int_D \phi_j(x) dx$ in (2.9a) and (2.9b), respectively.

Let $I(f)$ denote the universal kriging estimate of $\int_D f(x) dx$ obtained by the modifications just mentioned, and let $f_N(x^*)$ denote the kriging estimate of $f(x^*)$. Recall our assumption that $\int_D f dx = 1$. Then we have the following

Theorem 2.2.

$$I(f) = \int_D f_N(x) dx. \quad (2.25)$$

Proof. One may express (2.9a,b) in matrix form as

$$\underline{\lambda}(x^*) = \underline{A}^{-1} \underline{c}(x^*) \quad (2.26)$$

where $\underline{\lambda}(x^*) = (\lambda_1, \dots, \lambda_N, \mu_1, \dots, \mu_J)^T$, $c_1(x^*) = \gamma(x_i - x^*)$, $1 \leq i \leq N$,

$$c_{i+N}(x^*) = \phi_i(x^*), \quad 1 \leq i \leq J,$$

and

$$A_{ij} = \gamma(x_i - x_j), \quad 1 \leq i, j \leq N; \quad A_{j+N, i} = A_{i, j+N} = \phi_j(x_i), \quad 1 \leq i \leq N, \quad 1 \leq j \leq J.$$

From (2.3), we see that if we define $\underline{\beta} = (f(x_1), \dots, f(x_N), 0, \dots, 0)$, then

$$f(x^*) = \underline{\beta} \underline{A}^{-1} \underline{c}(x^*). \quad (2.27)$$

Now it is clear from (2.12) that for the integration problem, universal kriging equation may be represented as

$$I(f) = \underline{\beta} \underline{A}^{-1} \int_D \underline{c}(x) dx = \int_D \underline{\beta} \underline{A}^{-1} \underline{c}(x) dx = \int_D f_N(x) dx \quad (2.28)$$

and our proposition is established. □

The predicted mean square error was given in (2.12). But the following evident result is useful:

Corollary:

$$E\left[\left(I(f) - \int_D f(x) dx\right)^2\right] \leq \left[\int_D^{1/2} [(f_N(x) - f(x))^2] dx\right]^{1/2} \quad (2.29)$$

Proof.

$$\begin{aligned} E\left[\left(\int_D f_N(x) - \int_D f(x)\right)^2\right] &= \iint \text{cov}(f(x) - f_N(x), f(\hat{x}) - f_N(\hat{x})) dx d\hat{x} \\ &\leq \left[\int_D \text{var}(f_N(x))^{1/2} dx\right]^2. \end{aligned} \quad \square$$

From the corollary, it is apparent that earlier bounds with respect to convergence of kriging point estimates can be directly applied to bounding the convergence of integral estimates as the number of sample pairs increases. Moreover the above analysis is applicable to estimates of other linear functionals $L(f)$ of the

DATA QUALITY

The final theoretical topic we shall broach, in connection with kriging convergence, is the effect of incorrect variogram on the estimate $f_N(x^*)$. For simplicity, assume the no drift case. It is fairly clear that if the variogram is in error, there is little hope of estimating $E[(f(x^*) - f_N(x^*))^2]$ correctly.

Example. In this example, we show that it is possible for the kriging predictor to be exact, while the variogram (and hence the error estimate) to contain significant error. Suppose $\gamma_2 = b\gamma_1$, where b is any positive constant. If $\underline{\lambda} = (\lambda_1, \dots, \lambda_n)$ is the minimizer of (2.7), subject to the constraints (2.5), with $\gamma = \gamma_1$, then $\underline{\lambda}$ will also be the constrained minimizer of (2.7) with $\gamma = \gamma_2$. Thus if a presumed variogram is so much as approximately proportional to the correct one, the estimate $f_N(x^*)$ will be reliable. But from (2.10), one sees that (ignoring the drift term) the error estimate under γ_2 will differ from that under γ_1 by the scale factor b .

OF P... ..

Let $\underline{\lambda}^{(1)}$ and $\underline{x}^{(2)}$ be the solutions of the universal kriging equation (2.9a,b) under variograms γ_1 and γ_2 , respectively.

Suppose that for some positive number δ and all h , $||\gamma_1(h) - \gamma_2(h)|| < \delta$.

From a standard numerical analysis formula (e.g., Szidarovszky and Yakowitz (1978), p. 214), we have that for $\delta < ||A||/\Gamma(A)$,

$$||\underline{\lambda}^{(1)} - \underline{\lambda}^{(2)}|| < \Gamma(A) ||\underline{\lambda}^{(1)}|| \delta / (||A|| - \delta \Gamma(A)) \quad (2.30)$$

where A is the matrix determined in connection with (2.26) and

$$\Gamma(A) = ||A|| ||A^{-1}||$$

in the condition number. Some insight into the potential perniciousness of variogram error can be inferred from (2.30) by considering that the linear equation associated with least squares problems frequently is ill-conditioned because of collinearity effects.

This phenomenon is evidenced by large condition number $\Gamma(A)$.

Let E_1 and E_2 be expectations of square differences determined by γ_1 and γ_2 respectively and (2.8). From earlier developments, we can be assured that if the kriging equation (2.9) uses $\gamma = \gamma_2$,

$$E_2[(f(x^*) - f_N(x^*))^2] \rightarrow 0$$

provided only that $\{x_1\}$ is dense in X . In the noiseless case, E_1 and E_2 determine metrics d_1 and d_2 (2.3) on $V = \text{span}\{f(x) : x \in X\}$ according to

$$d_j(y, z) = (E_j[y - z]^2)^{1/2}, \quad j = 1, 2; y, z \in V. \quad (2.31)$$

Thus V is the smallest space containing all linear predictors.

The task of finding the circumstances relating to γ_1 and γ_2 , under which d_1 and d_2 determine equivalent topologies remains a subject for future research. At this point, one can quickly confirm that if $V_1 = \{af(x) : a \text{ real}, x \in X\}$ is dense in V , with respect to both

metrics, and if both variograms are continuous and have their unique minima at the origin, then (2.31) implies convergence with respect to E_1 . For under these assumptions, for any unbiased linear predictor $f_N(x^*)$ in V , there is a domain point x^N and a random variable $f(x^N) \in V_1$ such that

$$d_j(f(x^N), f_N(x^*)) \leq 1/N, \quad j = 1, 2. \quad (2.32)$$

Note that in view of (2.31),

$$d_j(f(x^N), f(x^*)) = 2\gamma_j(x^N - x^*), \quad j=1, 2. \quad (2.33)$$

So if $d_1(f(x^*), f_N(x^*)) \rightarrow 0$, then $x^N \rightarrow x^*$. Now (2.32) and (2.33) imply that $d_2(f(x^*), f_N(x^*)) \rightarrow 0$.

Unfortunately, it is not always the case that convergence in d_2 implies convergence with respect to d_1 .

Example. Let γ_2 be a bounded variogram and γ_1 and unbounded function (as in the Brownian motion case). Suppose that g_N converges to $f(x^*)$ with respect to both metrics d_1 and d_2 . Define

$$f_N(x^*) = (1 - \lambda_N) g_N + \lambda_N f(z_N) \quad (2.34)$$

where λ_N is a sequence of numbers converging to 0 and z_N are points in V such that $\gamma_1(x^* - z_N) > 1/\lambda_N$. Then still $d_2(f(x^*), f_N(x^*)) \rightarrow 0$, but $d_1(f(x^*), f_N(x^*))$ will be approximated by $(\lambda_N \gamma_1(x^* - z_N))$, which is bounded away from 0. An unsettling aspect of this example is that one may very well anticipate that for relatively large sample sizes, least squares estimates may well assign some weight to samples x_n with domain points far from x^* . These points could cause havoc if $E[(f(x_n) - f(x^*))^2]$ is much larger than anticipated. In kriging practice, it is common to assume that variograms are bounded by "sills".

3. Nonparametric Regression Application

The intention of the present section is to reveal extensions of nonparametric regression which make this approach more suited to Problems 1 and 2 of Section 1. In the section to follow, a comparison of properties of kriging with those of nonparametric regression will be offered.

The particular nonparametric regression (NPR) method to be investigated here is the kernel estimator proposed by Watson (1964). The two developments revealed here are (i) a formula for the asymptotic expected square error, and (ii) a data-based approximation of the mean squared error. The discussion closes by showing that the asymptotic convergence of the NPR estimates is, in a certain sense, optimal.

Let (X, Y) denote jointly distributed random variables. The dimension m of X is arbitrary, but Y is real. Nonparametric regression methods are intended for the problem of inferring the conditional expectation (i.e., regression function)

$$m(x) = E\{Y | X=x\} \quad (3.1)$$

on the basis of an observation $\{(x_1, y_1), \dots, (x_N, y_N)\}$ of the sequence (X_i, Y_i) .

To begin with, let us phrase Problem 1 of Section 1 in NPR terms. One presumes that the random vector (X, Y) satisfies

$$Y = m(X) + n(X), \quad (3.2)$$

where m is a fixed deterministic but unknown "regression function". In NPR, the distribution of the noise $n(x)$ may depend on the domain point x .

The Watson NPR estimator is

$$m_N(x) = \frac{\sum_{i=1}^N y_i k((x_i - x)/a_N)}{D_N(x)} \quad (3.3)$$

where $D_N(x) = \sum_{i=1}^N k((x_i - x)/a_N)$, a_N is a positive number, and $k(\cdot)$ is a probability density function chosen by the user. By way of convergence results, it is known (Schuster and Yakowitz (1979)) that if $\{a_N\}$, $k(x)$, and the (X,Y) variable satisfy certain lenient conditions, then for any given $\epsilon > 0$, there is some constant C such that for every N ,

$$P[\sup_x |m_N(x) - m(x)| > \epsilon] < C/(Na_N^2). \quad (3.4)$$

It is often not practical to compute the constant C and in any event, the bound above is typically pessimistic.

Since in kriging squared-error is the essence, our analysis at this point is directed toward establishing the behavior of $E[(m_N(x) - m(x))^2]$ as the number N of observations increases. Toward that end, let $h(x,y)$ and $g(x)$ be the pdf's of (X,Y) and X , respectively. Let $w(x) = \int y h(x,y) dy$, thus $m(x) = w(x)/g(x)$, and define for some m -tuple x , and $1 \leq i \leq N$

$$\begin{aligned}
 U_{Ni} &= k((x^* - x_i)/a_N)/a_N^m, & \text{ORIGINAL PAGE IS} \\
 & & \text{OF POOR QUALITY} \\
 V_{Ni} &= Y_i U_{Ni}, \\
 U_N &= 1/N \sum U_{Ni} : 1 \leq i \leq N, \\
 V_N &= 1/N \sum V_{Ni} : 1 \leq i \leq N.
 \end{aligned}
 \tag{3.5}$$

Throughout this section, we will assume that the kernel pdf $k(u)$ is selected so as to satisfy the properties (i) to (iv) below:

- (i) $k(u)$ and $\|uk(u)\|$ are bounded,
- (ii) $\int uk(u) du = 0$,
- (iii) $\int \|u\|^2 k(u) du < \infty$,
- (iv) the functions $g(x)$ and $w(x)$ are twice continuously differentiable and the second partial derivatives of $g(x)$ are bounded,
- (v) the second moment of Y is finite.

The pdf of the multivariate normal law satisfies properties (i) to (iii).

The convergence facts we will need are given in the statement below.

Theorem 1. Let m be the dimension of the sample vectors x_1, x_2, \dots and assume $g(x^*) > 0$. Then

- (a) $\text{var}(V_N)$ and $\text{var}(U_N)$ are both $O(1/(N a_N^m))$,
- (b) $(E[V_N] - w(x^*))$ and $(E[U_N] - g(x^*))$ are both $O(a_N^2)$.
- (c) If $a_N = N^{-(1/(m+4))}$, then for some sequence of events E_N such that $P[E_N] \rightarrow 1$,

$$E[(m_N(x^*) - m(x^*))^2 | E_N] = O(N^{-(1 + m/4)^{-1}}). \tag{3.6}$$

Proof. This theorem is very much inspired by developments of Senuster (1972). Thus part (a) is essentially formula (1) in the proof of his Lemma 1, but extended here to m variables. In

particular, after a change of variables to $u = (x_i - x^*)/a_N$ we have

$$\begin{aligned} E[(U_{Ni})^2] &= a_N^{-m} \int k(-u)^2 g(x^* - a_N u) du \\ &= \left(g(x^*)/a_N^m \right) \left[\int k^2(u) du + o(a_N) \right]. \end{aligned}$$

Similarly, one may confirm that

$$E[U_{Ni}] = g(x^*) \left[\int k(u) du + o(a_N) \right].$$

Now use that the variables are uncorrelated to get $\text{var}(U_N) = o((a_N^m N)^{-1})$. The demonstration for $\text{var}(V_N)$ proceeds in the same fashion. The proof of part (b) is essentially that of the first part of Lemma 2 in Schuster (1972). Thus after the change in variable, and use of assumed property (11) above,

$$\begin{aligned} E[U_{N1}] - g(x^*) &= \int k(-u) [g(x^* - a_N u) - g(x^*)] du \\ &\leq (a_N^2/2) \sup_x |g''(x)| \int |u|^2 k(u) du = o(a_N^2) \end{aligned}$$

Clearly, U_N and U_{N1} have the same expectation. The analysis of $E[V_N]$ proceeds in a similar fashion.

Toward demonstration of (c), define E_N to be the event that

$$U_N > (1/2)g(x^*) \text{ and } V_N \leq 2w(x^*).$$

In view of parts a and b and Chebyshev's inequality, the probability of E_N converges to 1. Also, note that $\text{var}(U_N | E_N) \leq \text{var}(U_N) / P[E_N]$ and

$$\text{var}(V_N | E_N) \leq \text{var}(V_N) / P[E_N].$$

Now under E_N ,

$$\begin{aligned} \pi_N(x^*) - \pi(x^*) &= (V_N g(x^*) - U_N w(x^*)) / U_N g(x^*) \\ &\leq (2/g(x^*)^2) (|w(x^*) (U_N - g(x^*))| + g(x^*) |V_N - w(x^*)|) \end{aligned}$$

Part (c) now is easily seen to be a consequence of (a) and (b).

OF POOR QUALITY

Our attention now turns to derivation of a data-based estimate of the mean squared error of the NPR point estimate $m_N(x)$:

$$\sigma^2(x) = E[(m_N(x) - m(x))^2 | X_j = x_j, 1 \leq j \leq N]. \quad (3.7)$$

Observe that since the terms $\{V_{Ni}\}_i$ in (3.5) are uncorrelated,

$$\sigma^2(x) \triangleq (1/D_N(x))^2 \left(\sum_{i=1}^N E[n(x_i)^2] k^2((x-x_i)/a_N) \right). \quad (3.8)$$

The only term in (3.8) which is not known to the statistician is $E[n(x_i)^2]$. But this can be approximated from the sample by defining α to be any positive number less than 1 and defining

$$\hat{E}[n(x_i)^2] = 1/N^\alpha \left(\sum_{j \in S(i,N)} (y_j - m_N(x_j))^2 \right), \quad (3.9)$$

where $S(i,N)$ is the set of indices of the N^α nearest neighbors in $\{x_k\}_{k=1}^N$ of x_i . Since in view of (3.4), (3.9) is converging, in N , to $m(x) = E(Y|X=x)$, uniformly in x , since with probability 1, the radii of the sets $S(j,N)$ become vanishingly small as $N \rightarrow \infty$, it is evident that the estimate

$$\hat{\sigma}^2(x) \triangleq (1/D_N(x))^2 \sum_{i=1}^N \hat{E}[n(x_i)^2] k^2((x-x_i)/a_N), \quad (3.10)$$

satisfies the relation

$$\hat{\sigma}^2(x) / \sigma^2(x) \rightarrow 1 \text{ as } N \rightarrow \infty, \text{ i.p.} \quad (3.11)$$

Note that the estimator $\hat{\sigma}^2(x)$ depends solely on the statisticians choices of $k(\cdot)$ and $\{a(n)\}$, and the observed sequence $\{(x_j, y_j)\}$. From the Theorem, one may conclude that if a_N tends to zero slightly faster than $(1/N)^{1/(m+4)}$, then the variance error part of (a) will dominate, yet need not seriously degrade the rate of convergence in (3.6). Under this circumstance, $\hat{\sigma}^2(x)$ will be an asymptotically square

One can confirm that for any $\delta > 0$, as $a_N \rightarrow 0$, the contribution in (3.10) of terms x_i such that $\|x - x_i\| > \delta$, becomes negligible, and in practice, we have found that

$$\hat{\sigma}^2(x) = \hat{E}[n(x)^2], \quad (3.12)$$

gives a reliable approximation of the error variance. Similarly, one can show that for any points x^1, x^2

$$\begin{aligned} \tilde{\text{Cov}}(x, x^1) &= \left[\frac{1}{D_N(x)} \frac{1}{D_N(x^1)} \sum_{i=1}^N k((x - x_i)/a_N) k((x^1 - x_i)/a_N) \right] \\ & \quad \left[\hat{E}(n(x)^2) \hat{E}(n(x^1)^2) \right]^{1/2}, \end{aligned} \quad (3.13)$$

is asymptotically accurate.

This relationship is useful in applying the NPR approach to Problem 2, of Section 1, as is now seen.

Our procedure for approaching Problem 2 is to apply numerical quadrature to the function $m_N(x)$. Specifically, let $\{(t_j, w_j)\}_{j=1}^M$ be quadrature points and weights for integrating over the desired domain D . The nonparametric estimate I_N is then defined by

$$I_N = \sum_{j=1}^M w_j m_N(t_j), \quad (3.14)$$

which is an estimate of $I = \sum_{j=1}^M w_j m(t_j)$.

The error

$$I_N - \int_D m(x) dx$$

has two components,

$$I_N - \int_D m(x) dx = [I_N - I] + [I - \int_D m(x) dx], \quad (3.15)$$

the first bracketed term being the error due to approximation of the function $n(x)$ by $m_N(x)$ in the quadrature formula, and the second

arising from quadrature truncation error in approximating the integral. Methods for bounding the latter source are found in the numerical analysis literature (e.g., Szidarovszky and Yakowitz, (1978), Chapter 3). For example, if D is an m -dimensional unit cube, and one applies a product trapezoidal quadrature rule (keeping in mind that the t_j 's in (3.14) can be chosen arbitrarily), one can verify that, provided $m(x)$ has continuous second derivatives,

$$| I - \int m(x) dx | = O(h^2),$$

h being the step size for the quadrature formula.

The variance of the first source is given by $\sum w_i w_j \text{Cov}(t_i, t_j)$, which can be approximated by

$$E[(I_N - I)^2] = \sum \sum w_i w_j \tilde{\text{Cov}}(t_i, t_j): 1 \leq i, j \leq M,$$

where the term $\tilde{\text{Cov}}$ is the covariance approximation given in (3.13). As we have noted, as $N \rightarrow \infty$, the covariance terms become negligible and useful approximation is that, in terms of (3.10),

$$E[(I_N - I)^2] = \sum_{i=1}^M w_i^2 \tilde{\sigma}^2(t_i). \quad (3.16)$$

The final consideration of this section concerns a certain optimality property of NPR convergence. In view of (3.6) and the Chebyshev inequality, one can conclude that for $r^* = 2/(m+4)$, and for any regression function $m(x)$ and noise process $n(x)$ satisfying the theorem hypothesis, if $a_N \rightarrow 0$ proportionally to $N^{-(m+4)^{-1}}$, then for $r = r^*$

$$\lim_{C \rightarrow \infty} \limsup_N [P[|m_N(x^*) - m(x^*)| > C N^r] \rightarrow 0. \quad (3.17)$$

Thus, in the terminology of Stone (1980), the NPR estimate achieves convergence rate r^* . But according to the Theorem of that work,

C-5

for any NPR estimator of a twice continuously differentiable regression function of m independent variables, $r^*=2/(m+4)$ is the optimal rate: there is no estimator for which (3.17) holds for some $r>r^*$.

4. A Comparison of Convergence Properties of Kriging and Nonparametric Regression

Assume that the intrinsic random function (IRF) hypothesis holds, and there is no drift ($J=1, \phi_1=1$). As mentioned at the close of Section 2, the nonparametric regression (NPR) approach is applicable, if the sample domain points $\{x_i\}_{i=1}^N$ are chosen randomly and if, $\{x_i\}_{i=1}^N$ with probability 1, the sample functions $\bar{f}(x)$ of the IRF are continuous at x^* , then $m_N(x)$ converges to $\bar{f}(x)$ in the mean. If the sample IRF's are twice-continuously differentiable, with probability 1, then Theorem 3.1 gives convergence rates.

Toward addressing parts b of Problems 1 and 2 of Section 1, we have provided error formulas (3.10) and (3.16) which are asymptotically accurate provided only that the sample functions are continuous at x^* . These convergence properties hold regardless of whether noise $n(x)$ in (1.1) is present. But all these statements have been predicated on the assumption that the $\{x_i\}_{i=1}^N$ values are actually a random sample. However, under fairly lenient assumptions, Schuster and Yakowitz (1980, Theorem 2) have shown in the univariate case that $m_N(x)$ converges uniformly in x to $\bar{f}(x)$ provided only that the x_i 's are dense. Undoubtedly such results can be extended to bear on kriging-type problems more forcefully, and citations of related results (especially concerning the Priestly-Chao estimator) are to be found in the above reference.

Now it is clear that if the variogram is known exactly, because the kriging estimator is the best unbiased linear estimator, then the expected square error of the kriging estimator $f_N(x)$ is no greater than that of the NPR estimator, which is also linear and unbiased. On the other hand, in the noisy case, it is not known at this point whether its asymptotic convergence rate is faster than the NPR rate given in Theorem 3.1. In summary, when the IRF hypothesis is true and the variogram is known to the statistician, the kriging estimate is better in the least squares sense than the NPR estimate, and its error estimators (2.10) and (2.13) are exact, whereas the NPR error estimators are only asymptotically accurate.

On the other hand, if the IRF hypothesis cannot be relied on, or even when it can, if the variogram is not known (even if it is known to be in one of the parametric families of Table 2.1), then nothing can be said about the convergence of either the kriging estimator or the error function, whereas NPR convergence conditions we have alluded to may well be satisfied.

5. Some Illustrative Computations

We hope to eventually publish results summarizing our extensive computational experimentation on kriging and alternative procedures. For now, we provide a brief illustration of the preceding material by reporting just a few computations. In this particular case study, the function $f(x)$ is chosen to exactly satisfy the kriging hypotheses:

It is a realization of the Gaussian process with mean 0 and variogram

$$\gamma(h) = C(1 - \exp(-25h)). \quad (5.1)$$

We have plotted $f(x)$ in Figure 1. The sample function $f(x)$ was simulated according to an algorithm described in Newman and Odell (1971) and is exact (within machine error) to the extent of one's being able to provide independent Gaussian observations. These we approximated by the Box-Muller algorithm (described in Yakowitz (1977)) using the CDC random number generator RANF.

In Table 2.2, we report the results of applying the kriging method with exponential variogram to 50 uniformly chosen domain points from the domain $X = [0,1]$. (RANF was used to obtain these points also.) In the first listing, we give the approximation at eight equidistant domain points, of the kriging algorithm in which the exponential parameter has been set to its correct value. This is, therefore, kriging under the ideal conditions of the variogram being known. In the second exponential listing, the variogram parameters a and w were obtained by least-squares fit according to current practice. In Table 2.3, we have repeated the calculations, but Gaussian noise $n(x_1)$, with standard deviation $\sigma = 0.5$, was added to each value $f(x_1)$. In Table 2.3, we have repeated the calculations, using exactly the same (x_1, y_1) sample values as in the construction for Table 2.2, but here we have assumed that the variogram is spherical (the parameters again being calibrated by a least squares procedure). Also, we have applied the same simulated data to the Watson nonparametric regression method.

One will notice that in all cases, the estimation capabilities exhibited by the various rules are quite comparable. Interestingly enough, the spherical variogram is also competitive, even though the model is wrong. But, especially in the noisy case, the spherical rule is much less accurate than the other rules in approximating the errors.

Deihomme (1979) has claimed that classical function interpolation and approximation methods are not effective with intrinsic random functions. Our experience with cubic splines, Lagrange interpolation, and least squares approximation concurs with this assessment. In Table 2.4, we present the estimates obtained from using the IMSL cubic spline package on the data points used for calculations in the preceding tables.

We applied the kriging integration algorithm to the function $f(x)$ which has served as basis for the calculations reported in the preceding tables. The same data pairs $\{(x_i, y_i)\}$ were used. The results of the integration estimation studies are summarized in Table 2.5 below.

DOMAIN POINT X^*	TRUE VALUE $f(X^*)$	PREDICTED VALUES $f_{50}(X^*)$	EXPECTED ERROR $E[(f(X^*) - f_{50}(X^*))^2]^{1/2}$
.111111	-.273470	-.422838	.191929
.222222	-.009991	.032374	.622137
.333333	.203601	.353256	.129905
.444444	.175779	.176909	.014628
.555556	-.621823	-.379835	.298123
.666667	.010971	-.069117	.224106
.777778	-.141964	-.096465	.120230
.888889	-.165034	-.165653	.007817

$\sigma = 0$, Exponential Covariogram, $\alpha = 1/25$

.111111	-.273470	-.413973	.233340
.222222	-.009991	.029094	.710130
.333333	.203601	.348190	.159101
.444444	.175779	.176738	.018131
.555556	-.621823	-.372394	.363309
.666667	.010971	-.057460	.273555
.777778	-.141964	-.096163	.148534
.888889	-.165034	-.165523	.009691

$\sigma = 0$, Exponential Covariogram, Calibrated α

.111111	-.273470	-.492881	.191929
.222222	-.009991	.119208	.622137
.333333	.203601	.177926	.129905
.444444	.175779	.174030	.014628
.555556	-.621823	-.309545	.298123
.666667	.010971	.311225	.224106
.777778	-.141964	.482942	.120230
.888889	-.165034	.195491	.007817

$\sigma = 0.5$, Exponential Covariogram, $\alpha = 1/25$

.111111	-.273470	-.309462	.638155
.222222	-.009991	.097315	1.014027
.333333	.203601	.629600	.478127
.444444	.175779	.092593	.066360
.555556	-.621823	-.200567	.835892
.666667	.010971	.216833	.722449
.777778	-.141964	.448980	.469842
.888889	-.165034	.197299	.035730

$\sigma = 0.5$, Exponential Covariogram, Calibrated α

TABLE 2.2
RANDOM FUNCTION ESTIMATES, I

DOMAIN POINT X*	TRUE VALUE f(X*)	PREDICTED VALUE f ₅₀ (X*)	EXPECTED ERROR E[(f(X*)-f ₅₀ (X*)) ²] ^{1/2}
.111111	-.273470	-.343883	.017824
.222222	-.009991	-.097980	.014013
.333333	.203601	.201102	.00511
.444444	.175779	.049139	.007597
.555556	-.621823	-.550244	.007565
.666667	.010971	-.087334	.027576
.777778	-.141964	-.088491	.006461
.888889	-.165034	-.120925	.027717

σ = 0, Spherical Variogram

.111111	-.273470	-.313063	.126485
.222222	-.009991	.028381	.107186
.333333	.203601	.177298	.140293
.444444	.175779	.072291	.091596
.555556	-.621823	-.516766	.110714
.666667	.010971	.100016	.133669
.777778	-.141964	.113751	.031628
.888889	-.165034	-.100722	.177456

σ = 0, Watson Algorithm

.111111	-.273470	.148942	.347720
.222222	-.009991	.214582	.161325
.333333	.203601	.427310	.091228
.444444	.175779	.234166	.035164
.555556	-.621823	-.620842	.086362
.666667	.010971	.354670	.101062
.777778	-.141964	-.202988	.079503
.888889	-.165034	-.331417	.026907

σ = 0.5, Spherical Variogram

.111111	-.273470	.004953	.144841
.222222	-.009991	-.270671	.162684
.333333	.203601	-.044319	.122492
.444444	.175779	.049391	.147927
.555556	-.621823	-.096836	.200769
.666667	.010971	-.102636	.252475
.777778	-.141964	.173377	.197700
.888889	-.165034	.121846	.200929

σ = 0.5, Watson Algorithm

TABLE 2.3
RANDOM FUNCTION ESTIMATES, II

OF FOUR QUANTILES

DOMAIN POINT X*	TRUE VALUE f(X*)	SPLINE VALUE
.111111	-.273470	-.537764
.222222	-.009991	.244992
.333333	.203601	.321591
.444444	.175779	.177214
.555555	-.621823	-.339041
.666667	.010971	-.075058
.777778	-.141964	-.156291
.888889	-.165034	-.171764

$\sigma = 0$

.111111	-.273470	1.226453
.222222	-.009991	-.976504
.333333	.203601	.875519
.444444	.175779	-.082533
.555555	-.621823	-.806521
.666667	.010971	.653228
.777778	-.141964	.571922
.888889	-.165034	.176924

$\sigma = 0.5$

TABLE 2.4
SPLINE ESTIMATES

Exact Value of $\int f(x) dx = -0.089$

	No Noise		Noise, $\sigma = 0.5$	
	Approximation	Estimated Standard Deviation of Error	Approximation	Estimated Standard Deviation of Error
Exponential Variogram, with $a = 1/25$	-0.051	0.93	0.126	0.927
Exponential Variogram, fitted parameter	-0.053	0.94	0.088	0.94
Spherical Variogram	-0.043	0.13	0.079	0.34
Watson Method	-0.056	0.19	0.069	0.21

TABLE 2.5
Integration Estimation Experiments

OF POOR QUALITY

5. Acknowledgements

This work is the product of evolution and labor over several years. Many kriging partisans, most notably, G. De Marsilly, J.P. Delhomme, G. Gambolati, and S. Neuman have been kind enough to patiently explain their viewpoints on kriging in conversations with the first author. Also, the first author is grateful for fruitful discussions about kriging with P. K. Bhattacharya, J. L. Denny, and E. Schuster.

This collaborative research was made possible by the NSF cooperative grant (with the Hungarian Mining Authority) ENG Int. 78-12184, and additionally the first author received support for this work from NSF grants ENG 76-20280, 78-07358, and CME 7905010.

ORIGINAL PAGE IS
OF POOR QUALITY

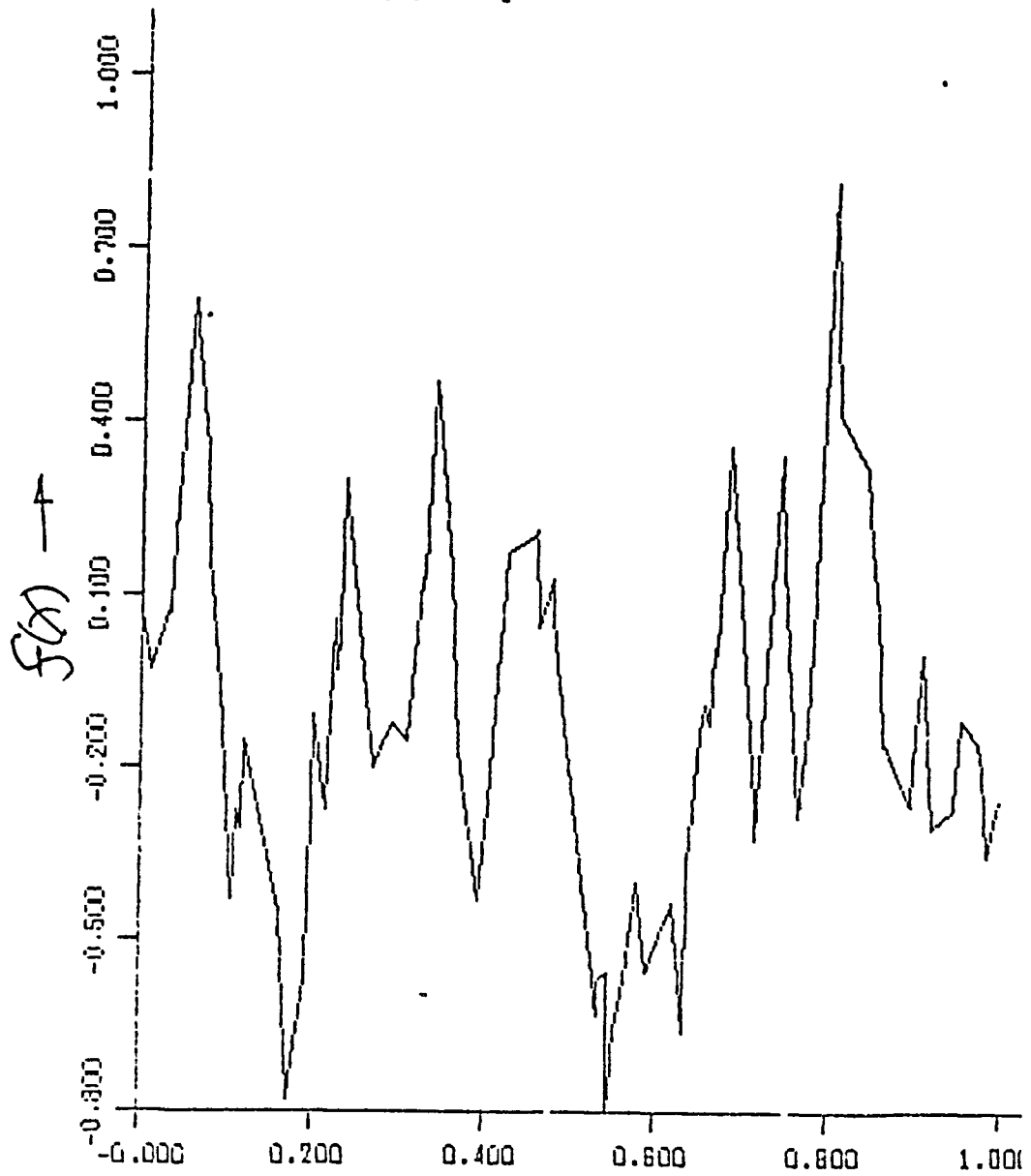


Figure 1

Graph of Target Function

X →

REFERENCES

- Bakr, A. A., L. W. Gelhar, A. L. Gutjahr and J. R. MacMilliam (1978).
Stochastic Analysis of Spatial Variability in Subsurface Flows
1. Comparison of One- and Three-Dimensional Flows. Water
Resources Research. 14(2):263-271.
- Chirlin, G. R. and G. Dagan (1980), Theoretical Head Variograms for
Steady Flow in Statistical Homogeneous Aquifers, Water Resour.
Res., 16(6), 1001-10015.
- David, M., Geostatistical Ore Reserve Estimation, Elsevier, New
York, 1977.
- Delhomme, J. (1979), Spatial Variability and Uncertainty in Ground-
water Flow Parameters: a Geostatistical Approach, Water Resour.
Research 15(2), 269-280.
- Delhomme, J. P. (1978), Kriging in the Hydrosiences, Advances in
Water Resources, 1(5), 251-266.
- Dendrou, B. A. and En N. Houstis (1978), An Inference-Finite
Element Model for Field Problems, Appl. Math. Modelling, 2,
109-114.
- Fisher, L. and Yakowitz, S. (1976), Uniform Convergence of the
Potential Function Algorithm. SIAM J. Control 14 95-103.
- Gambolati G. and G. Volpi (1979), A Conceptual Deterministic Analysis
of the Kriging Technique in Hydrology, Water Res. Research, 15(3),
625-629.
- Gambolati, G. and V. Giampiero (1979). Groundwater Contour Mapping
in Venice by Stochastic Interpolators 1. Theory. Water Resources
Research. 15(2):281-290.
- Gelhar, L. W., A. L. Gutjahr and R. L. Naff (1979). Stochastic
Analysis of Macro-Dispersion in a Stratified Aquifer. Water
Resources Research. 15(6):1387-1397.
- Giampiero, V. and G. Gambolati (1979). Groundwater Contour Mapping
in Venice by Stochastic Interpolators 2. Results. Water Resources
Research. 15(2):291-297.
- Gutjahr, A. L., L. W. Gelhar, A. A. Bahr and J. R. MacMillan (1978).
Stochastic Analysis of Spatial Variability in Subsurface Flows
2. Evaluation and Application. Water Resources Research. 14(5):
953-959.
- Huijbregts, C. J. (1975). Regionalized Variables and Quantitative
Analysis of Spatial Data. Eds. J. C. Davis, and M. J. McCullagh.
John Wiley and Sons, Inc., New York.

- Journel, A. G. and Ch. J. Huljbrechts (1978), Mining Geostatistics, Academic Press, N.Y.
- Journel, A. (1977). Kriging in Terms of Projections, J. Math. Geol., 9(6), 563-586.
- Journel, A. (1974), Geostatistics for Conditional Simulations of Ore Bodies, Econ. Geo., 69(5), 673-687.
- Krige, D. G., A Statistical Approach to some Mine Valuations and Allied Problems on the Witwaterstrand, Unpublished Master's Thesis, University of Witwaterstrand, South Africa, 1951.
- Krige, D. K. (1966). Two-Dimensional Weighted Moving Average Trend Surfaces for Ore Valuation. Journal of the South African Institute of Mining and Metallurgy. pp. 13-79.
- Matheron, G. (1973), The Intrinsic Random Functions and their Applications, Adv. Appl. Prob., 5, 439-468.
- Matheron, G. (1971), The Theory of Regionalized Variables and its Applications. Les Cahiers du CMM Fasc. no. 5, ENSMP, Paris, 211 p.
- Matheron, G. (1963). Principles of Geostatistics. Economic Geology. 58:1246-1266.
- Newman, T. and P. Odell, (1971), The Generation of Random Variates, Griffin, London.
- Olea, R. A. (1974). Optimal Contour Mapping Using Universal Kriging. Journal of Geophysical Research. 79(5):695-702.
- Parthasarathy, K. R., and P. K. Bhattacharya (1961), Some Limit Theorems in Regression Theory, Sankhyā, Series A. 23, 91-102.
- Rendu, J. (1980), Disjunctive Kriging: Comparison of Theory with Actual Results, Mathematical Geology, 12(4), 305-320.
- Sacks, J. and C. Spiegelman (1980), Consistent Window Estimation in Nonparametric Regression, Ann. Math. Statist., 9(2), 240-246.
- Schuster, E. F. (1972), Joint Asymptotic Distribution of the Estimated Regression Function at a Finite Number of Distinct Points, Ann. Math. Statist., 43(1), 84-88.
- Schuster, E. F. and S. Yakowitz (1979), Contributions to the Theory of Nonparametric Regression, with Applications to System Identification, Ann. Statist., 7(1), 139-149.
- Stone, C. (1980). Optimal Rates of Convergence for Nonparametric Estimators, Ann. Statist. 8 (6), 1348-1360.
- Stone, C. J. (1977). Consistent Nonparametric Regression. Ann Statist., 5 595-620.
- Szidarovszky, F., and S. Yakowitz (1978), Principles and Procedures of Numerical Analysis, Plenum Press, New York.

- Szidarovszky, F., and S. Yakowitz (1981), Some Mathematical Properties of the Kriging Method (in Hungarian), accepted for publ. in Banvaszati Lapok (Mining Journal), Budapest, Hungary.
- Villeneuve, J. P., G. Morin, B. Bobee, D. Lebanc, and J. P. Delhomme, (1979) Kriging in the Design of Streamflow Sampling Networks, Water Resour. Res., 15(6), 1833-1840.
- Watson, G. (1977), Review of Advanced Geostatistics in the Mining Industry, J. American Statistical Assoc., 72, 687-688.
- Watson, G. S. (1964). Smooth Regression Analysis. Sankhyā Ser. A 26 359-372.
- Yakowitz, S., J. Krimmel, and F. Szidarovszky (1978), Weighted Monte Carlo Integration, SIAM J. on Numerical Analysis, 15(6), 1289-1300.
- Yakowitz, S. (1977), Computational Probability and Simulation, Addison Wesley, Reading, Mass.

N83

15788

UNCLAS

DEPARTMENT OF STATISTICS

University of Wisconsin
1210 W. Dayton St.
Madison, WI 53706

March 1982

ESTIMATION OF DIVERGENCE AND VORTICITY
USING MULTIDIMENSIONAL SMOOTHING SPLINES

James G. Wendelberger

University of Wisconsin-Madison

This manuscript was prepared in conjunction with an invited talk for the NASA workshop on "Density Estimation and Function Smoothing" held at the Texas A & M University March 11-13, 1982. This research was supported by NASA under Grant No. NAG5-128 and by the Office of Naval Research under Contract No. N00014-77-C-0675.

ABSTRACT

Laplacian smoothing splines, smoothing splines on the sphere and smoothing pseudo splines on the sphere are presented. The method of generalized cross validation to choose the smoothing parameter is described. An application of these methods to estimate divergence and vorticity of the atmosphere from wind speed and wind direction is provided.

1. Introduction

This report portrays the status of current research into a meteorological application which involves the use of multidimensional smoothing splines. Aspects of meteorology, theoretical and applied mathematics, statistics, numerical analysis and computer science are involved in the analysis. A more detailed dissertation involving this problem is provided in Wahba and Wendelberger (1980), Wendelberger (1981) and Wendelberger (1982). The relevance to problems encountered in remote sensing are mentioned in a very general way throughout. Research topics involving the application of multidimensional smoothing splines are provided.

To review the work being done in this area there are sections about Laplacian smoothing splines, smoothing pseudo splines on the sphere and the method of generalized cross validation. These three sections are followed by one which involves the analysis of meteorological data which is of the type which may be encountered in the application of remote sensing. The last section proposes some future research areas.

2. The Laplacian Smoothing Spline

A Laplacian smoothing spline (LSS) is a function defined from Euclidean d -space, R^d , to R which arises as the estimate from the statistical model presented below. The term model is meant in the broad sense of Box (1981). In that sense we tentatively entertain the assumptions, provide a solution using the data and then check the validity of the assumptions. In this section we present the assumptions which this model entertains and provide a solution using the data. The question of model validity will not be dealt with here.

In the model, the data $z_i \in R$, $i=1, \dots, N$ consist of a fixed component and a random component. The fixed (or signal) component, Lif , in its most general

form, is a continuous linear functional L_i , $i=1, \dots, N$, of a function $f \in X$, X the appropriate Sobolev space, Adams (1975), to R . The random (or noise) component $e_i \in R$ satisfies

$$E e_i = 0, \quad i=1, \dots, N, \quad 2.1$$

$$E e_i^2 = \text{Var } e_i = \sigma_i^2 \sigma^2, \quad i=1, \dots, N, \quad 2.2$$

for σ^2 , σ_i^2 , constants with σ^2 unknown, σ_i^2 known and the

$$e_i, \quad i=1, \dots, N \text{ are independent.} \quad 2.3$$

In 2.1 and 2.2 E means mathematical expectation with respect to the error distribution of e_i . In 2.2 the σ_i are known weights which should be thought of as relative measurement error variances. The fixed and random components are additive;

$$z_i = L_i f + e_i, \quad i=1, \dots, N. \quad 2.4$$

We concern ourselves here with the evaluation functionals, $L_i f = f(t_i)$, where $t_i \in R^d$, $i=1, \dots, N$ and the t_i are considered to be known without error. Then 2.4 becomes

$$z_i = f(t_i) + e_i, \quad i=1, \dots, N. \quad 2.5$$

Applications of remote sensing may involve continuous linear functionals other than the evaluation functionals. For a further discussion of the use of general continuous linear functionals see Wahba and Wendelberger (1980).

To recover an estimate of $f \in X$, say g , from the observations $z = (z_1, \dots, z_N)^T$ we require that f be smooth. By smooth it is meant that $J_m(f)$ is small where

$$J_m(f) = \sum_{v=1}^{M'} \frac{m!}{\alpha_{1,v}! \dots \alpha_{d,v}!} \int_{R^d} \left[\frac{\partial^m f(t)}{\partial t_1^{\alpha_{1,v}} \dots \partial t_d^{\alpha_{d,v}}} \right]^2 dt, \quad 2.6$$

for $M' = \binom{m+d-1}{d-1}$, $t=(t_1, \dots, t_d)^T$ and the $\alpha_{1,v}, \dots, \alpha_{d,v}$ are the M' unique

combinations of $\{0, 1, \dots, m\}$ such that $\alpha_{1,v} + \dots + \alpha_{d,v} = m$. The smoothness function is induced by the Sobolev space X of which f is a member.

Besides being smooth f should also be close to the data. To measure closeness define

$$C(f) = \sum_{i=1}^N ([f(t_i) - z_i]/\sigma_i)^2. \quad 2.7$$

As defined here closeness and smoothness are conflicting criteria. To measure the tradeoff between the two we introduce the parameter λ . The choice of λ will be discussed in section 4. The estimate g of f is chosen as the minimizer of

$$C(f) + \lambda J_m(f). \quad 2.8$$

The minimizer of 2.8 can be shown to be of the form

$$g(t) = \sum_{i=1}^N c_i n_{J_m}(t, t_i) + \sum_{v=1}^M d_v \phi_v(t) \quad 2.9$$

where

$\phi_v(t)$ = the M polynomials of total degree less than m which span P_d^{m-1} , 2.10

$$M = \binom{d+m-1}{d}, \quad 2.11$$

P_d^{m-1} = the space of all polynomials of total degree less than m ,

η_{J_m} = a function of $|t-t_i|$ which depends on J_m and is rigorously defined in Wahba and Wendelberger (1980),

$c = (c_1, \dots, c_N)^T$, $d = (d_1, \dots, d_M)^T$ are constants which arise as the solution to the linear system

$$(K + \lambda \lambda D_\sigma^2) c + Td = Z \quad 2.12$$

and

$$T^T c = 0 \quad 2.13$$

where $D_\sigma^2 = \text{diag}(\sigma_1^2, \dots, \sigma_N^2)$ and the N by N matrix K and N by M matrix T depend only on t_i , $i=1, \dots, N$ and $J_m(\cdot)$.

The estimate g along with the assumptions stated in this section and those made in section 4 involving the choice of λ constitute the Laplacian smoothing spline model.

The Laplacian smoothing spline is given by g in 2.9. In remote sensing applications which involve a small section of a sphere (the earth), the Laplacian smoothing spline is appropriate. However, for applications which

5. Acknowledgements

This work is the product of evolution and labor over several years. Many kriging partisans, most notably, G. De Marsilly, J.P. Delhomme, G. Gambolati, and S. Neuman have been kind enough to patiently explain their viewpoints on kriging in conversations with the first author. Also, the first author is grateful for fruitful discussions about kriging with P. K. Bhattacharya, J. L. Denny, and E. Schuster.

This collaborative research was made possible by the NSF cooperative grant (with the Hungarian Mining Authority) ENG Int. 78-12184, and additionally the first author received support for this work from NSF grants ENG 76-20280, 78-07358, and CME 7905010.

ORIGINAL PAGE IS
OF POOR QUALITY

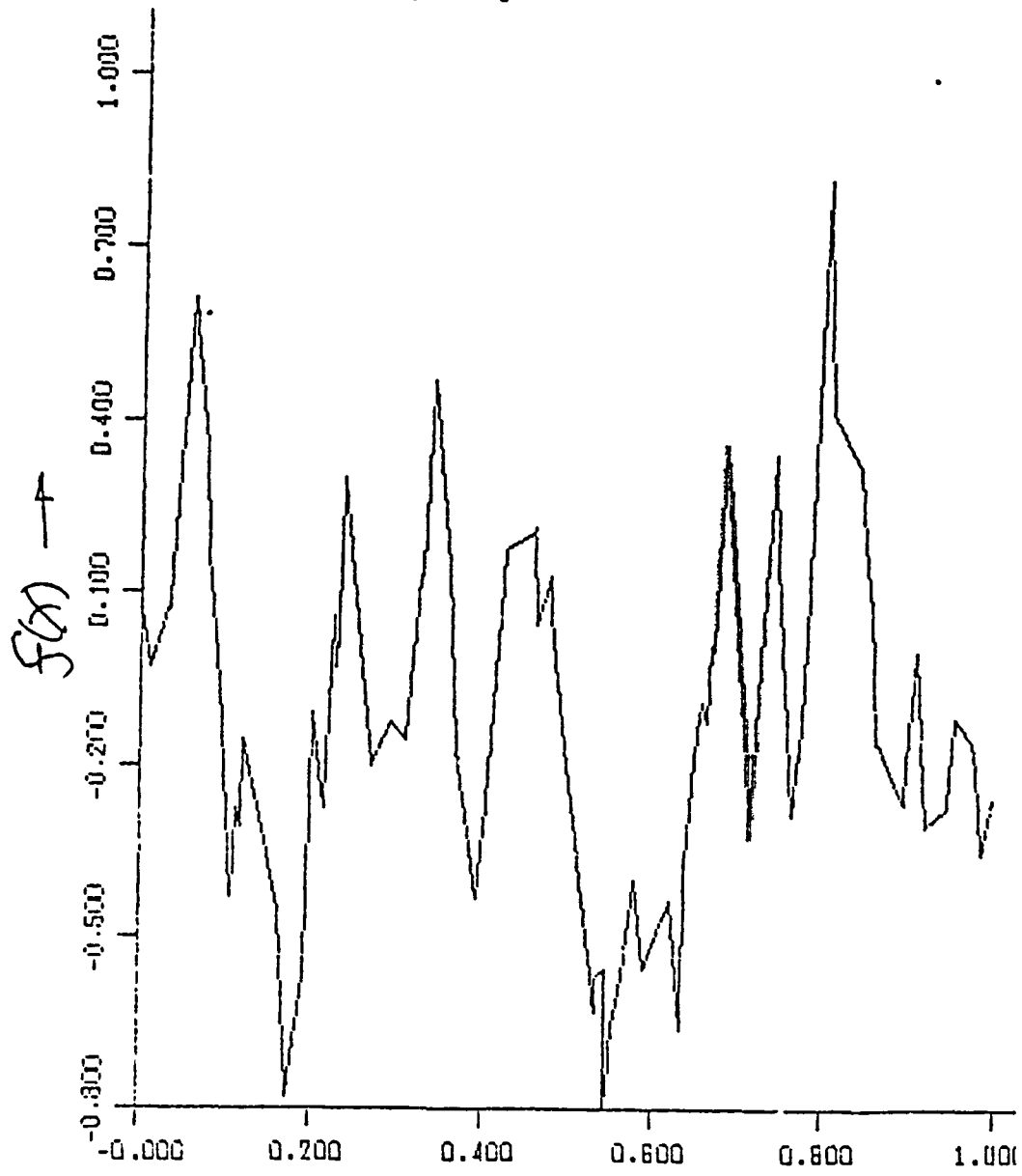


Figure 1
Graph of Target Function

$x \rightarrow$

REFERENCES

- Bakr, A. A., L. W. Gelhar, A. L. Gutjahr and J. R. MacMilliam (1978). Stochastic Analysis of Spatial Variability in Subsurface Flows 1. Comparison of One- and Three-Dimensional Flows. Water Resources Research. 14(2):263-271.
- Chirlin, G. R. and G. Dagan (1980), Theoretical Head Variograms for Steady Flow in Statistical Homogeneous Aquifers, Water Resour. Res., 16(6), 1001-10015.
- David, M., Geostatistical Ore Reserve Estimation, Elsevier, New York, 1977.
- Delhomme, J. (1979), Spatial Variability and Uncertainty in Groundwater Flow Parameters: a Geostatistical Approach, Water Resour. Research 15(2), 269-280.
- Delhomme, J. P. (1978), Kriging in the Hydrosciences, Advances in Water Resources, 1(5), 251-266.
- Dendrou, B. A. and En N. Houstis (1978), An Inference-Finite Element Model for Field Problems, Appl. Math. Modelling, 2, 109-114.
- Fisher, L. and Yakowitz, S. (1976), Uniform Convergence of the Potential Function Algorithm. SIAM J. Control 14 95-103.
- Gambolati G. and G. Volpi (1979), A Conceptual Deterministic Analysis of the Kriging Technique in Hydrology, Water Res. Research, 15(3), 625-629.
- Gambolati, G. and V. Giampiero (1979). Groundwater Contour Mapping in Venice by Stochastic Interpolators 1. Theory. Water Resources Research. 15(2):281-290.
- Gelhar, L. W., A. L. Gutjahr and R. L. Naiff (1979). Stochastic Analysis of Macro-Dispersion in a Stratified Aquifer. Water Resources Research. 15(6):1387-1397.
- Giampiero, V. and G. Gambolati (1979). Groundwater Contour Mapping in Venice by Stochastic Interpolators 2. Results. Water Resources Research. 15(2):291-297.
- Gutjahr, A. L., L. W. Gelhar, A. A. Bahr and J. R. MacMillan (1978). Stochastic Analysis of Spatial Variability in Subsurface Flows 2. Evaluation and Application. Water Resources Research. 14(5): 953-959.
- Huijbregts, C. J. (1975). Regionalized Variables and Quantitative Analysis of Spatial Data. Eds. J. C. Davis, and M. J. McCullagh. John Wiley and Sons, Inc., New York.

- Journel, A. G. and Ch. J. Huijbregts (1978), Mining Geostatistics, Academic Press, N.Y.
- Journel, A. (1977). Kriging in Terms of Projections, J. Math. Geol., 9(6), 563-586.
- Journel, A. (1974), Geostatistics for Conditional Simulations of Ore Bodies, Econ. Geo., 69(5), 673-687.
- Krige, D. G., A Statistical Approach to some Mine Valuations and Allied Problems on the Witwaterstrand, Unpublished Master's Thesis, University of Witwaterstrand, South Africa, 1951.
- Krige, D. K. (1966). Two-Dimensional Weighted Moving Average Trend Surfaces for Ore Valuation. Journal of the South African Institute of Mining and Metallurgy. pp. 13-79.
- Matheron, G. (1973), The Intrinsic Random Functions and their Applications, Adv. Appl. Prob., 5, 439-468.
- Matheron, G. (1971), The Theory of Regionalized Variables and its Applications. Les Cahiers du CMM. Fasc. no. 5, ENSMP, Paris, 211 p.
- Matheron, G. (1963). Principles of Geostatistics. Economic Geology. 58:1246-1266.
- Newman, T. and P. Odell, (1971), The Generation of Random Variates, Griffin, London.
- Olea, R. A. (1974). Optimal Contour Mapping Using Universal Kriging. Journal of Geophysical Research. 79(5):695-702.
- Parthasarthy, K. R., and P. K. Bhattacharya (1961), Some Limit Theorems in Regression Theory, Sankhya, Series A. 23, 91-102.
- Rendu, J. (1980), Disjunctive Kriging: Comparison of Theory with Actual Results, Mathematical Geology, 12(4), 305-320.
- Sacks, J. and C. Spiegelman (1980), Consistent Window Estimation in Nonparametric Regression, Ann. Math. Statist., 9(2), 240-246.
- Schuster, E. F. (1972), Joint Asymptotic Distribution of the Estimated Regression Function at a Finite Number of Distinct Points, Ann. Math. Statist., 43(1), 84-88.
- Schuster, E. F. and S. Yakowitz (1979), Contributions to the Theory of Nonparametric Regression, with Applications to System Identification, Ann. Statist., 7(1), 139-149.
- Stone, C. (1980). Optimal Rates of Convergence for Nonparametric Estimators, Ann. Statist. 8 (6), 1348-1360.
- Stone, C. J. (1977). Consistent Nonparametric Regression. Ann Statist., 5 595-620.
- Szidarovszky, F., and S. Yakowitz (1978), Principles and Procedures of Numerical Analysis, Plenum Press, New York.

- Szidarovszky, F., and S. Yakowitz (1981), Some Mathematical Properties of the Kriging Method (in Hungarian), accepted for publ. in Bányászati Lapok (Mining Journal), Budapest, Hungary.
- Villeneuve, J. P., G. Morin, B. Bobee, D. Lebanc, and J. P. Delhomme, (1979) Kriging in the Design of Streamflow Sampling Networks, Water Resour. Res., 15(6), 1833-1840.
- Watson, G. (1977), Review of Advanced Geostatistics in the Mining Industry, J. American Statistical Assoc., 72, 687-688.
- Watson, G. S. (1964). Smooth Regression Analysis. Sankhyā Ser. A 26 359-372.
- Yakowitz, S., J. Krimmel, and F. Szidarovszky (1978), Weighted Monte Carlo Integration, SIAM J. on Numerical Analysis, 15(6), 1289-1300.
- Yakowitz, S. (1977), Computational Probability and Simulation, Addison Wesley, Reading, Mass.

N83

157888

UNCLAS

N83 15788⁸⁶ D14

DEPARTMENT OF STATISTICS

University of Wisconsin
1210 W. Dayton St.
Madison, WI 53706

March 1982

ESTIMATION OF DIVERGENCE AND VORTICITY
USING MULTIDIMENSIONAL SMOOTHING SPLINES

James G. Wendelberger

University of Wisconsin-Madison

This manuscript was prepared in conjunction with an invited talk for the NASA workshop on "Density Estimation and Function Smoothing" held at the Texas A & M University March 11-13, 1982. This research was supported by NASA under Grant No. NAG5-128 and by the Office of Naval Research under Contract No. N00014-77-C-0675.

ABSTRACT

Laplacian smoothing splines, smoothing splines on the sphere and smoothing pseudo splines on the sphere are presented. The method of generalized cross validation to choose the smoothing parameter is described. An application of these methods to estimate divergence and vorticity of the atmosphere from wind speed and wind direction is provided.

1. Introduction

This report portrays the status of current research into a meteorological application which involves the use of multidimensional smoothing splines. Aspects of meteorology, theoretical and applied mathematics, statistics, numerical analysis and computer science are involved in the analysis. A more detailed dissertation involving this problem is provided in Wahba and Wendelberger (1980), Wendelberger (1981) and Wendelberger (1982). The relevance to problems encountered in remote sensing are mentioned in a very general way throughout. Research topics involving the application of multidimensional smoothing splines are provided.

To review the work being done in this area there are sections about Laplacian smoothing splines, smoothing pseudo splines on the sphere and the method of generalized cross validation. These three sections are followed by one which involves the analysis of meteorological data which is of the type which may be encountered in the application of remote sensing. The last section proposes some future research areas.

2. The Laplacian Smoothing Spline

A Laplacian smoothing spline (LSS) is a function defined from Euclidean d -space, R^d , to R which arises as the estimate from the statistical model presented below. The term model is meant in the broad sense of Box (1981). In that sense we tentatively entertain the assumptions, provide a solution using the data and then check the validity of the assumptions. In this section we present the assumptions which this model entertains and provide a solution using the data. The question of model validity will not be dealt with here.

In the model, the data $z_i \in R$, $i=1, \dots, N$ consist of a fixed component and a random component. The fixed (or signal) component, Lif , in its most general

form, is a continuous linear functional L_i , $i=1, \dots, N$, of a function $f \in X$, X the appropriate Sobolev space, Adams (1975), to R . The random (or noise) component $e_i \in R$ satisfies

$$E e_i = 0, \quad i=1, \dots, N, \quad 2.1$$

$$E e_i^2 = \text{Var } e_i = \sigma_i^2, \quad i=1, \dots, N, \quad 2.2$$

for σ_i^2 , constants with σ_i^2 unknown, σ_i^2 known and the

$$e_i, \quad i=1, \dots, N \text{ are independent.} \quad 2.3$$

In 2.1 and 2.2 E means mathematical expectation with respect to the error distribution of e_i . In 2.2 the σ_i are known weights which should be thought of as relative measurement error variances. The fixed and random components are additive;

$$z_i = L_i f + e_i, \quad i=1, \dots, N. \quad 2.4$$

We concern ourselves here with the evaluation functionals, $L_i f = f(t_i)$, where $t_i \in R^d$, $i=1, \dots, N$ and the t_i are considered to be known without error. Then 2.4 becomes

$$z_i = f(t_i) + e_i, \quad i=1, \dots, N. \quad 2.5$$

Applications of remote sensing may involve continuous linear functionals other than the evaluation functionals. For a further discussion of the use of general continuous linear functionals see Wahba and Wendelberger (1980).

To recover an estimate of $f \in X$, say g , from the observations $z = (z_1, \dots, z_N)^T$ we require that f be smooth. By smooth it is meant that $J_m(f)$ is small where

$$J_m(f) = \sum_{v=1}^{M'} \frac{m!}{\alpha_{1,v}! \dots \alpha_{d,v}!} \int_{R^d} \left[\frac{\partial^m f(t)}{\partial t_1^{\alpha_{1,v}} \dots \partial t_d^{\alpha_{d,v}}} \right]^2 dt, \quad 2.6$$

for $M' = \binom{m+d-1}{d-1}$, $t=(t_1, \dots, t_d)^T$ and the $\alpha_{1,v}, \dots, \alpha_{d,v}$ are the M' unique

combinations of $\{0, 1, \dots, m\}$ such that $\alpha_{1,v} + \dots + \alpha_{d,v} = m$. The smoothness function is induced by the Sobolev space X of which f is a member.

Besides being smooth f should also be close to the data. To measure closeness define

$$C(f) = \sum_{i=1}^N ([f(t_i) - z_i]/\sigma_i)^2. \quad 2.7$$

As defined here closeness and smoothness are conflicting criteria. To measure the tradeoff between the two we introduce the parameter λ . The choice of λ will be discussed in section 4. The estimate g of f is chosen as the minimizer of

$$C(f) + \lambda J_m(f). \quad 2.8$$

The minimizer of 2.8 can be shown to be of the form

$$g(t) = \sum_{i=1}^N c_{iN} \phi_{iN}(t, t_i) + \sum_{v=1}^M d_v \phi_v(t) \quad 2.9$$

where

$\phi_v(t)$ = the M polynomials of total degree less than m which span P_d^{m-1} , 2.10

$$M = \binom{d+m-1}{d}, \quad \text{2.11}$$

P_d^{m-1} = the space of all polynomials of total degree less than m ,

η_{J_m} = a function of $|t-t_i|$ which depends on J_m and is rigorously defined in Wahba and Wendelberger (1980),

$c = (c_1, \dots, c_N)^T$, $d = (d_1, \dots, d_M)^T$ are constants which arise as the solution to the linear system

$$(K + N\lambda D_\sigma^2) c + Td = Z \quad \text{2.12}$$

and

$$T^T c = 0 \quad \text{2.13}$$

where $D_\sigma^2 = \text{diag}(\sigma_1^2, \dots, \sigma_N^2)$ and the N by N matrix K and N by M matrix T depend only on t_i , $i=1, \dots, N$ and $J_m(\cdot)$.

The estimate g along with the assumptions stated in this section and those made in section 4 involving the choice of λ constitute the Laplacian smoothing spline model.

The Laplacian smoothing spline is given by g in 2.9. In remote sensing applications which involve a small section of a sphere (the earth), the Laplacian smoothing spline is appropriate. However, for applications which

involve a large area of a sphere we require splines which have the surface of the sphere, rather than R^d , as their domain. These splines are developed in the next section.

3. Smoothing Splines on the Sphere

Smoothing splines on the sphere, as investigated by Wahba (1981), are developed both as an extension of one dimensional periodic polynomial splines and as a restriction of three dimensional thin plate (Laplacian) smoothing splines to the surface of the sphere. The derivation of smoothing splines on the sphere parallels that of Laplacian smoothing splines. In this section we provide the modifications of section 2 which are required to obtain smoothing splines on the sphere.

The first modification is that the independent variable space, R^d , is replaced by the surface of the sphere, S . This means that the t_i in 2.5 become $t_i \in S$, $i = 1, \dots, N$. In particular $t_i = (\phi_i, \lambda_i)^T$, $\phi_i =$ latitude and $\lambda_i =$ longitude, $i = 1, \dots, N$.

The second modification is that in 2.11 $M = 1$. This means in 2.9 and 2.10 there is only one polynomial $\phi_1(t) = 1$. Intuitively, this arises because of the necessity of having a periodic solution.

The third modification is that $J_m(\cdot)$ in 2.6 is replaced by $K_m(\cdot)$. $K_m(\cdot)$ is a restriction of $J_m(\cdot)$ to S . For the specific form of $K_m(\cdot)$ see Wahba (1981).

The fourth modification is that $\eta_{j_m}^\lambda(t, t_i)$ both in 2.9 and in the definition of K in 2.12 is replaced by

$$\eta_{K_m}^\lambda(t, t_i) = \sum_{v=1}^{\infty} \frac{2v+1}{v^m(v+1)^m} P_v(\cos(A(t, t_i))), \quad 3.1$$

where $\cos(A(t,t_i))$ equals the cosine of the angle between t and t_i . Also, $P_\nu(\cdot)$ is the ν -th Legendre polynomial.

Modifications one through four provide the smoothing spline on the sphere which is given analogously to 2.9 as

$$h(t) = \sum_{i=1}^N c_i \eta_{K_m}(t, t_i) + d_1, \quad m = 2, 3, \dots \quad 3.2$$

To obtain and evaluate the smoothing spline on the sphere, 3.1 must be evaluated. Wahba (1981) notes that the series given in 3.1 cannot be expressed in terms of elementary functions. To compute smoothing splines on the sphere, the accurate and fast evaluation of 3.1 is necessary. To alleviate the difficulties which this entails, she derives smoothing pseudo splines on the sphere.

To obtain these splines $K_m(\cdot)$ is replaced by a topologically equivalent norm $L_m(\cdot)$. In both 3.1 and 3.2 $K_m(\cdot)$ is replaced by $L_m(\cdot)$, with specific expressions for $\eta_{L_m}(t, t_i)$ given in Wahba (1981). For illustrative purposes we provide for $m = 2$

$$\eta_{L_2}(t, t_i) = \ln(1 + (2/(1-z))^{1/2}) [3z^2 - 2z - 1]/2 - 6[(1-z)/2]^{3/2} + 2 - 3z/2, \quad 3.3$$

with $z = \cos(A(t, t_i))$. The smoothing pseudo spline on the sphere is thus easily computed by using expressions like 3.3 to obtain $\eta_{L_m}(t, t_i)$.

4. Generalized Cross Validation

In applications the smoothing parameter λ is unknown. To determine an estimate of this parameter, Craven and Wahba (1979) and Golub, Heath and Wahba (1979) have suggested the use of generalized cross validation. To enhance the understanding of this method a short synopsis of its development is given.

The method of cross validation (presented here as related to LSS's) is developed in response to the question: How well may one expect LSS's to predict the true functional value $f(t)$ at some point t ?

Simple cross validation (SCV) suggests predicting the true functional values of data different from that used in the analysis to assess this predictive ability. In SCV's simplest form this entails dividing the sample into two pieces of similar size, using one section for optimization and the other for testing. In addition, in order to gain more information from the data, the two pieces may be interchanged and the optimization and testing performed on each.

SCV is alright if there is an ample supply of data so that halving or doubling the data has little effect on the quality of the estimator. To lessen this effect Mosteller and Tukey (1968) propose single cross validation (1CV), (called ordinary cross validation by Wahba (1979)), which is described suitably by them as follows:

"Suppose that we set aside one individual case, optimize for what is left, then test on the set-aside case. Repeating this for every case squeezes the data almost dry. If we have to go through the full optimization calculation every time, the extra computation may be hard to face. Occasionally, one can easily calculate, either exactly or to an adequate approximation, what the effect of dropping a specific and very small part

of the data will be on the optimized result. This adjusted optimized result can then be compared with the values for the omitted individual. That is, we make one optimization for all the data, followed by one repetition per case of a much simpler calculation, a calculation of the effect of dropping each individual, followed by one test of that individual. When practical, this approach is attractive."

To describe 1CV mathematically we require some notation. Let $f_{\lambda}^{(j)}$ be the solution to the minimization of 2.8 with the j th point removed from the analysis. Similarly, $D_{\sigma}^{(j)}$ is the $N-1$ by $N-1$ matrix composed of D_{σ} with its j -th row and column removed. To "test on the set aside case" we require that $[(f_{\lambda}^{(j)}(t_j) - z_j)/\sigma_j]^2$ be small. "Repeating this for every case" and averaging to yield an overall test gives

$$V_m^0(\lambda) = (1/N) \sum_{j=1}^N [(f_{\lambda}^{(j)}(t_j) - z_j)/\sigma_j]^2. \quad 4.1$$

1CV uses the λ which minimizes $V_m^0(\lambda)$, Wahba and Wold (1975).

To minimize $V_m^0(\lambda)$ directly is not a trivial computational matter. For each proposed value of λ a system of the form 2.12 and 2.13 (of order $N+M-1$ instead of $N+M$) must be solved for each of the N values left out of the analysis. This entails solving a linear system of order $N+M-1$ N times! As noted earlier, "if we have to go through the full optimization calculation every time, the extra computation may be hard to face." Following the idea of Mosteller and Tukey we seek a computational simplification for the minimizer of $V_m^0(\lambda)$.

The simplified form for 1CV was first noted by Craven and Wahba (1979), Golub, Heath and Wahba (1979) and given in a slightly more general form in Wahba and Wendelberger (1980). The 1CV function may be written

ORIGINAL FACT OF
OF POOR QUALITY

$$V_m^0(\lambda) = (1/N) \sum_{j=1}^N [f_\lambda(t_j) - z_j / (\sigma_j(1 - a_{jj}(\lambda)))]^2. \quad 4.2$$

$a_{jj}(\lambda)$ is the j th diagonal element of $A_m(\lambda)$ which is defined by

$$A_m(\lambda)z = \begin{bmatrix} f_\lambda(t_1) \\ \vdots \\ f_\lambda(t_N) \end{bmatrix}$$

where g is the solution of 2.8. $A_m(\lambda)$ may be thought of as mapping the vector z into the smoothed values.

In this form "we make one optimization for all the data" by calculating g which is then "followed by one repetition per case of a much simpler calculation, a calculation of the effect of dropping each individual." Here find $a_{jj}(\lambda)$ and use (4.2).

Evaluation of this formulation of $V_m^0(\lambda)$ involves solving a linear system of size $N+M$ to find g and one of size N to find $a_{jj}(\lambda)$. This is a considerable improvement over using 4.1 directly. Because of a mathematical simplification the amount of computation needed to minimize $V_m^0(\lambda)$ can be substantially reduced. From a practical point of view this makes the use of cross validation very attractive.

When applying cross validation to problems other than LSS's, this last step of finding "what the effect of dropping a specific and very small part of the data will be on the optimized result" is very important and should not be overlooked. In fact, this step often makes cross validation computationally feasible, whereas, without this insight it may be impractical.

Finding the minimizer of $V_m^0(\lambda)$ requires evaluation of $V_m^0(\lambda)$ at different values of λ as determined by a search routine. Hence, although the minimization is possible, we need to repeatedly solve large linear systems with the number of solutions times being a function of the search routine employed.

In $V_m^0(\lambda)$ of 4.1 each deviation of $f_{\lambda}^{(1)}(t_i)$ from the observed value z_i is treated symmetrically. This choice is arbitrary and is chosen for simplicity. A more general approach is to weight each term of 4.1 or equivalently 4.2 to yield

$$V_m(f_{\lambda}) = (1/N) \sum_{i=1}^N W_i [(f_{\lambda}(t_i) - z_i) / (\sigma_i(1 - a_{ij}(\lambda)))]^2. \quad 4.3$$

Before discussing the choice of these weights, the following definition is needed.

Definition:

$$R_m(\lambda) = E(1/N) \sum_{i=1}^N [(f(t_i) - g(t_i)) / \sigma_i]^2$$

is the expected weighted (by σ_i) mean squared error between the true function (f) and the spline (g) evaluated at the independent variables (t_i). E denotes mathematical expectation with respect to the error distribution of the random errors as described in the model of Section 2.

If we want $R_m(\lambda)$ to be small, then the generalized cross validation value of λ should be used as the smoothing parameter value. Using ICV as motivation Craven and Wahba (1979) and Golub, Heath and Wahba (1979) have shown that the λ which minimizes $V_m(\lambda)$ with weights

$$W_i = (1 - a_{ij}(\lambda))^2 / (1 - N^{-1} \sum_{j=1}^N a_{jj}(\lambda))^2$$

is an estimate of the λ which minimizes $R_m(\lambda)$. Using these weights in 4.3 gives the generalised cross validation function (GCVF)

$$V_m(\lambda) = (1/N) \sum_{i=1}^N [(g(t_i) - z_i) / (\sigma_i (1 - N^{-1} \sum_{j=i}^N a_{jj}(\lambda)))]^2. \quad 4.4$$

The minimizer of 4.4 is called the GCV estimate of λ .

The GCVF can be rewritten as

$$V_m(\lambda) = (1/N) \|D\sigma^{-1}(I - A_m(\lambda))z\|^2 / ((1/N) \text{Tr}(I - A_m(\lambda)))^2, \quad 4.5$$

where Tr is the trace.

Wahba (1981) has proposed

$$\sigma_e^2 = \|D\sigma^{-1}(I - A_m(\lambda))z\|^2 / \text{Tr}(I - A_m(\lambda)) \quad 4.6$$

as an estimate of the error variance σ^2 . This leads us to consider $df_e = \text{Tr}(I - A_m(\lambda))$ as the equivalent degrees of freedom of error Wahba (1982). Using these notions we rewrite the GCVF as

$$V_m(\lambda) = N\sigma_e^2 / df_e. \quad 4.7$$

The method of GCV may be viewed as minimizing the estimated error variance per error equivalent degrees of freedom.

5. Estimation of Height, Wind, Divergence and Vorticity

In this section we provide a preliminary report of the analysis of some meteorological data. For a discussion of the analysis of Monte Carlo experiments using Laplacian smoothing splines see Wendelberger (1981) and Wahba and Wendelberger (1980). The data to be analysed are obtained from the irregularly spaced North American radiosonde network during the Ohio storm of 00 Z January 25, 1978.

The height, h_i , wind speed and wind direction are reported with measurement error at the 850, 700, 500, 400, 300, 250, 200, 150 and 100 mb pressure levels. To analyse the wind the u_i (east) component and the v_i (north) component are obtained from the wind speed and wind direction measurements. Using those stations and levels for which all three components, (h_i, u_i, v_i) are obtained yields 112, 117, 116, 116, 113, 114, 109, 108 and 93 observations, respectively, for each pressure level.

Using the Laplacian smoothing spline model the method of sections 2 and 4 with $m = 4$ provides three fitted surfaces h_p , u_p and v_p for each pressure level p . Figure 1 provides the height field, h_p , for $p = 850, 500$ and 200 mb. The synoptic patterns are in general agreement with the National Meteorological Center's analysis. Figure 2 gives the isotachs and streamlines for u_p and v_p with $p = 850, 500$ and 200 mb. The isotachs are levels of constant wind speed and the streamlines denote the wind direction.

The vorticity, V , and horizontal divergence, D , may be obtained from

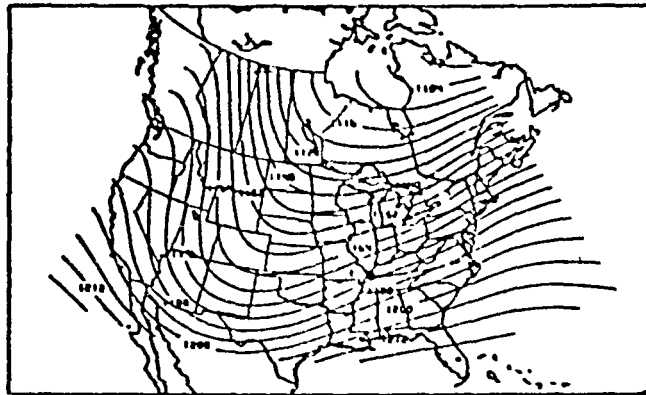
$$V = [(\partial v / \partial \lambda) / \cos \phi - \partial u / \partial \phi + u \cdot \tan \phi] / R \quad 5.1$$

and

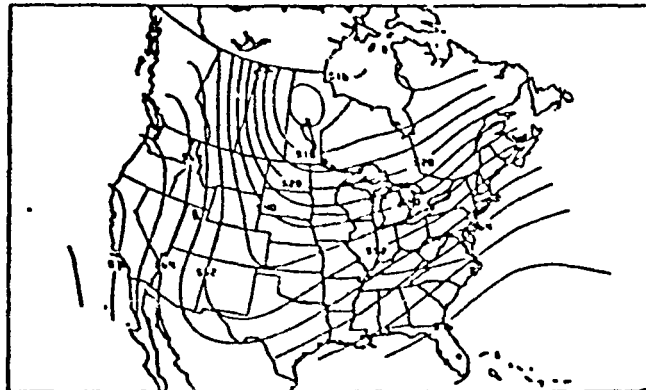
$$D = [(\partial u / \partial \lambda) / \cos \phi - \partial v / \partial \phi - v \cdot \tan \phi] / R \quad 5.2$$

where R is the radius of the earth, $\phi =$ latitude and $\lambda =$ longitude. Figure 3 is obtained from 5.1 and 5.2 using u_p and v_p for $p = 850, 500$ and 200 mb. The 500 mb vorticity pattern is in excellent agreement with the National Meteorological Center's analysis which is unavailable for comparison at the other levels.

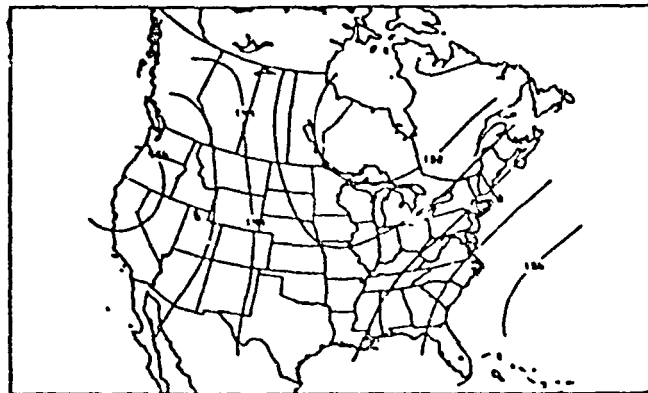
ORIGINAL PAGE IS
OF POOR QUALITY



(1a) 200 mb

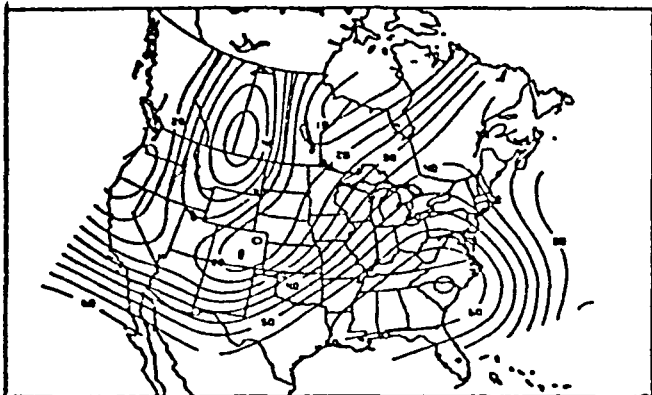


(1b) 500 mb

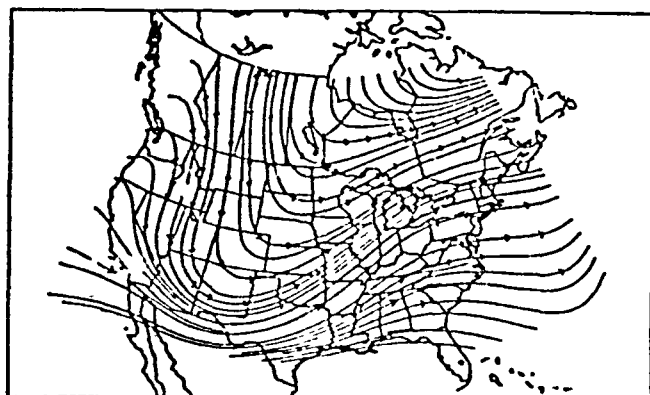


(1c) 850 mb

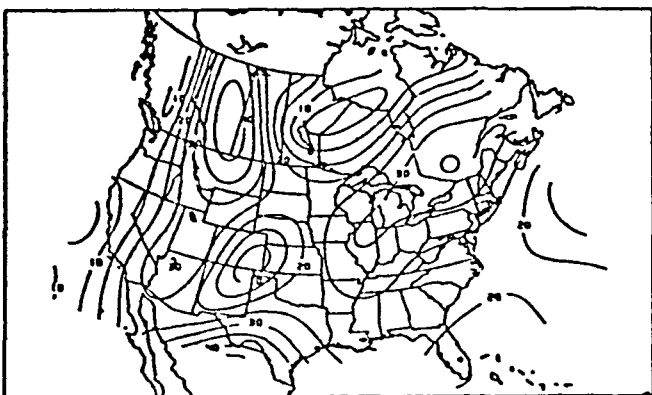
Figure 1: Height Fields, X 10 meters



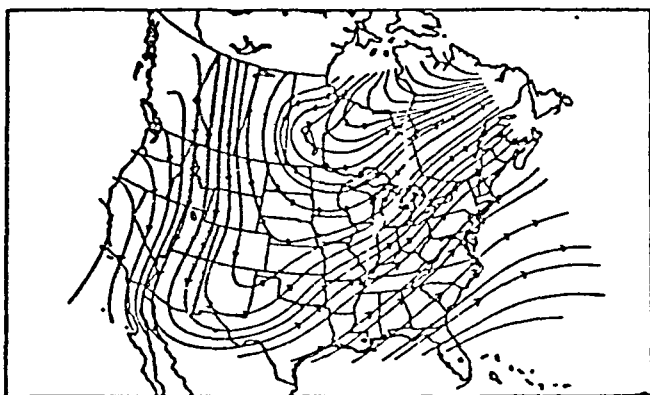
(2a) 200 mb Isotachs



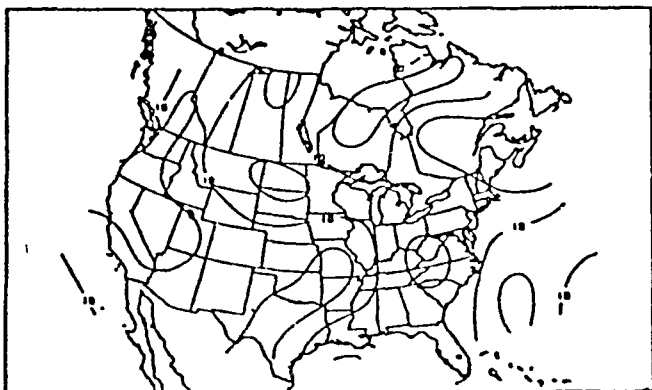
(2b) 200 mb Streamlines



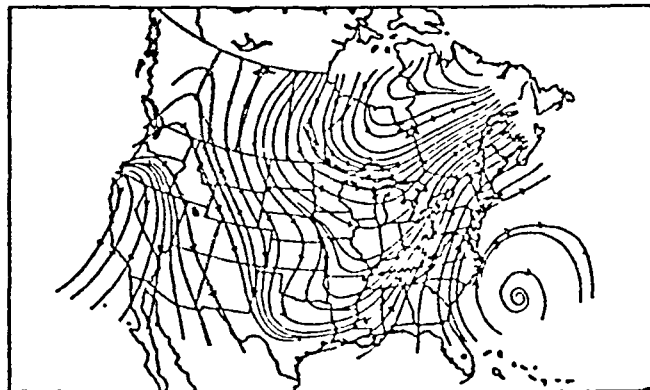
(2c) 500 mb Isotachs



(2d) 500 mb Streamlines

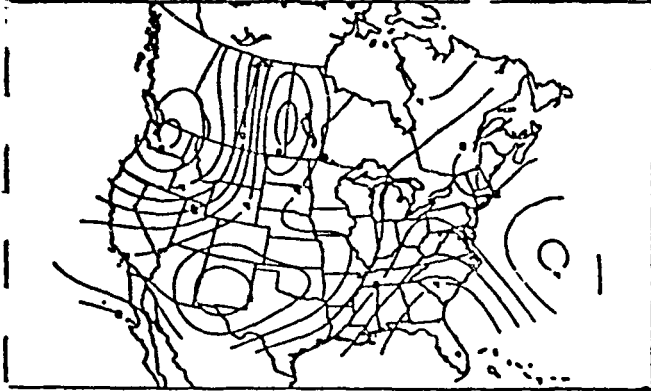


(2e) 850 mb Isotachs

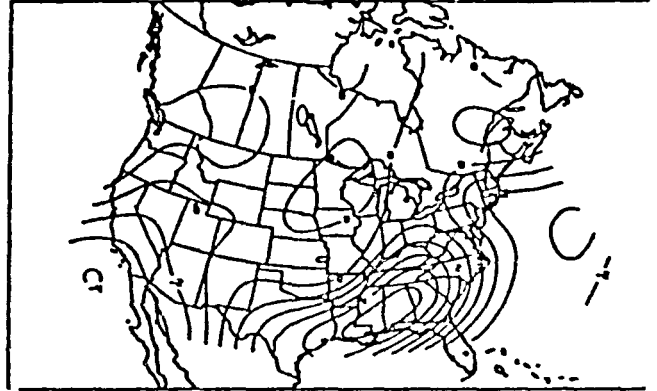


(2f) 850 mb Streamlines

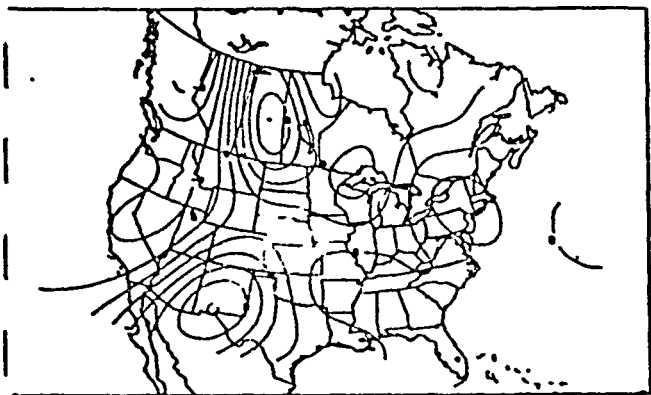
Figure 2: Isotachs (m/sec) and streamlines.



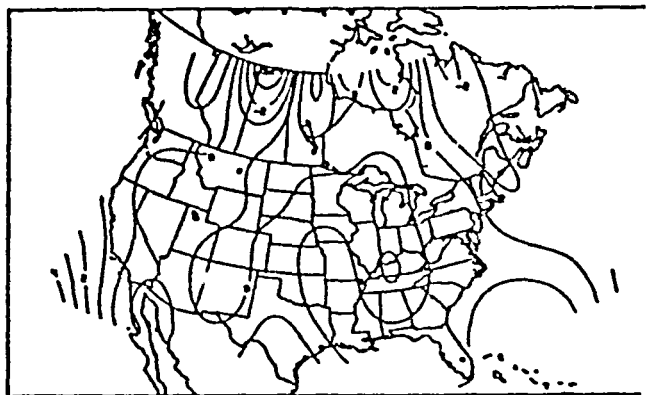
(3a) 200 mb Vorticity



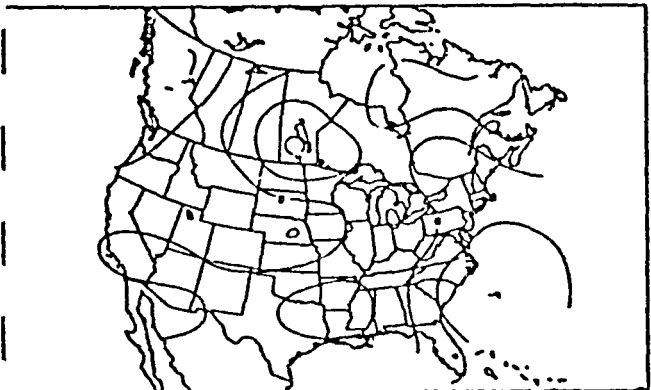
(3b) 200 mb Divergence



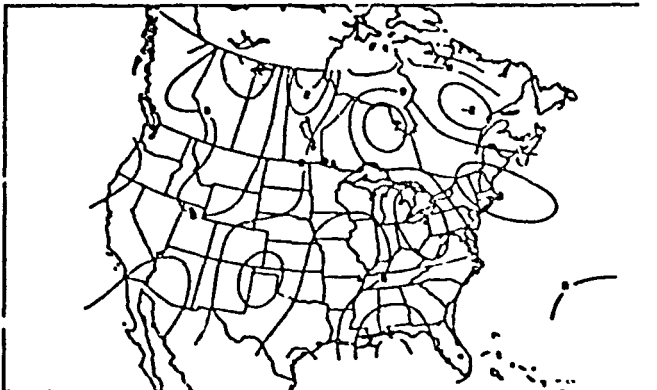
(3c) 500 mb Vorticity



(3d) 500 mb Divergence



(3e) 850 mb Vorticity



(3f) 850 mb Divergence

Figure 3: Vorticity and divergence, $\times 10^{-5}/\text{sec}$.

The results presented here will be supplemented with estimates of the accuracy of these fields by the method presented in Wahba (1981). The smoothing pseudo splines on the sphere will be employed to obtain h , u , v and the resulting divergence and vorticity estimates.

6. Further Research

In this section we list some further research ideas.

The current computational method used for Laplacian smoothing splines requires a spectral decomposition of an $N-M$ by $N-M$ matrix, Wendelberger (1981). It seems likely that the calculation of all $N-M$ of the eigenvalues and eigenvectors is unnecessary. An algorithm which determines how many eigenvalues are needed would be extremely useful; then a truncation algorithm could be obtained to compute the spline, Bates and Wahba (1982).

Often, given N observations for which the analysis has been performed, we may need to update or downdate this set of observations by the inclusion or exclusion of a single observation. An algorithm which does not require the spectral decomposition to be performed on the new $N-M+1$ by $N-M+1$ or $N-M-1$ by $N-M-1$ matrix would be very valuable. We could then generalize this to updating and downdating by a small number of points. The usefulness of this type of algorithm is very apparent in the example provided in section 5.

In remote sensing applications different continuous linear functionals L_j will be required. These need to be identified and their fast and accurate computational algorithms need to be designed. For a specific example see Nychka (1983).

In remote sensing applications, experiments need to be designed which will demonstrate the utility of smoothing splines. These will include Monte

Carlo runs with data similar to that obtained in practice and confidence statements about the estimates obtained.

Methods to check the validity of model assumptions must be devised. A probability plot of the residuals is one such method, see Wendelberger (1981) and Wendelberger (1982).

ACKNOWLEDGMENT

The work described here benefits greatly from close collaboration with Prof. Donald Johnson and Prof. Grace Wahba. However, the author accepts sole responsibility for the presentation given here.

This research is sponsored by NASA under Grant No. NAG5-128 and by the Office of Naval Research under Contract No. N00014-77-C-0675.

REFERENCES

- Adams, R.A., 1975: Sobolev Spaces. Academic Press. pp. 268.
- Bates, D. and G. Wahba, 1982: In Preparation.
- Box, G.E.P., 1980: Sampling and Bayes' Inference in Scientific Modelling and Robustness. Journal of the Royal Statistical Society, Series A, Vol. 143, Part 4, pp. 383-430.
- Craven, P. and G. Wahba, 1979: Smoothing Noisy Data with Spline Functions: Estimating the Correct Degree of Smoothing by the Method of Generalized Cross-Validation. Numer. Math., 31, 377-403.
- Golub, G., M. Heath, and G. Wahba, 1979: Generalized Cross-Validation as a Method for Choosing a Good Ridge Parameter. Technometrics 21, 2 pp. 215-223.
- Mosteller, F. and J. W. Tukey, 1968: Data Analysis Including Statistics. Handbook of Social Psychology, Vol. 2, Reading, MA, Addison-Wesley, 80-203.
- Nychka, D. 1983: Thesis to Appear. Dept. of Statistics, University of Wisconsin.
- Wahba, G., 1979: How to Smooth Curves and Surfaces with Splines and Cross-Validation. Tech Report #555, Dept. of Statistics, University of Wisconsin.
- Wahba, G. 1981: Spline Interpolation and Smoothing on the Sphere. SIAM J. Sci. Stat. Comput., Vol. 2, No. 1, pp. 5-16.
- Wahba, G., 1981: Bayesian Confidence Intervals for the Cross Validated Smoothing Spline. Tech. Report No. 645, Dept. of Statistics, University of Wisconsin-Madison.
- Wahba, G., 1982: Personal Communication.
- Wahba, G. and J. Wendelberger, 1980: Some New Mathematical Methods for Variational Objective Analysis Using Splines and Cross Validation. Mon. Wea. Rev., Vol. 108, 8, 1122-1143.
- Wahba, G. and S. Wold, 1975: A Completely Automatic French Curve: Fitting Spline Functions by Cross-Validation. Communications in Statistics, 4, 1-17.
- Wendelberger, J., 1981: The Computation of Laplacian Smoothing Splines with Examples. Tech. Report No. 648, Dept. of Statistics, University of Wisconsin-Madison.
- Wendelberger J., 1982: Multidimensional Smoothing Splines and Their Application. Thesis to Appear. Dept. of Statistics, University of Wisconsin-Madison.

N83

15789

UNCLAS

N83 15789

407

D15

DEPARTMENT OF STATISTICS

University of Wisconsin
1210 W. Dayton St.
Madison, WI 53706

TECHNICAL REPORT NO. 648

September 1981

THE COMPUTATION OF LAPLACIAN
SMOOTHING SPLINES WITH EXAMPLES

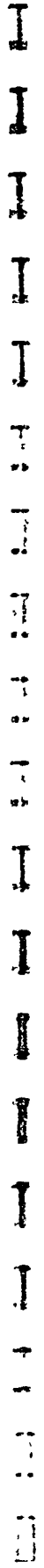
James G. Wendelberger

University of Wisconsin-Madison

This research was supported by NASA under Grant No. NAG5-128 and by the
Office of Naval Research under Contract No. N00014-77-C-0675.

ABSTRACT

Laplacian Smoothing Splines (LSS) are presented as generalizations of graduation, cubic and thin plate splines. The method of generalized cross validation (GCV) to choose the smoothing parameter is described. GCV is used in the algorithm for the computation of LSS's. An outline of a computer program which implements this algorithm is presented along with a description of the use of the program. Examples in one, two and three dimensions demonstrate how to obtain estimates of function values with confidence intervals and estimates of first and second derivatives. Probability plots are used as a diagnostic tool to check for model inadequacy.



1. Motivation

A Laplacian smoothing spline (LSS) is a statistical tool used to model a smooth but otherwise unknown function. The fitted spline provides an analytic function which may be utilized to estimate derivatives, integrals or values of the underlying function. For data analysis purposes a graphical display of the fitted spline (or cross sections for multidimensional problems) often provides insight which might otherwise remain masked by the irregularly spaced, multidimensional and "noisy" data. The residuals, which are the observed values of the dependent variable minus the corresponding fitted spline values, may be utilized as an aid in model checking. A probability plot of the residuals provides a vehicle to detect possibly discrepant observations (outliers). With the above ideas as the eventual objective we first elucidate the functional form of the LSS and then describe an algorithm for its computation.

When someone mentions a line, cosine or an exponential we all have a visual image of "feel" for the function in question. Using the following example we hope to provide an intuitive feeling for an LSS.

In one dimension imagine a long, thin, perfectly rigid rod (a line) lying on a frictionless plane with coordinate axes (t,z) . We represent this rod as a function of t , say $g(t)$. Assume that we are given N points in the plane $\{(t,z):(t,z)=(t_i,z_i), i=1,\dots,N\}$. The t_i are considered to be distinct and known without error. The z_i are measurements of a true but unknown function f evaluated at t_i plus some "noise" e_i . The e_i are independent random variables, each having mean zero and finite variance.

With the previous setup imagine that an ideal spring is attached to data point (t_i, z_i) and to the rod $(t_i, g(t_i))$ for each $i, i=1, \dots, N$. This fixes the springs to remain parallel to the ordinate axis. What position will the rod $g(t)$ assume?

Physics provides a means to answer this question. The rod will assume the position which minimizes the energy of the springs. The energy of an ideal spring is equal to some positive constant k_i (called the spring constant) times the square of the length it is stretched. Thus the cumulative energy of the N springs is

$$\sum_{i=1}^N k_i (z_i - g(t_i))^2 .$$

This is minimized when g is the least squares line (provided we restrict g to be rigid) therefore the least squares line is the position the rod will assume if $k_i = k_0, i=1, \dots, N, k_0$ some constant. If the k_i are not all equal then the rod will assume the position of the weighted least squares line. Notice that this spring idea provides an intuitive explanation for minimizing the residual sum of squares in regression.

The situation is analogous in two dimensions: a thin plate of infinite rigidity (not bendable) would assume the position of the least squares plane. The situation in three dimensions, although not as easy to visualize, is analogous. There are further restrictions on the t_i which are rigorously given in (2.6).

We have thus far assumed that the rod is rigid. This is not necessary and may not be a good representation of the physical phenomenon under

consideration. So we relax the rigidity assumption and assume that the rod is flexible. If zero energy were required to flex the rod then the minimum energy position which the rod would assume is that of a function of interpolation. Since the residuals are zero, this configuration has zero energy and thus is a minimum. By this explanation it is readily seen that the function thus obtained is not unique. This anomaly will be alleviated by requiring energy to flex the rod.

Consider the more realistic case where the rod is flexible and takes energy to flex. The spring of a diving board is testimony to this. Note that the bending energy of a rod is $(\rho/\sigma^2)J_2(g)$, where ρ/σ^2 is a constant and

$$J_2(g) = \int_{-\infty}^{\infty} [g^{(2)}(x)]^2 dx . \quad (1.1)$$

Therefore the bending energy is proportional to curvature which may be measured as $J_2(g)$ in (1.1).

To find the position which the rod will assume under these conditions is equivalent to finding the function g which will minimize the total energy of the system

$$\sum_{i=1}^N k_i (z_i - g(t_i))^2 + (\rho/\sigma^2) J_2(g) \quad (1.2)$$

or equivalently the minimizer of

$$(1/N) \sum_{i=1}^N \sigma^2 k_i (z_i - g(t_i))^2 + (\rho/N) J_2(g) . \quad (1.3)$$

The function from a certain class of functions, X , which minimizes (1.3) can be shown to be a piecewise cubic spline. The function space X is

ORIGINAL PAGE IS
OF POOR QUALITY

rigorously defined in Wahba and Wendelberger (1980). Here x should be thought of as a space of smooth functions which map R^d into R^1 . There is much literature about cubic splines in one dimension. To this author's knowledge the earliest work on LSS's is that of Schoenberg (1964); other important work on splines is given in Craven and Wahba (1979), Duchon (1976), Prenter (1975), and Reinsch (1967).

The one dimensional case generalizes to two dimensions. In two dimensions the splines are called thin plate splines because of the analogy of minimizing the energy of a thin plate of infinite extent. The earliest suggested application of thin plate smoothing splines seems to have been by Harder and Desmarais (1972). They suggested that spring forces may be applied at the points of interpolation. This inspired the spring analogy given here. This spring concept is equivalent to LSS's in either one or two dimensions (with $m=2$ in (2.1)). Much recent work on LSS's has been done by Wahba (see Wahba (1979) and the references cited there).

In two dimensions $J_2(g)$ becomes

$$J_2(g) = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} \sum_{v=0}^2 \binom{2}{v} \left[\frac{\partial^2 g(x_1, x_2)}{\partial x_1^v \partial x_2^{2-v}} \right]^2 dx_1 dx_2. \quad (1.4)$$

$J_2(g)$ is proportional to the bending energy of a thin plate (under simplifying assumptions); for details see Meinguet (1979). However, in two dimensions the solution is no longer a piecewise cubic but rather takes the form

$$g(\underline{t}) = \sum_{i=1}^N c_i \tau_i^2 \ln(\tau_i) + d_0 + d_1 x_1 + d_2 x_2, \quad (1.5)$$

where τ_i is the Euclidean distance between \underline{t} and \underline{t}_i , that is $\tau_i^2 = |\underline{t} - \underline{t}_i|^2$

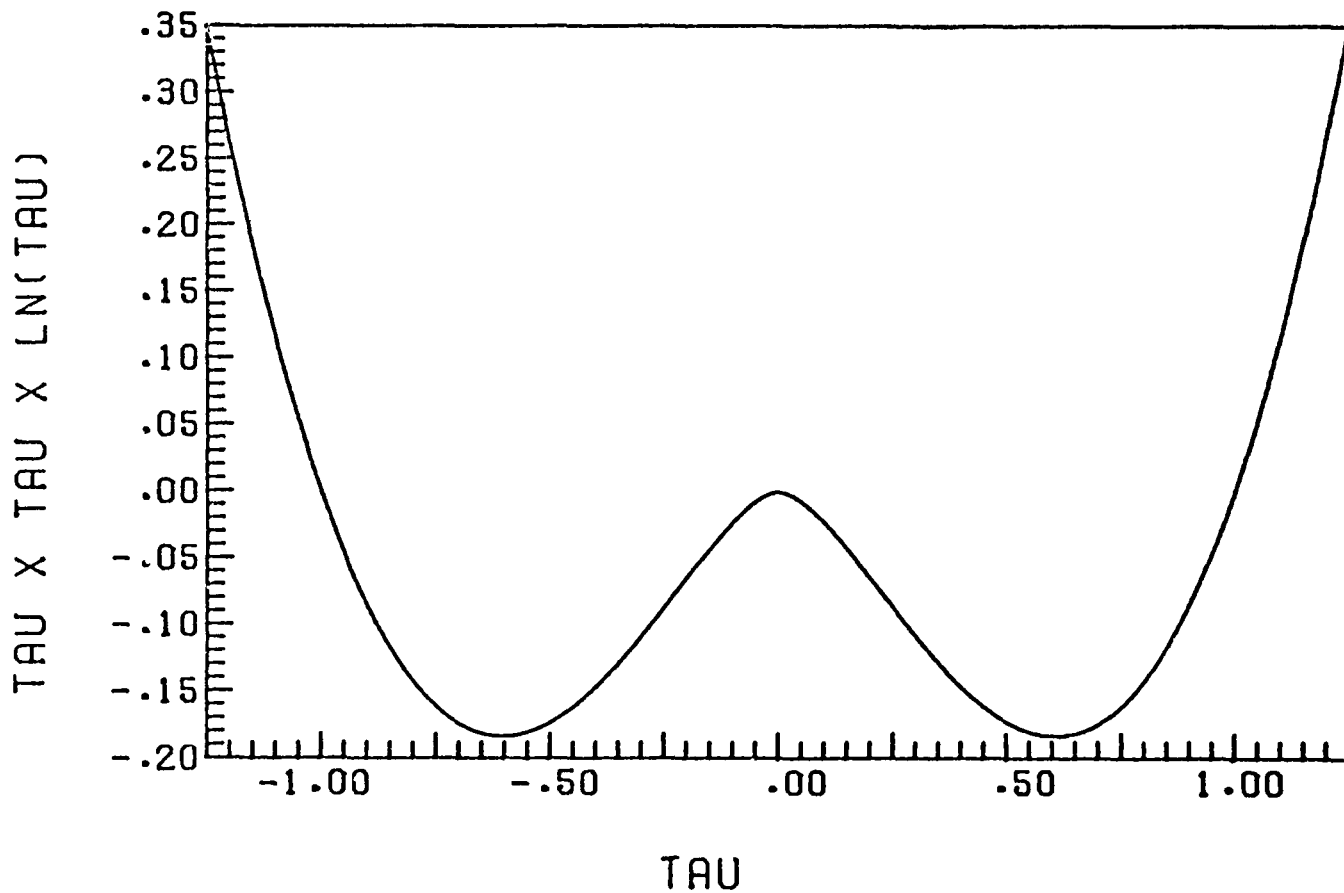
$= (t_{i1} - x_1)^2 + (t_{i2} - x_2)^2$; t_{ij} is the j^{th} component of \underline{t}_i , $j=1,2$,
 $\underline{t}_i = (x_1, x_2)$; c_i^v and d_v are constants, $i=1, \dots, N$, $v=0,1,2$.

To aid in understanding (1.5) the function $\tau_0^2 \ln(\tau_0)$ is plotted in Figure 1.1 for $\underline{t}_0 = (0,0)$ and $x_2 = 0$. Rotation of this function around the ordinate axis and centering at the point \underline{t}_i will produce the radially symmetric function $\tau_i^2 \ln(\tau_i)$. Using (1.5) an LSS is seen to be composed of a linear combination of these radially symmetric functions plus a plane. The plane has zero bending energy but generally does have nonzero spring energy. Linear combinations of the radially symmetric functions can be forced to interpolate the points and hence may have zero spring energy but generally have nonzero bending energy. This tradeoff between bending and spring energy, or smoothness and infidelity to the data (terminology of Wahba (1979)), leads one to consider the minimization problem of Section 2 as a generalization of these ideas. The one and two dimension examples with $m=2$ are special cases of this generalization.

We see that the motivation for one and two dimensional LSS's is quite simple (at least for $m=2$). Attach springs to the data points, constrain them to lie perpendicular to the independent variable space R^d , then let the curve or surface conform by simple bending to the minimum energy configuration.

The Laplacian smoothing spline was suggested by Duchon (1976) as a multidimensional generalization of the thin plate (or "plaques minces"), $d=2$, interpolating spline. An LSS is also a multivariate generalization of the one dimensional, $d=1$, "graduation" spline of Schoenberg (1964). Furthermore, the "graduation" spline is a generalization of the familiar cubic smoothing

TAU VS. TAU X TAU X LN(TAU)



ORIGINAL PAGE IS
OF POOR QUALITY

Figure 1.1: τ_0 vs. $\tau_0^2 \ln \tau_0$.

spline. The terminology "Laplacian smoothing spline" was suggested by Professor I. J. Schoenberg. An explanation for using the term "Laplacian" is given in Wahba (1979).

2. Characterization

Let $z_i = f(t_i) + e_i$, $i=1, \dots, N$. The $t_i \in R^d$ are known exactly. We assume that the function f is smooth but otherwise unknown. By smooth it is meant that the function is well approximated by a function $g \in X$; X is rigorously defined in Wahba and Wendelberger (1980). X may be thought of as a space of functions which approximate well a large class of functions of which f is a member. The e_i are independent, zero mean and finite variance random variables with variance-covariance matrix $\sigma^2 D_\sigma^{-2} = \sigma^2 \text{diag}(\sigma_1^2, \dots, \sigma_N^2)$. Here σ^2 is an unknown constant. For example, if we know that all the variances are equal then we may take $1.0 = \sigma_1^2 = \dots = \sigma_N^2$ in what follows. The σ_i^2 used here are inversely proportional to the k_i of Section 1, that is, $k_i = (\sigma \sigma_i)^{-2}$. The σ_i^2 may be thought of as relative weights of the measurement errors e_i . The z_i are observed dependent variables in R^1 and the corresponding t_i are independent variables in R^d , $i=1, \dots, N$.

A Laplacian smoothing spline is the function g which is the solution to the problem.

Find $g \in X$, X a suitable function space, such that

$$N^{-1} \|D_\sigma^{-1}(z-g)\|^2 + (\rho/N) J_m(g) \quad (2.1)$$

attains its minimum. Here define

$$\begin{aligned} z &= (z_1, \dots, z_N)^T, \quad g = (g_1, \dots, g_N)^T, \quad g_i = g(t_i), \quad \|D_\sigma^{-1}(z-g)\|^2 \\ &= (z-g)^T D_\sigma^{-2} (z-g), \quad D_\sigma^{-2} = \text{diag}(\sigma_1^{-2}, \dots, \sigma_N^{-2}), \end{aligned}$$

where superscript T means transpose throughout. Also,

$$J_m(g) = \sum_{v=1}^M \frac{m!}{\alpha_{1,v}! \dots \alpha_{d,v}!} \int_{-\infty}^{\infty} \dots \int_{-\infty}^{\infty} \left[\frac{\partial^m g(t)}{\partial x_1^{\alpha_{1,v}} \dots \partial x_d^{\alpha_{d,v}}} \right]^2 dx_1 \dots dx_d; \quad (2.2)$$

ORIGINAL PAGE IS
OF POOR QUALITY

$\underline{t} = (x_1, \dots, x_d)^T$; $M' = \binom{m+d-1}{d-1}$; the $\alpha_{1,v}, \dots, \alpha_{d,v}$ are the M' unique combinations of $\{0, 1, \dots, m\}$ such that $\alpha_{1,v} + \dots + \alpha_{d,v} = m$.

In the case presented earlier with $d=2$ and $m=2$ we have $M'=3$ and $(\alpha_{1,v}, \alpha_{2,v})$ takes on the M' unique values $(1,1)$, $(2,0)$ and $(0,2)$. In this case (2.2) reduces to (1.4).

The solution to the minimization problem is unique and given in (2.3).

$$g(\underline{t}) = \sum_{i=1}^N c_i \theta_{m,d} \tau_i^{2m-d} (\ln \tau_i) I_e(d) + \sum_{v=1}^M d_v \phi_v(\underline{t}), \quad (2.3)$$

where I_e is the indicator function of even integers, that is $I_e(d)=1$, for d even and $I_e(d)=0$, for d odd;

$$\theta_{m,d} = \begin{cases} (-1)^{d/2+1+m} / (2^{2m-1} \pi^{d/2} (m-1)! (m-d/2)!), & d \text{ even} \\ \Gamma(d/2-m) / (2^{2m} \pi^{d/2} (m-1)!), & d \text{ odd} \end{cases} \quad (2.4)$$

and ϕ_v are the polynomials of total degree less than m ,

$$\phi_v(\underline{t}) = \phi_v(x_1, \dots, x_d) = x_1^{p_{1v}} \dots x_d^{p_{dv}}. \quad (2.5)$$

Here the ϕ_v are unique; $p_{iv} > 0$, $i=1, \dots, d$ and $p_{1v} + \dots + p_{dv} < m$, $v=1, \dots, M$, $M = \binom{m+d-1}{d}$. Define the M by d matrix P to have iv^{th} element p_{iv} . Also, $2m-d > 0$ and (2.6) holds.

$$\sum_{v=1}^M a_v \phi_v(\underline{t}_i) = 0, \quad i=1, \dots, N \text{ implies } a_v = 0, \quad v=1, \dots, M. \quad (2.6)$$

(Condition (2.6) requires that the matrix T_σ of Section 5 step (ii) be of rank M .) $\underline{c} = (c_1, \dots, c_N)^T$ and $\underline{d} = (d_1, \dots, d_N)^T$ are obtained by solving the linear system

$$(K + \rho \sigma^2 D_\sigma^2) \underline{c} + T_\sigma \underline{d} = \underline{z} \quad (2.7)$$

and

$$T^T \underline{c} = 0. \quad (2.8)$$

In (2.7) K is the N by N matrix with ij th element $\theta_{m,d} \tau_{ij}^{2m-d} (\ln(\tau_{ij})) I_e(d)$. In (2.7) and (2.8) T is the N by M matrix with iv th element $\phi_v(t_i)$. In (2.7) D_σ^2 is the N by N diagonal matrix with ii th entry σ_i^2 . σ^2 is an unknown proportionality constant which along with ρ is absorbed into λ using $N\lambda = \rho\sigma^2$ to yield (2.9) from (2.7).

$$(K + N\lambda D_\sigma^2) \underline{c} + T \underline{d} = \underline{z} \quad (2.9)$$

The approach of Harder and Desmarais (1972) provides us with a physical interpretation of the parameters at least in the $d=2$ case. $\rho = N\lambda\sigma^{-2}$ is the plate "rigidity" which is a constant. The value of ρ depends on the material and the thickness of the plate. The spring constant k_j is equal to the reciprocal of the variance or $(\sigma\sigma_j)^{-2}$. The "load" at the j th point is $P_j = \rho c_j = (\sigma\sigma_j)^{-2} r_j = k_j r_j$, where r_j is the unnormalized or unscaled residual at that point; i.e., $r_j = z_j - g(t_j)$, $j=1, \dots, N$ or $\underline{r} = \underline{z} - K\underline{c} - T\underline{d}$.

For a discussion of a more general problem and the derivation of the solution the reader is referred to Wahba and Wendelberger (1980). We note here that if the e_i are not independent but instead have positive definite covariance matrix proportional to Σ then D_σ^2 and D_σ^{-1} are everywhere replaced by Σ and the symmetric inverse square root $\Sigma^{-1/2}$ to obtain the solution.

To this point we have assumed knowledge of the smoothness parameter λ . However it is generally unknown. Before describing a method to dynamically choose λ from the data at hand we provide an example to exhibit its influence on the LSS.

3. Example 1—Variation of the LSS with λ , $d=1$.

A company which makes and repairs small computers wants to forecast the number of service engineers that it will require over the next few years. To do this requires, among other things, knowledge of the length of a service call. The length of a call is a function of the number of components within the computer which must be repaired or replaced. The information in Table 3.1 was collected on 24 service calls; the data are from Chatterjee and Price (1977). We would like to fit a spline to the data in order to forecast the length of a service call.

We fit a spline to the data using the algorithm given in Section 5. The smoothness parameter, λ , is dynamically chosen from the data using the method of generalized cross validation (GCV). By showing the influence of λ on the LSS of this example we hope to provide a clearer understanding of the role of GCV in choosing the smoothness parameter. The results of the following sections will be easier to understand with this example in mind. Exactly what the GCV choice of λ is will be presented in Section 4.

Figure 3.1 shows a plot of the data and the corresponding spline for five different values of λ . Because there are only 24 observations of which only 17 have unique independent variables we should not be surprised if the GCV estimate (to be described in Section 4) of λ , which is a large sample result, does not perform well. The confidence intervals are calculated using method of Wahba (1981); the formula used for their computation is given in Example 2 of Section 6.

TABLE 3.1

EXAMPLE 1 - REPAIR TIMES

Length of Calls (Minutes)	Units Repaired (Number)
23	1
29	2
49	3
64	4
74	4
87	5
96	6
97	6
109	7
119	8
149	9
145	9
154	10
166	10
162	11
174	11
180	12
176	12
179	14
193	16
193	17
195	18
198	18
205	20

M = 2
LAMBDA = 0.00

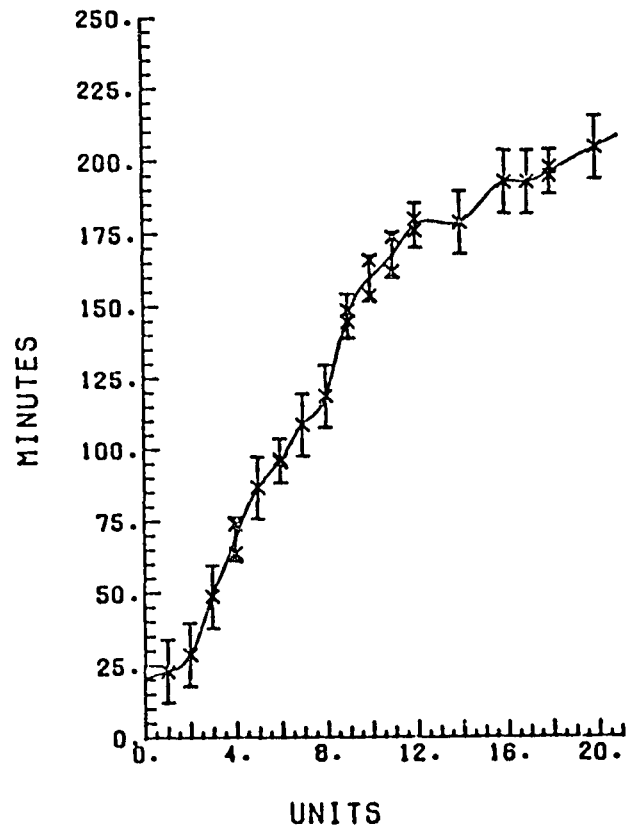


Figure 3.1a: Example 1 with $\lambda = 0.00$

M = 2
LAMBDA = .0166

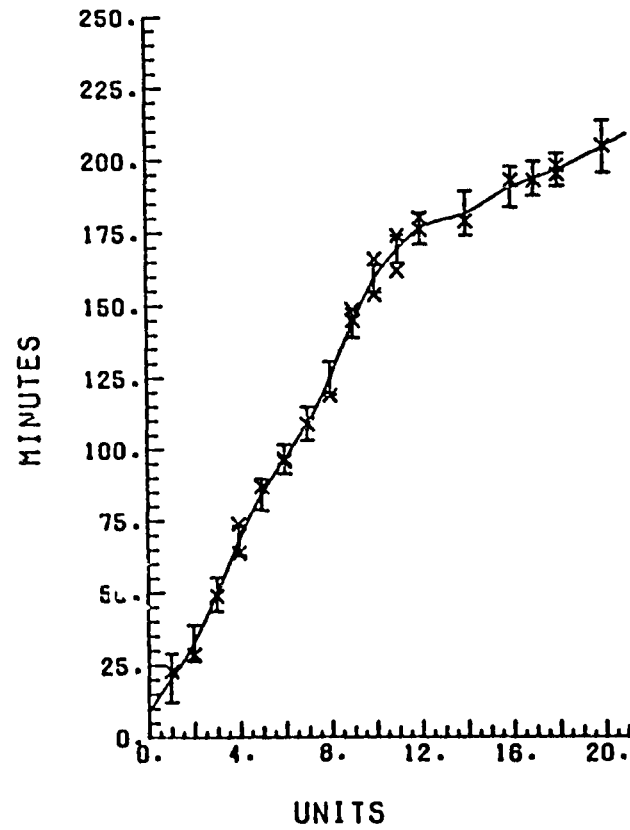


Figure 3.1b: Example 1 with $\lambda = 0.0166$

ORIGINAL PAGE IS
OF POOR QUALITY

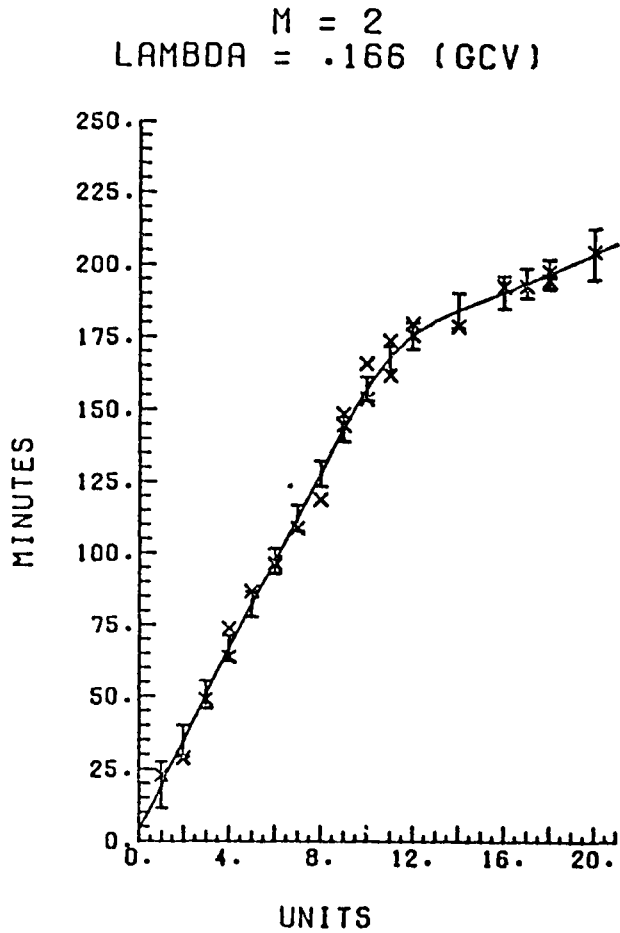


Figure 3.1c: Example 1 with $\lambda = .166$,
the GCV choice

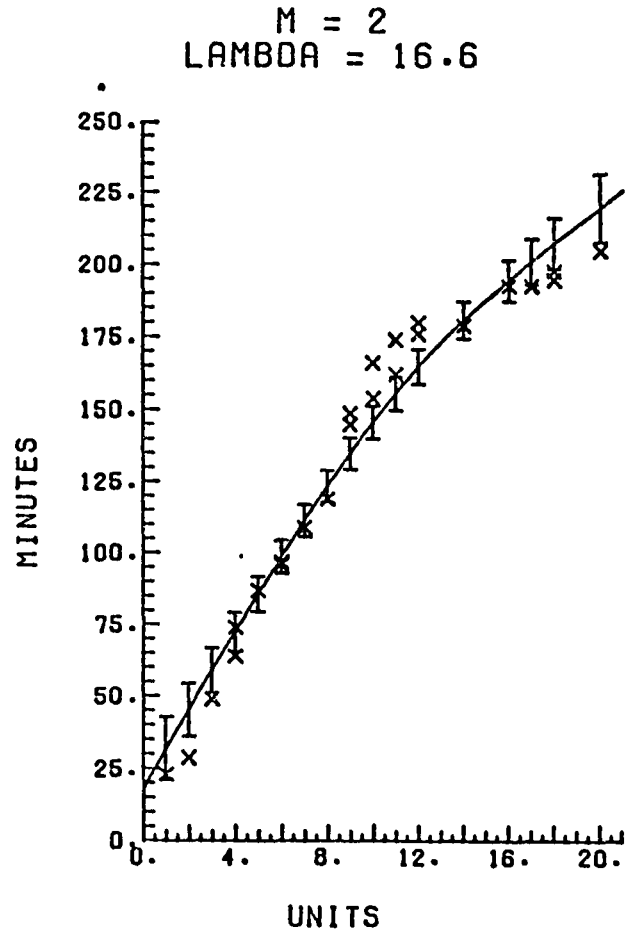
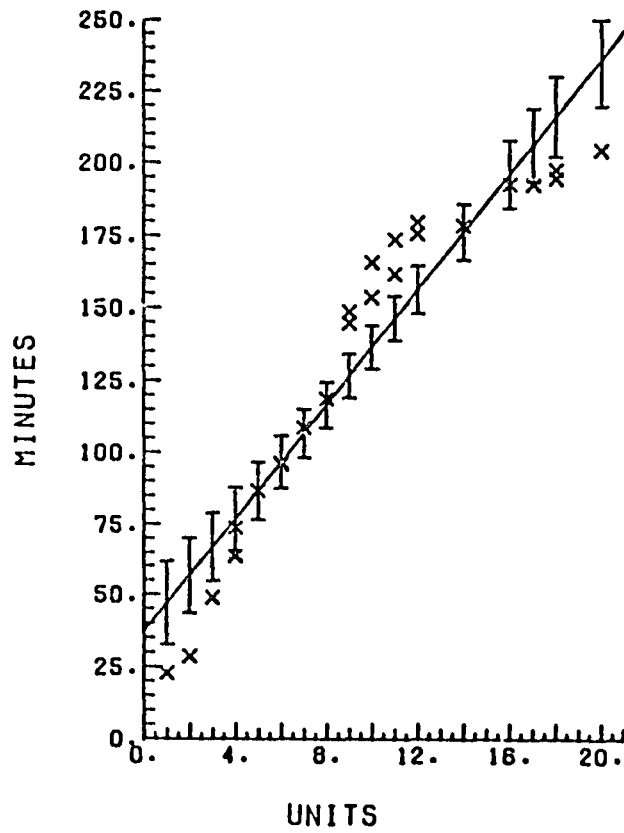


Figure 3.1d: Example 1 with $\lambda = 16.6$

ORIGINAL PAGE IS
OF POOR QUALITY

M = 2
LAMBDA = INFINITY



ORIGINAL PAGE IS
OF POOR QUALITY

Figure 3.1e: Example 1 with $\lambda = \infty$

Considering the brief explanation of the problem given here the GCV choice of λ , as used in Figure 3.1c, seems reasonable to use in predicting the number of minutes spent. The GCV choice of λ appears to be the most visually pleasing and consistent with how we would expect the number of minutes spent on a service call to be related to the number of computer components repaired.

4. Generalized Cross Validation

In the example of Section 3 the smoothing parameter λ is unknown. To determine an estimate of this parameter Craven and Wahba (1979) and Wahba and Wold (1979) have suggested the use of generalized cross validation. A short synopsis of the development of this method is given to enhance the understanding of it.

The method of cross validation (presented here as related to LSS's) is developed in response to the question: How well may one expect LSS's to predict the true functional value $g(t)$ at some point t ?

Simple cross validation (SCV) suggests predicting the true functional values of data different from that used in the analysis to assess this predictive ability. In its simplest form this entails dividing the sample into two pieces of similar size using one section for optimization and the other for testing. In addition to this, in order to gain more information from the data, the two pieces may be interchanged and the optimization and testing performed on each.

SCV is alright if there is an ample supply of data so that halving or doubling it has little effect on the quality of the estimator. To lessen this effect Mosteller and Tukey (1968) propose single cross validation (1CV), (called ordinary cross validation by Wahba (1979)), which is described suitably by them as follows:

"Suppose that we set aside one individual case, optimize for what is left, then test on the set-aside case. Repeating this for every case squeezes the data almost dry. If we have to go through the full

ORIGINAL PAGE IS
OF POOR QUALITY

optimization calculation every time, the extra computation may be hard to face. Occasionally, one can easily calculate, either exactly or to an adequate approximation, what the effect of dropping a specific and very small part of the data will be on the optimized result. This adjusted optimized result can then be compared with the values for the omitted individual. That is, we make one optimization for all the data, followed by one repetition per case of a much simpler calculation, a calculation of the effect of dropping each individual, followed by one test of that individual. When practical, this approach is attractive."

To describe ICV mathematically we require some notation. Let $g_\lambda(j)$ be the solution to the minimization of (2.1) with the j^{th} point removed from the analysis. Similarly, $D_\sigma(j)$ is the $N-1$ by $N-1$ matrix composed of D_σ with its j^{th} row and column removed. To "test on the set aside case" we require that $[(g_\lambda(j)(t_j) - z_j)/\sigma_j]^2$ be small. "Repeating this for every case" and averaging to yield an overall test gives

$$V_m^0(\lambda) = (1/N) \sum_{j=1}^N [(g_\lambda(j)(t_j) - z_j)/\sigma_j]^2 . \quad (4.1)$$

ICV uses the λ which minimizes $V_m^0(\lambda)$.

To minimize $V_m^0(\lambda)$ directly is not a trivial computational matter. For each proposed value of λ a system of the form (2.8) and (2.9) (of order $N+M-1$ instead of $N+M$) must be solved for each of the N values left out of the analysis. This entails solving a linear system of order $N+M-1$ N times! As noted earlier "if we have to go through the full optimization calculation every time, the extra computation may be hard to face." Following the idea of Mosteller and Tukey we seek a computational simplification for the minimizer of $V_m^0(\lambda)$.

ORIGINAL PAGE IS
OF POOR QUALITY

The simplified form for 1CV was first noted by Craven and Wahba (1979) and given in a slightly more general form in Wahba and Wendelberger (1980). The 1CV function may be written

$$V_m^0(\lambda) = (1/N) \sum_{j=1}^N [(g_\lambda(t_j) - z_j) / (\sigma_j(1-a_{jj}(\lambda)))]^2 . \quad (4.2)$$

$a_{jj}(\lambda)$ is the j th diagonal element of $A_m(\lambda)$ which is defined by

$$A_m(\lambda)z = \begin{pmatrix} g_\lambda(t_1) \\ \vdots \\ g_\lambda(t_N) \end{pmatrix}$$

where g_λ is the solution of (2.1). $A_m(\lambda)$ may be thought of as mapping the vector z into the smoothed values.

In this form "we make one optimization for all the data" by calculating g_λ then "followed by one repetition per case of a much simpler calculation, a calculation of the effect of dropping each individual." Here find $a_{jj}(\lambda)$ and use (4.2).

Evaluation of this formulation of $V_m^0(\lambda)$ involves solving a linear system of size $N+M$ to find g_λ and one of size N to find $a_{jj}(\lambda)$. This is a considerable improvement over that of using (4.1) directly. Because of a mathematical simplification the amount of computation needed to minimize $V_m^0(\lambda)$ can be substantially reduced. From a practical point of view this makes the use of cross validation very attractive.

When applying cross validation to problems other than LSS's this last step of finding "what the effect of dropping a specific and very small part of the data will be on the optimized result" is very important and should not be

overlooked. In fact, this step often makes cross validation computationally feasible whereas without this insight it may be impractical.

Finding the minimizer of $V_m^0(\lambda)$ requires its evaluation at different values of λ as determined by a search routine. Hence, although the minimization is possible we need to repeatedly solve large linear systems with the number of solution times being a function of the search routine employed.

In $V_m^0(\lambda)$ of (4.1) each deviation of $g_\lambda(t_j)$ from the observed value z_j is treated symmetrically. This choice is arbitrary and is chosen for simplicity. A more general approach is to weight each term of (4.1) or equivalently (4.2) to yield

$$V_m(\lambda) = (1/N) \sum_{i=1}^N W_i [(g_\lambda(t_i) - z_i) / (\sigma_i(1-a_{i1}(\lambda)))]^2. \quad (4.3)$$

Before a discussion of the choice of these weights the following definition is needed.

Definition:

$$R_m(\lambda) = E(1/N) \sum_{i=1}^N [(f(t_i) - g_\lambda(t_i)) / \sigma_i]^2$$

is the expected weighted (by σ_i) mean squared error between the true function (f) and the spline (g_λ) evaluated at the independent variables (t_i). Here E denotes mathematical expectation with respect to the error distribution of the random errors as described in the model of Section 2.

If we want $R_m(\lambda)$ to be small then the generalized cross validation value of λ should be used as the smoothing parameter value. Using ICV as motivation

Craven and Wahba (1979) and Golub, Heath and Wahba (1979) have shown that the λ which minimizes $V_m(\lambda)$ with weights

$$W_i = (1 - a_{ii}(\lambda))^2 / (1 - N^{-1} \sum_{j=1}^N a_{jj}(\lambda))^2$$

is an estimate of the λ which minimizes $R_m(\lambda)$. Using these weights in (4.3) gives the generalized cross validation function (GCVF)

$$V_m(\lambda) = (1/N) \sum_{i=1}^N [(g_\lambda(\underline{t}_i) - z_i) / (\sigma_i (1 - N^{-1} \sum_{j=1}^N a_{jj}(\lambda)))]^2. \quad (4.4)$$

The minimizer of (4.4) is called the GCV estimate of λ .

The GCVF can be rewritten as

$$V_m(\lambda) = (1/N) \| |D_\sigma^{-1} (I - A_m(\lambda)) \underline{z}| \|^2 / ((1/N) \text{Tr}(I - A_m(\lambda)))^2; \quad (4.5)$$

where Tr is the trace.

Wahba (1981) has proposed

$$\sigma_e^2 = \| |D_\sigma^{-1} (I - A_m(\lambda)) \underline{z}| \|^2 / \text{Tr}(I - A_m(\lambda)) \quad (4.6)$$

as an estimate of the error variance σ^2 . This leads us to consider $df_e = \text{Tr}(I - A_m(\lambda))$ as the degrees of freedom of error. Using these notions we rewrite the GCVF as

$$V_m(\lambda) = N \sigma_e^2 / df_e. \quad (4.7)$$

The method of GCV may be viewed as minimizing the estimated error variance per error degrees of freedom. This may further be thought of as a form of parsimonious model selection.

In the next section we see that the computation of $V_m(\lambda)$ is reduced to essentially the singular value (or eigenvalue-eigenvector) decomposition of a symmetric positive definite $N-M$ by $N-M$ matrix (M is usually a small integer). The above decomposition makes it possible to form $V_m(\lambda)$ by simple scalar operations for each value of λ . Thus we have taken the ideas of Mosteller and Tukey one step further. This algorithm is much simpler than the original analysis at essentially the cost of a one time eigenvalue-eigenvector decomposition; i.e., changing the dependent variable (but not the independent variables) does not necessitate another spectral decomposition. Thus, many data sets which have identical independent variables but different dependent variables may be analyzed quite easily and inexpensively.

When using GCV with a small sample size we may run into problems. The most frequent small sample problem with GCV is that $\lambda = 0$ or $\lambda = \infty$ is chosen when physical considerations dictate that it should not be. $\lambda = 0$ implies that we are interpolating the dependent variable. This should be done if the true underlying rigidity ρ is zero. λ equal to infinity implies that we are fitting a polynomial of degree $m-1$ by least squares. This should be done if either the variance is large (relative to the dependent variable) or if the true underlying rigidity is infinite (i.e., the true model is a polynomial). If it is clear from other considerations that the value of λ chosen is not indicative of the actual underlying mechanism then that particular value should not be used and the model assumptions should be checked for violations.

The choice of m can also be made by GCV, see Lucas (1978) and Wahba and Wendelberger (1980).

5. Algorithm

The user must supply N independent variables, $\underline{t}_i \in \mathbb{R}^d$, $i=1, \dots, N$, and their corresponding dependent variables, $z_i \in \mathbb{R}^d$, $i=1, \dots, N$ to compute the LSS at a point $\underline{t} \in \mathbb{R}^d$. Assume that the model described in Section 2 holds. In particular, assume the independent variables \underline{t}_i are known without error and the dependent variables z_i consist of the true function value at \underline{t}_i , $f(\underline{t}_i)$, plus "noise," e_i , $z_i = f(\underline{t}_i) + e_i$. The e_i are independent with finite variance $\sigma^2 \sigma_i^2$, σ^2 an unknown constant.

To produce the coefficients \underline{c} and \underline{d} needed to evaluate the spline we solve the linear system of equations

$$(K + N\lambda^* D_{\sigma^2}) \underline{c} + T \underline{d} = \underline{z}$$

and

$$T^T \underline{c} = 0.$$

In this system λ^* is the optimal value of the smoothing parameter λ as determined by the generalized cross validation function. If λ^* is known then the solution of the above linear system could be accomplished for relatively large values of N . However, it is usually unknown and must be calculated in order to solve the system of equations.

The method currently used to determine λ^* requires the solution of a symmetric $N \times M$ dimensional eigenvalue-eigenvector problem. This is the current computational barrier to solving problems with large numbers of observations.

The algorithm presented in Wahba and Wendelberger (1980) requires the inversion of a matrix of order M and two eigenvalue-eigenvector decompositions of symmetric matrices, one N by N and the other (positive definite)

N-M by N-M. The algorithm presented here requires the solution of a triangular system of order M, the QR-decomposition of an N by M matrix and the singular value (or eigenvalue-eigenvector) decomposition of a symmetric positive definite N-M by N-M matrix. This algorithm is faster and requires fewer operations, primarily because of the replacement of one N by N eigenvalue-eigenvector decomposition by the QR-decomposition of an N by M matrix ($M < N$).

This algorithm provides for replicated points. A replicated point is one for which there is more than one observation of the dependent variable for a particular value of the independent variable. Let the total number of unique (independent variable) points be N_N and define $N_0 = N - M - N_N$. Then the computational algorithm is as follows:

(i) Compute $T_\sigma = D_\sigma^{-1}T$.

(ii) Perform the QR-decomposition described in Dongarra, et al., (1979), of T_σ .

$$T_\sigma = (Q_1, Q_2) \times (R^T, 0)^T .$$

(iii) Calculate $B = Q_2^T D_\sigma^{-1} K D_\sigma^{-1} Q_2$.

(iv) Decompose $B = (U_1, U_2) D_B (U_1, U_2)^T$,

using the singular value decomposition of B, as described by Golub and Reinsch (1970) or using the spectral decomposition of B as described by Smith, et al., (1976); where

$D_{B'}$ - diagonal matrix of the eigenvalues (b_i) of B, which is of dimension $N-M$ by $N-M$,

D_B - diagonal matrix of the positive eigenvalues (b_i) of B, which is of dimension N_N by N_N ,

U_1 - the eigenvectors of the positive eigenvalues of B, which is of dimension $(N-M)$ by N_N , and

U_2 - the eigenvectors of the zero eigenvalues of B, which is of dimension $(N-M)$ by N_0 .

(v) Form $\underline{w} = U_1^T Q_2^T D_{\sigma}^{-1} \underline{z}$,
 $\underline{w}^T = (w_1, \dots, w_{N_N})$.

(vi) Obtain λ^* as the minimizer of

$$v_1(\lambda) = \begin{cases} N \sum_{i=1}^{N_N} [w_i / (b_i / N + \lambda)]^2 / \left(\sum_{i=1}^{N_N} (1 / (b_i / N + \lambda)) \right)^2, & \lambda \neq \infty \text{ and } N-M = N_N \\ N [\underline{z}_{\sigma}^T Q_2 Q_2^T \underline{z}_{\sigma} - \underline{w}^T \underline{w} + \lambda^2 \sum_{i=1}^{N_N} (w_i / (b_i / N + \lambda))^2] + & (5.1) \\ (N-M-N_N + \lambda \sum_{i=1}^{N_N} (1 / (b_i / N + \lambda)))^2, & \lambda \neq \infty \text{ and } N-M \neq N_N \\ N \underline{z}_{\sigma}^T Q_2 Q_2^T \underline{z}_{\sigma} / (N-M)^2, & \lambda = \infty \end{cases}$$

where $\underline{z}_{\sigma} = D_{\sigma}^{-1} \underline{z}$.

(vii) Calculate

$$\underline{c} = \begin{cases} D_{\sigma}^{-1} Q_2 U_1 D_B^{-1} U_1^T Q_2^T \underline{z}_{\sigma} , & \lambda = 0 \\ D_{\sigma}^{-1} Q_2 U_1 [(D_B + N\lambda I)^{-1}] U_1^T Q_2^T \underline{z}_{\sigma} , \\ & 0 < \lambda < \infty \text{ and } N-M = N_N \\ D_{\sigma}^{-1} Q_2 U_1 [(D_B + N\lambda I)^{-1} - (N\lambda)^{-1} I] U_1^T Q_2^T \underline{z}_{\sigma} \\ & + (N\lambda)^{-1} D_{\sigma}^{-1} Q_2 Q_2^T \underline{z}_{\sigma} , & 0 < \lambda < \infty \text{ and } N-M \neq N_N \\ 0 , & \lambda = \infty . \end{cases} \quad (5.2)$$

(viii) Solve the triangular system.

$$R \underline{d} = Q_1^T D_{\sigma}^{-1} (\underline{z} - K \underline{c}) \text{ for } \underline{d},$$

$$\underline{d}^T = (d_1, \dots, d_M) .$$

An important aspect of this method is the relatively small cost of reconstructing a new LSS using the identical independent variables while changing only the dependent variables. To see this notice that the bulk of the computational effort is in steps (i) through (iv) which do not require knowledge of the dependent variables. These steps depend upon the independent variables and D_{σ} . To construct a second LSS with the same independent variables and identical D_{σ} we need only save the matrices U_1 , D_{σ} , D_B , Q_1 , Q_2 and R . With these matrices we perform steps (v) through (viii) to produce a spline for another set of dependent variables, say \underline{z}' , with little additional computational effort.

The fact that obtaining another spline from \underline{z}' is easy requires further consideration. It is made possible because of the necessity to minimize the

GCVF. This minimization provides the mechanism to easily calculate \underline{c} and \underline{d} in steps (vii) and (viii) of the algorithm. If λ^* was somehow known a priori then we could go right ahead and solve the linear system (2.8) and (2.9) at a much less one time cost. However, even with λ^* known, if we had many new data sets \underline{z}' then for some number of them it indeed would be easier to do the spectral decomposition once and for all.

Instead of saving U_1 , D_σ , D_B , Q_1 , Q_2 and R we actually save $Q_2 U_1$, D_σ , D_B , $Q_1^T D_\sigma^{-1} K$ and the QR-decomposition of T_σ to retrieve R , $Q_2 Q_2^T$ and Q_1 . By using these matrices we can perform steps (v) through (viii) quite inexpensively. The QR-decomposition can be stored in the storage which has been allocated for T_σ plus M additional storage locations. $Q_1^T D_\sigma^{-1} K$ is retained so that it is unnecessary to reevaluate K .

6. Example 2--Franke's Principal Test Function, $d=2$.

Example 2 is a Monte Carlo experiment to demonstrate the surface ($d=2$) which may be obtained by using an LSS with GCV. The "principal test function" of Franke (1979) is used as the true function f . This surface consists of two Gaussian peaks and one Gaussian dip superimposed on a surface sloping towards the first quadrant. The surface is defined by

$$\begin{aligned} f(x,y) = & .75 \exp -[[(9x-2)^2 + (9y-2)^2] / 4] \\ & + .75 \exp -[[(9x+1)^2 / 49] + [(9y+1) / 10]] \\ & + .50 \exp -[[(9x-7)^2 + (9y-3)^2] / 4] \\ & - .20 \exp -[[(9x-4)^2 + (9y-7)^2] \end{aligned}$$

A plot of the surface f is given in Figure 6.1.

The surface is reconstructed from 169 "noisy" observations on the grid

$$G = \left\{ \underline{t}_i \mid \underline{t}_i = \left(\frac{2j-1}{26}, \frac{2k-1}{26} \right), i=13(j-1)+k; j,k=1,\dots,13 \right\} .$$

The "noisy" observations are

$$z_i = f(\underline{t}_i) + e_i \text{ with } e_i \sim N(0, \sigma^2), i=1, \dots, 169, \sigma^2 = (.03)^2.$$

The e_i are generated by the pseudo random number generator RAENBR at the Madison Academic Computing Center, MACC (1978). The LSS with $m=2$ and the smoothing parameter chosen by GCV is plotted in Figure 6.2. The closeness of fit can be qualitatively seen by overlaying Figure 6.2 on Figure 6.1.

For this example the calculated $\sigma_e^2 = (.026)^2$, (using (4.6)), compares favorably with the true $\sigma^2 = (.03)^2$. Using σ_e^2 to obtain confidence intervals for the true curve at the grid points G as in Wahba (1981) gives the 95% confidence intervals

ORIGINAL PAGE IS
OF POOR QUALITY

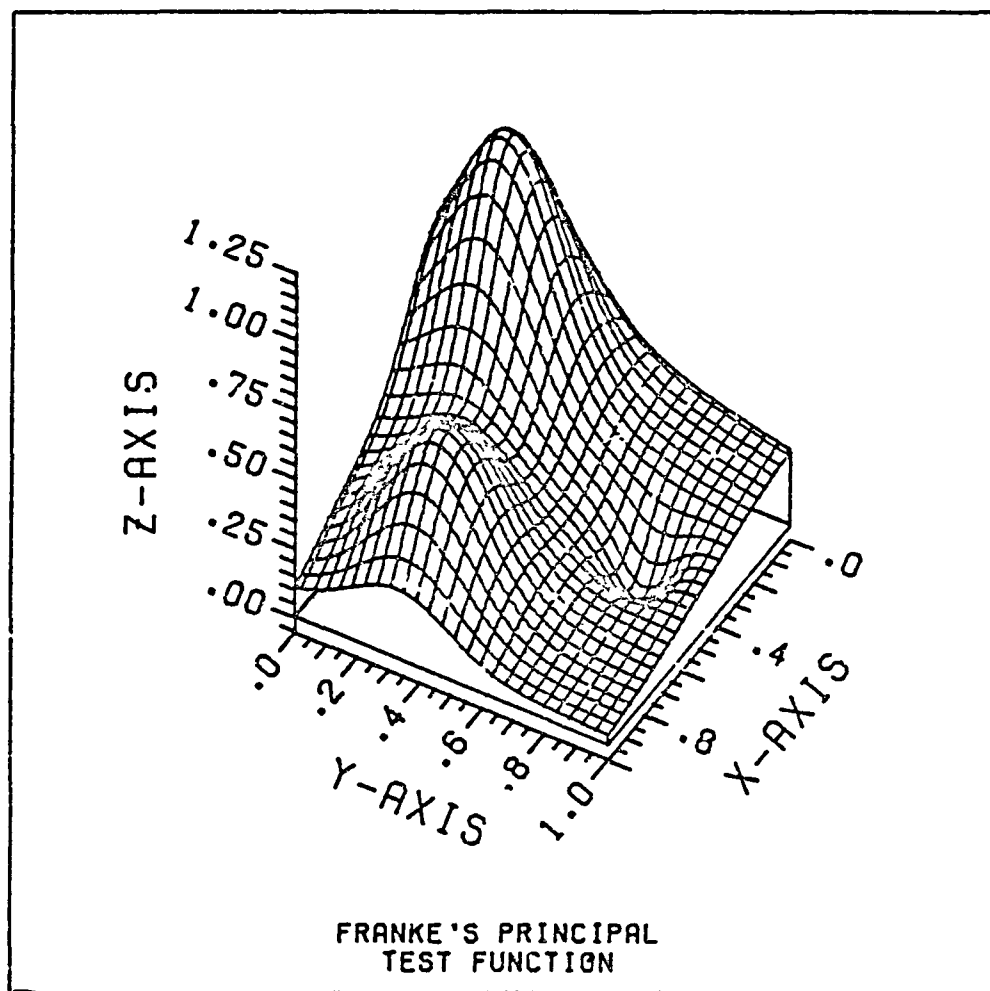


Figure 6.1: Example 2--Frank 's Principle Test Function

ORIGINAL PAGE IS
OF POOR QUALITY

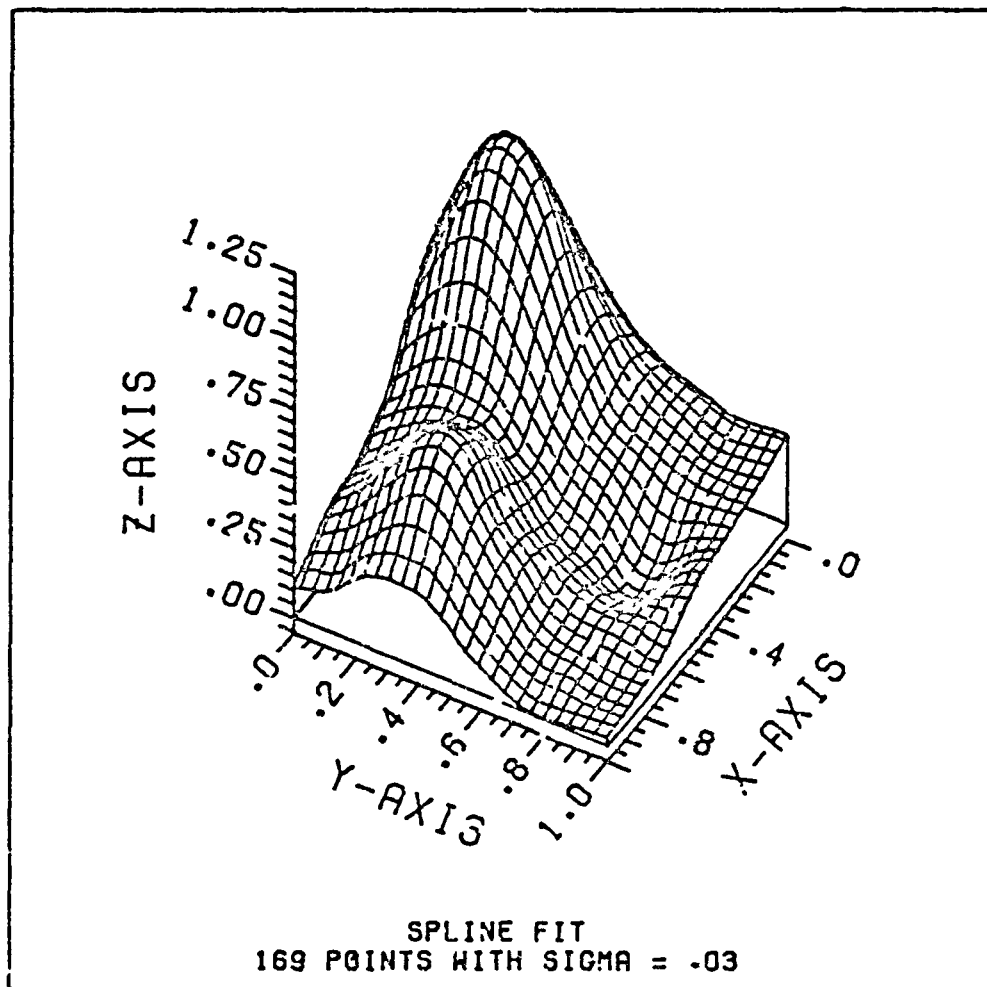


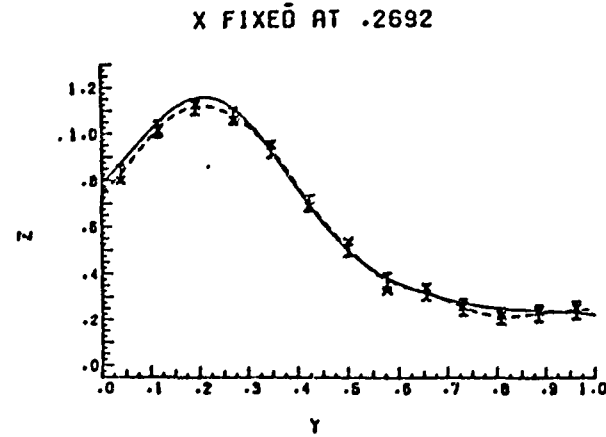
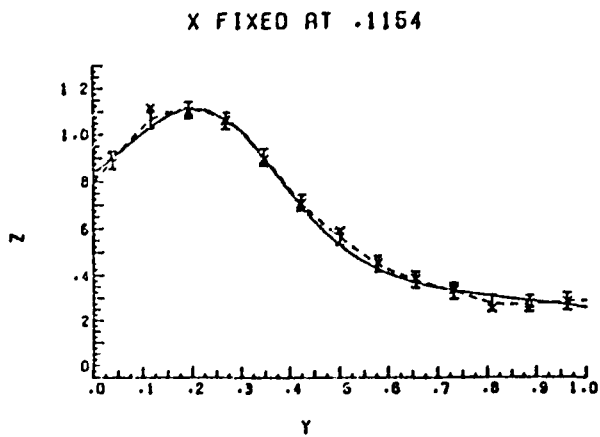
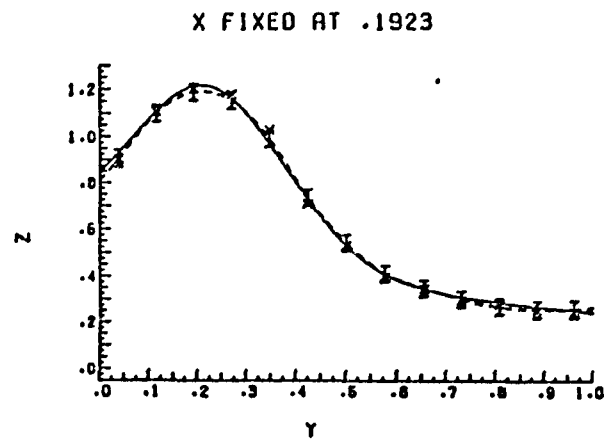
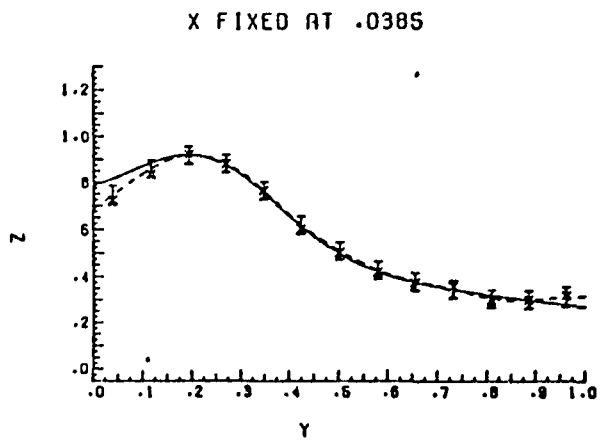
Figure 6.2: Plot of the $m = 2$, GCV λ , spline fit to Franke's Principal Test Function from 169 "noisy" points with $\sigma = .03$

$$g_{\lambda^*}(t_i) \pm 1.96\sigma_e\sigma_1(a_{1i}(\lambda^*))^{1/2}, \quad i=1,\dots,N.$$

Figure 6.3 gives the cross section along the grid showing the true curve, spline fit, observation and 95% confidence interval at each point for each value of x_1 , $i=1,\dots,13$.

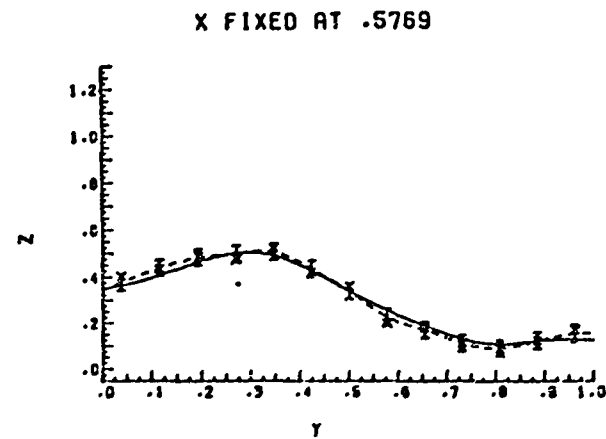
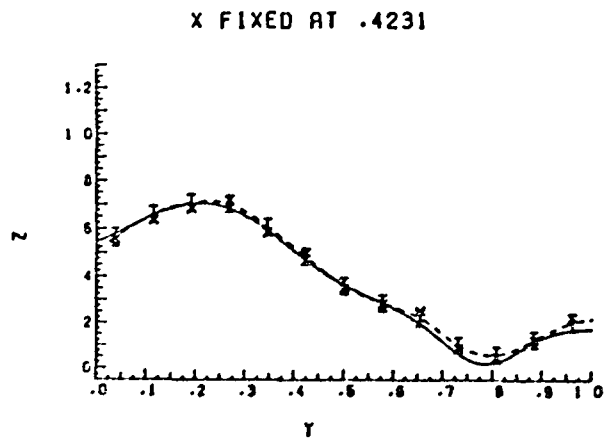
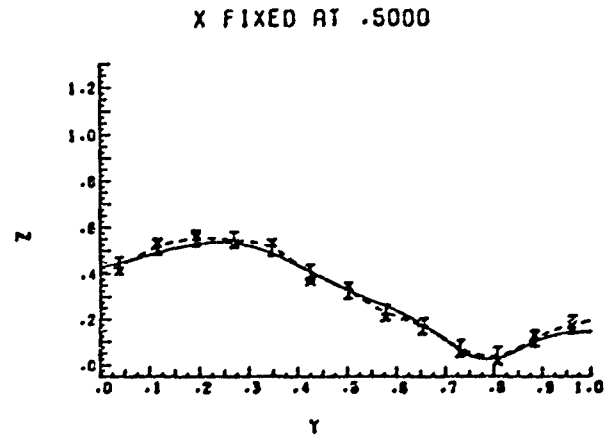
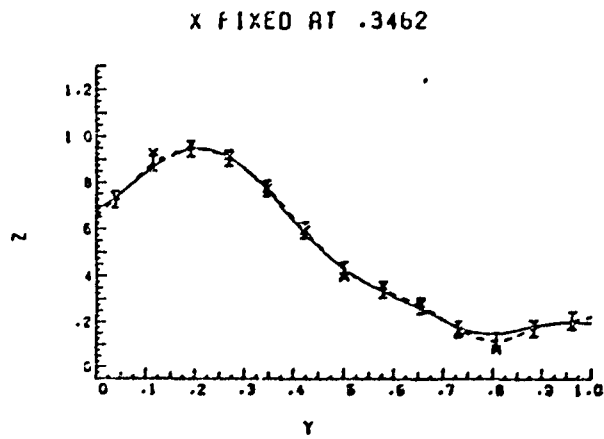
The number of 95% confidence intervals which cover the true surface is known because the true surface is known. For this example 162 or 95.9% of the intervals cover the true surface. This is a favorable comparison since the expected number is 161. This example was not chosen because of this agreement but rather was the only one run by prior decision.

The example given here uses points on a grid only for clarity of display. For other $d=2$ Monte Carlo results see Wahba and Wendelberger (1980). The meteorological example given there uses irregularly spaced points.



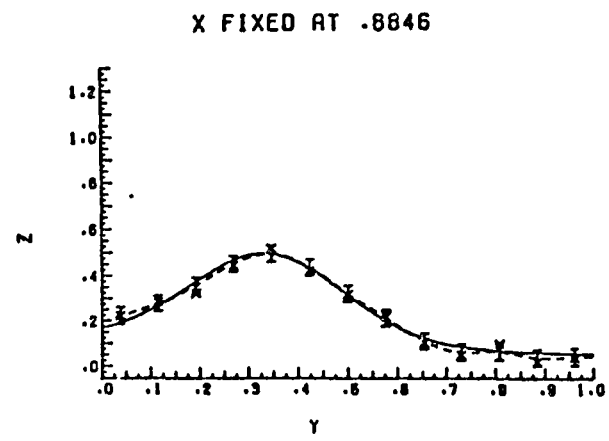
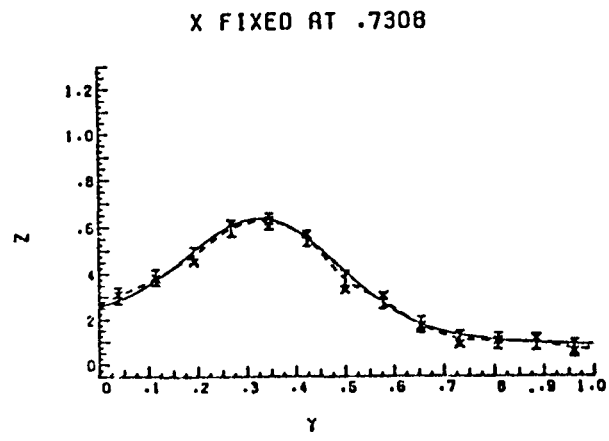
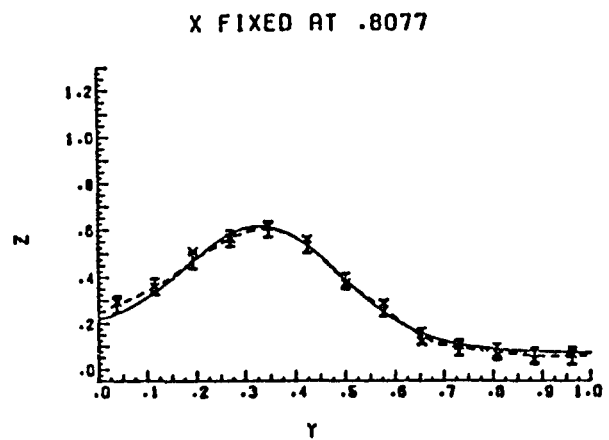
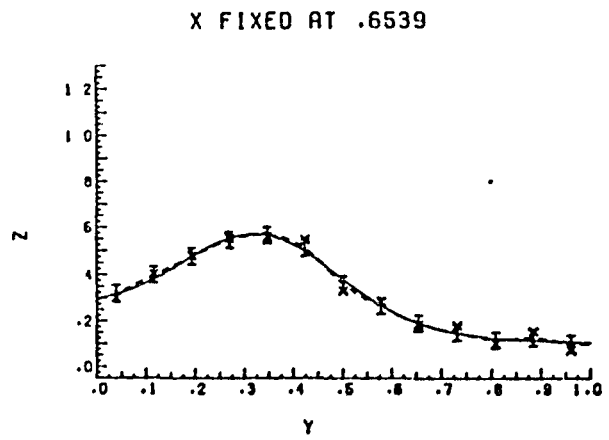
ORIGINAL PAGE IS
OF POOR QUALITY

Figure 6.3: 13 Cross sections of Example 2--true curve (solid line), spline fit (dashed line)



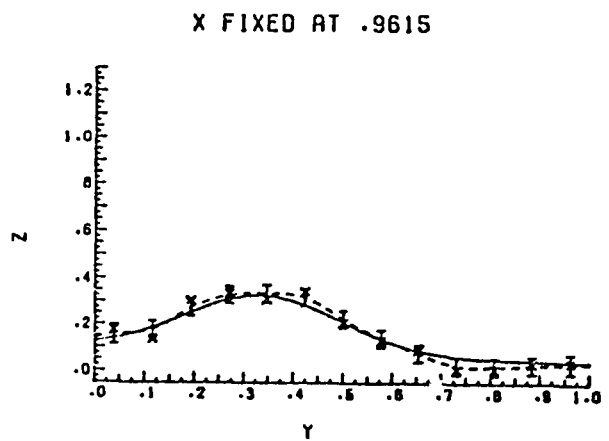
ORIGINAL PAGE IS
OF POOR QUALITY

Figure 6.3: 13 Cross sections of Example 2--true curve (solid line), spline fit (dashed line)



ORIGINAL PAGE IS
OF POOR QUALITY

Figure 6.3: 13 Cross sections of Example 2--true curve (solid line), spline fit (dashed line)



ORIGINAL PLOT OF POOR QUALITY

Figure 6.3: 13 Cross sections of Example 2--true curve (solid line), spline fit (dashed line)

ORIGINAL PAGE IS
OF POOR QUALITY

7. Example 3—Derivatives and Outliers, $d=3$.

Example 3 is a Monte Carlo experiment with $d=3$ and true function

$$f(x_1, x_2, x_3) = (2\pi)^{-3/2} \exp [(x_1^2 + 4x_2^2 + 9x_3^2)/(-2)].$$

Contours of f , f' and f'' are given as the solid lines in Figures 7.4, 7.5 and 7.6.

Three hundred points t_i , $i=1, \dots, 300$ are taken from a uniform distribution in $R = \{(x_1, x_2, x_3) | -2 < x_1 < 2, -1 < x_2 < 1, -2/3 < x_3 < 2/3\}$. The true function f is evaluated at each of the points t_i and added to a Gaussian pseudo random variable with standard deviation $\sigma = .0025$ to yield observation z_i . The peak height of f is approximately .0634. σ is roughly 4% of the peak height and therefore these data have a "typical" noise level.

A value of $m=4$ was chosen for this example in order that the second derivative of the spline could be used as an estimate of the second derivative of f . If k is the order of the derivative desired then $2m-2k-d$ must be positive. Here $2 \times 4 - 2 \times 2 - 3 = 1 > 0$ and so the second derivative of the LSS will be a good estimate of the second derivative of f ; for details see Wahba and Wendelberger (1980).

The estimate σ_e for this experiment is .0024 which agrees nicely with the true value of .0025.

Contours of the true function and the fitted spline, g_{λ^*} , are plotted in Figure 7.4 for 4 values of x_3 . Because of the symmetry of the true surface it was not plotted for negative values of x_3 . The true function and the fitted spline are close to one another near the center of the region and this closeness degrades as we approach the boundary in each of the three directions.

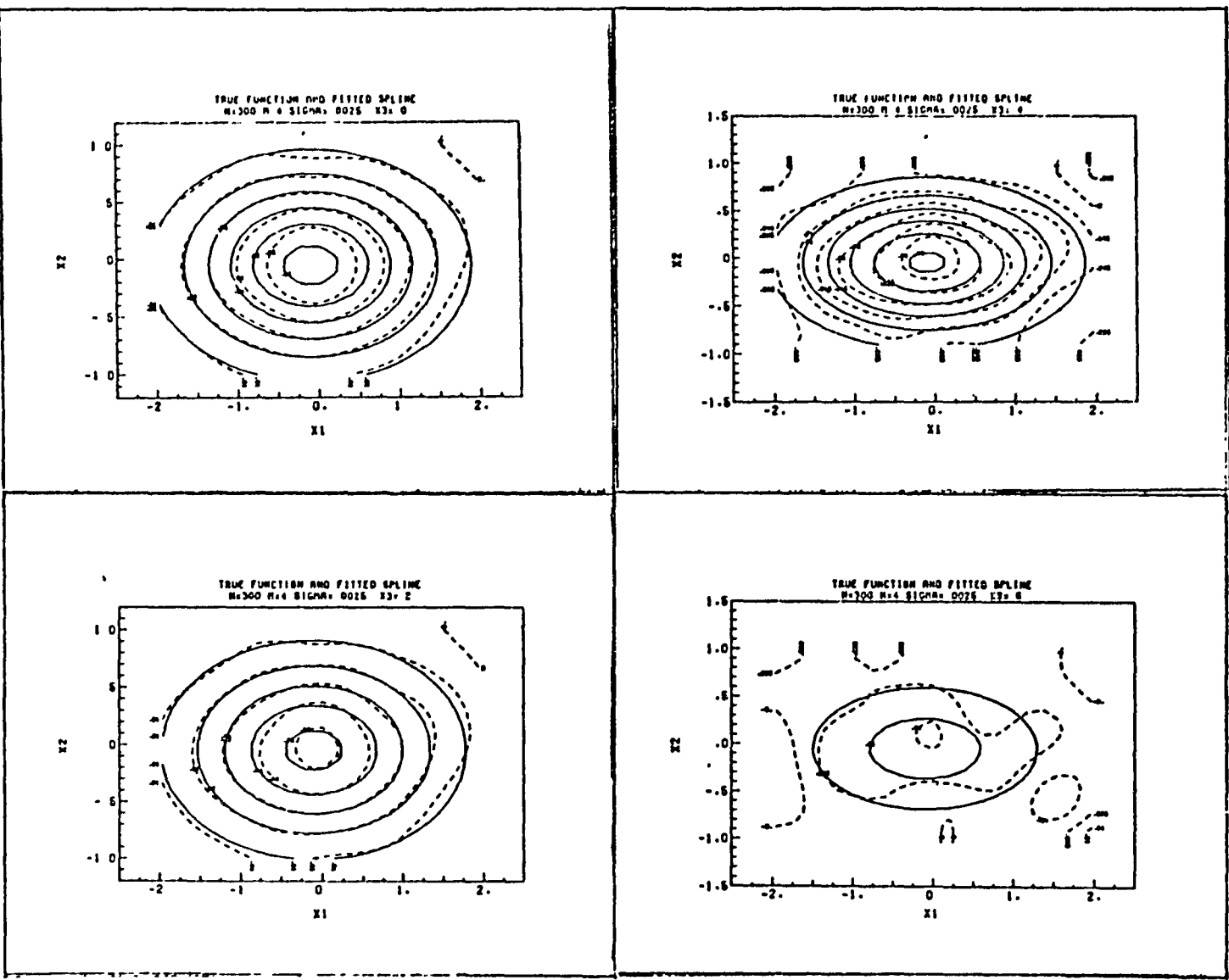


Figure 7.4: Example 3--solid line is f, dashed line is g.

The contours of the derivatives of f and g_{λ}^* with respect to x_1 , x_2 and x_3 are given in Figures 7.5a, 7.5b and 7.5c, respectively. The contours of the second derivatives of f and g_{λ}^* with respect to x_1x_1 , x_1x_2 , x_1x_3 , x_2x_2 , x_2x_3 and x_3x_3 are given in Figures 7.6a, 7.6b, 7.6c, 7.6d, 7.6e, and 7.6f, respectively. The same qualitative behavior is displayed by these derivatives as of the function with the degradation occurring relatively more rapidly as the boundary is approached. Figure 7.6f which is $(\partial^2)/(\partial x_3 \partial x_3)$ of f and g_{λ}^* displays a particularly good fit near the center of R .

LSS's may be utilized to detect outliers in multidimensional noisy data provided that the model of Section 2 is (nearly) appropriate. The model requires that the observations are unbiased, i.e., that $Ez = f$. The errors should be additive and have a known relative error structure, D_{σ} . For the purpose of the outlier study here we shall further assume that each error e_i has a Gaussian distribution.

To what extent the assumption of normality may be relaxed in practice requires further study. The smoothness assumption requires that $f(\underline{t})$ is a smooth function of \underline{t} . This rules out "cliff" functions or those with discontinuities. By using a probability plot of the residuals the example discussed here, which satisfies the above requirements, will be used to demonstrate an outlier detection method.

Data sets with outliers need to be constructed. To accomplish this choose the two points of \underline{t}_i , $i=1, \dots, 300$ which are nearest to and farthest from the origin, which is the center of the data region. These two points are $\underline{t}_k = (-.056, -.032, -.042)$ and $\underline{t}_1 = (1.985, -.879, -.325)$, respectively. To

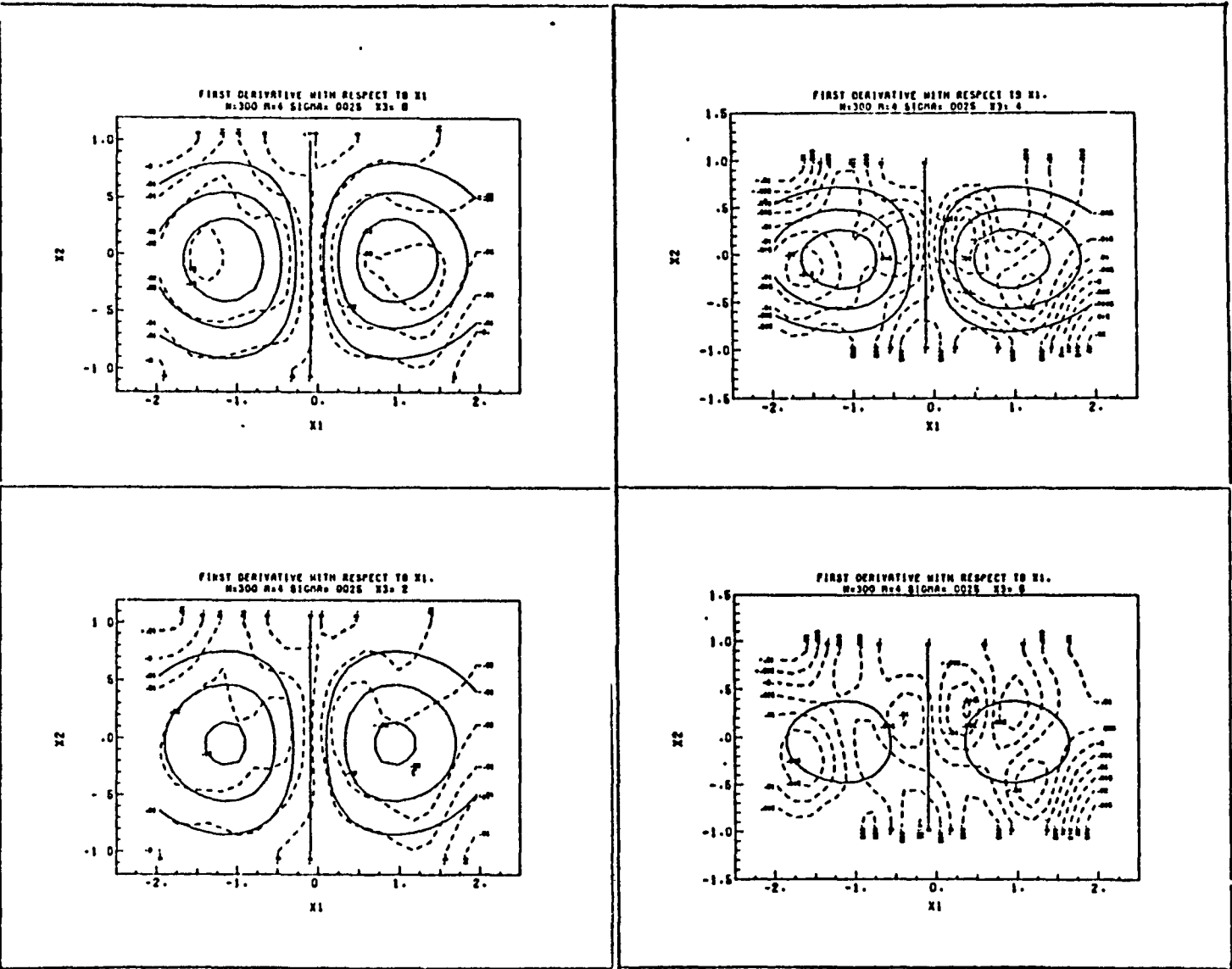


Figure 7.5a: Example 3--Solid line is df/dx_1 , dashed line is dg/dx_1 .

ORIGINAL PAGE IS
OF POOR QUALITY

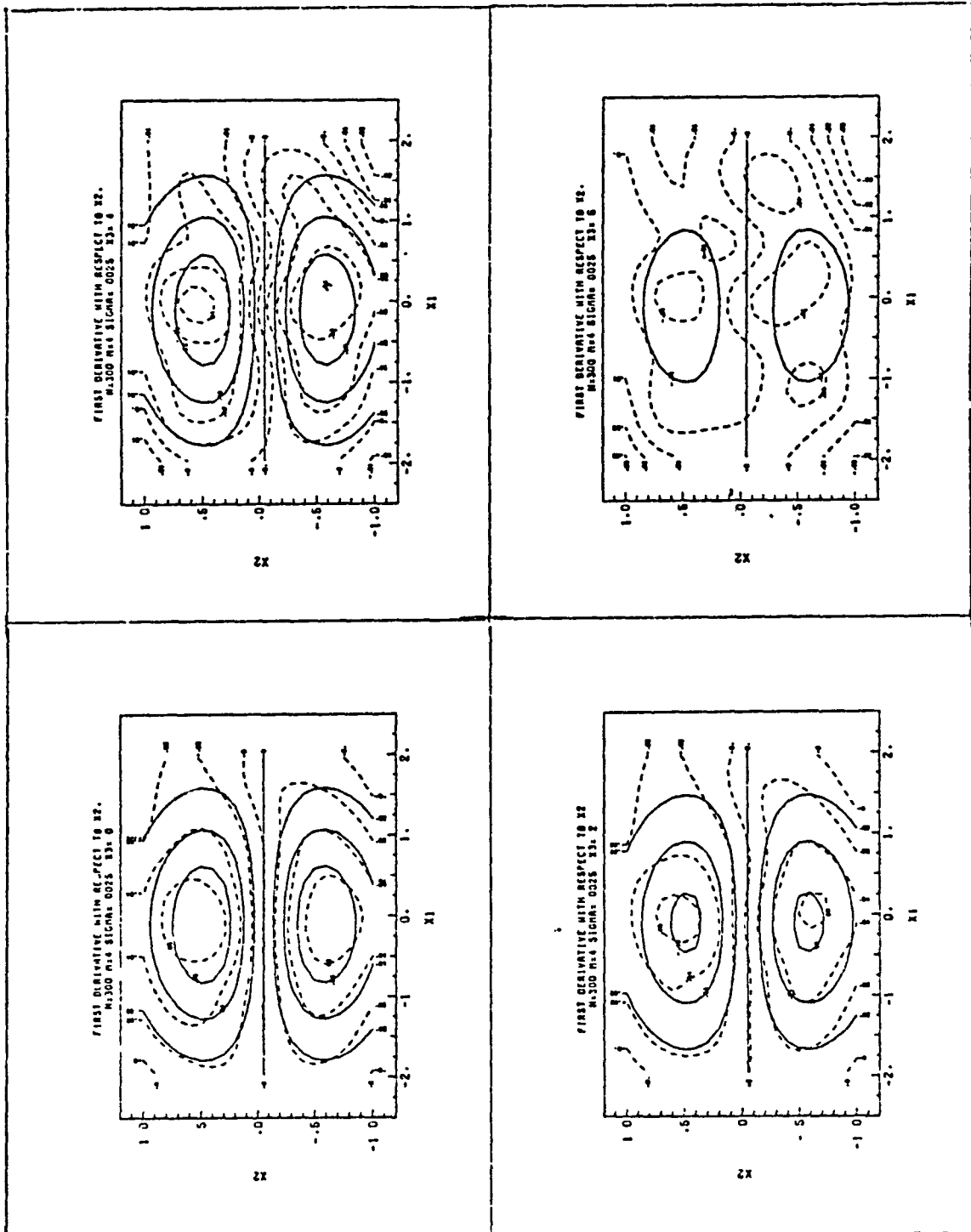


Figure 7.5b: Example 3--solid line is df/dx_2 , dashed line is dg/dx_2 .

ORIGINAL FACE IS
OF POOR QUALITY

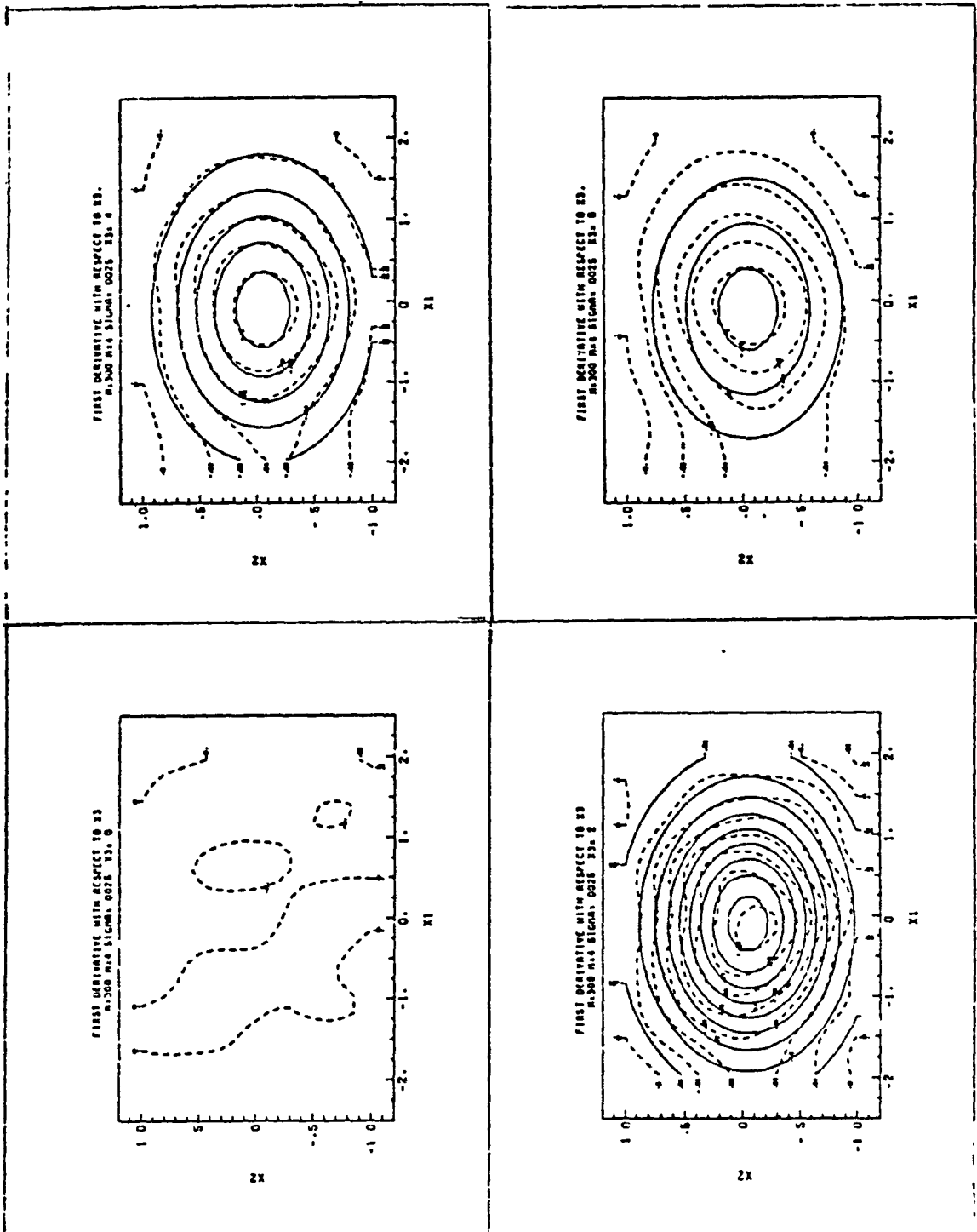


Figure 7.5c: Example 3--solid line is df/dx_3 , dashed line is dg/dx_3 .

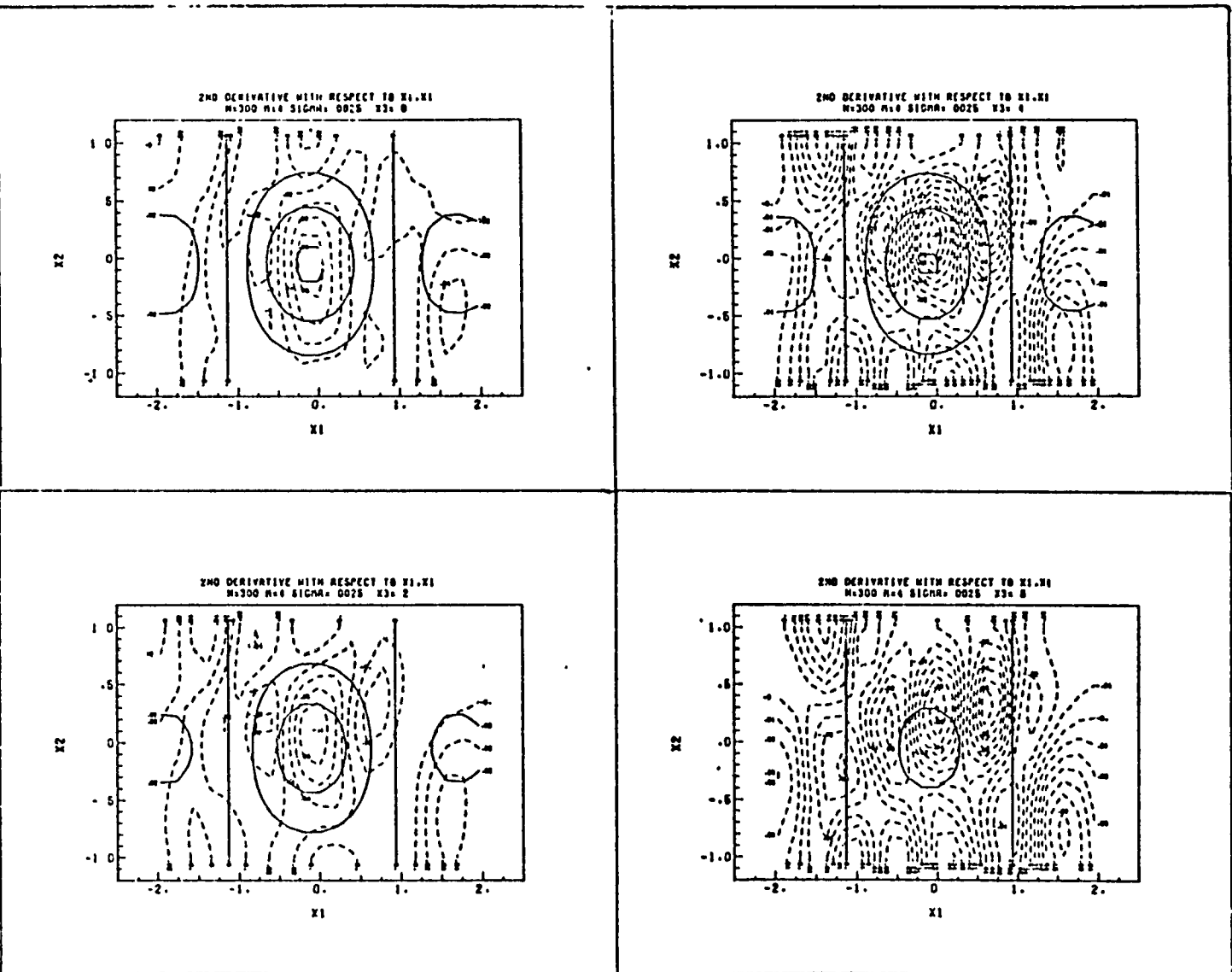
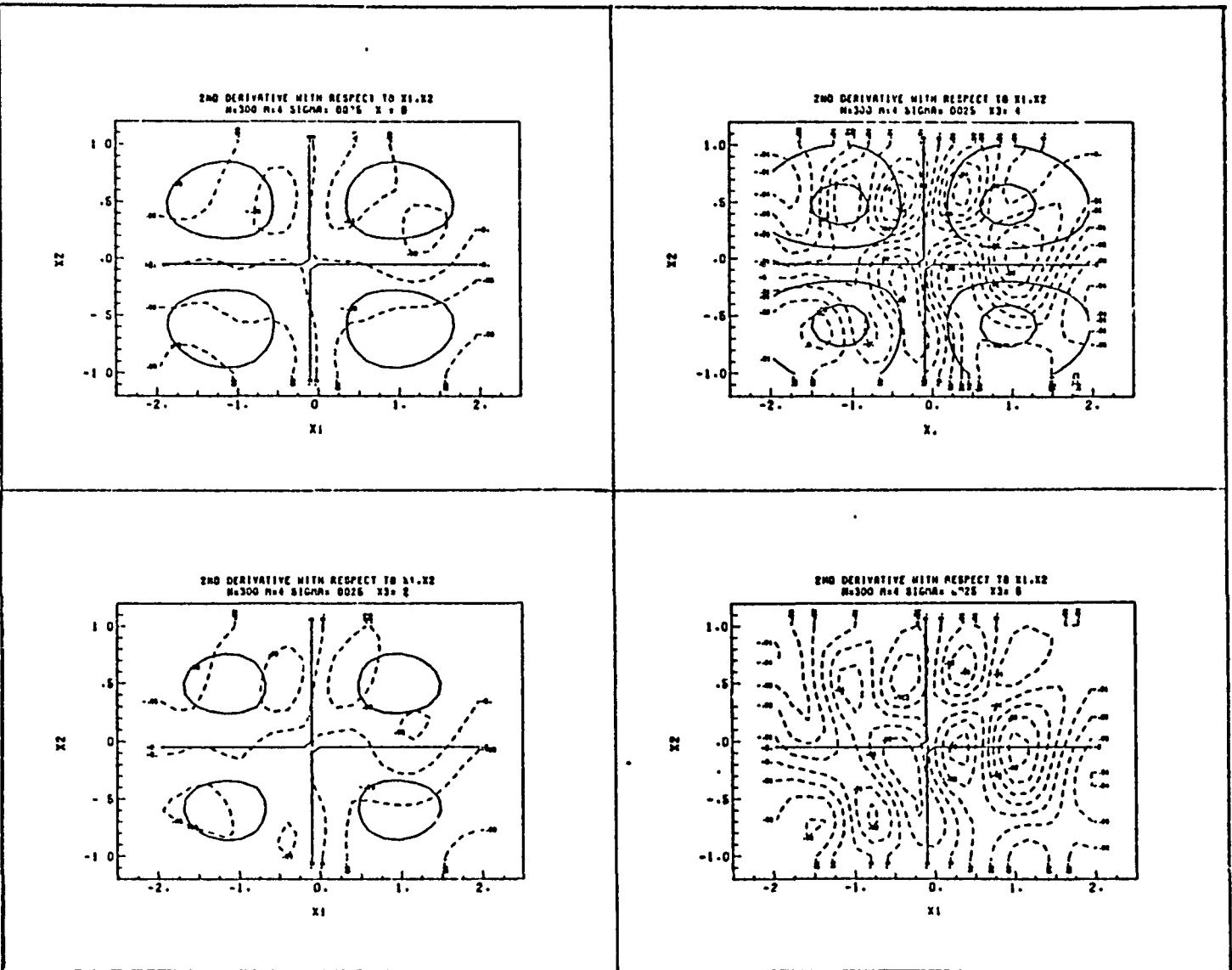


Figure 7.6a: Example 3--solid line is d^2f/dx_1^2 , dashed line is d^2g/dx_1^2 .

Figure 7.6b: Example 3--solid line is d^2f/dx_1dx_2 , dashed line is d^2g/dx_1dx_2 .



ORIGINAL PAGE IS
OF POOR QUALITY

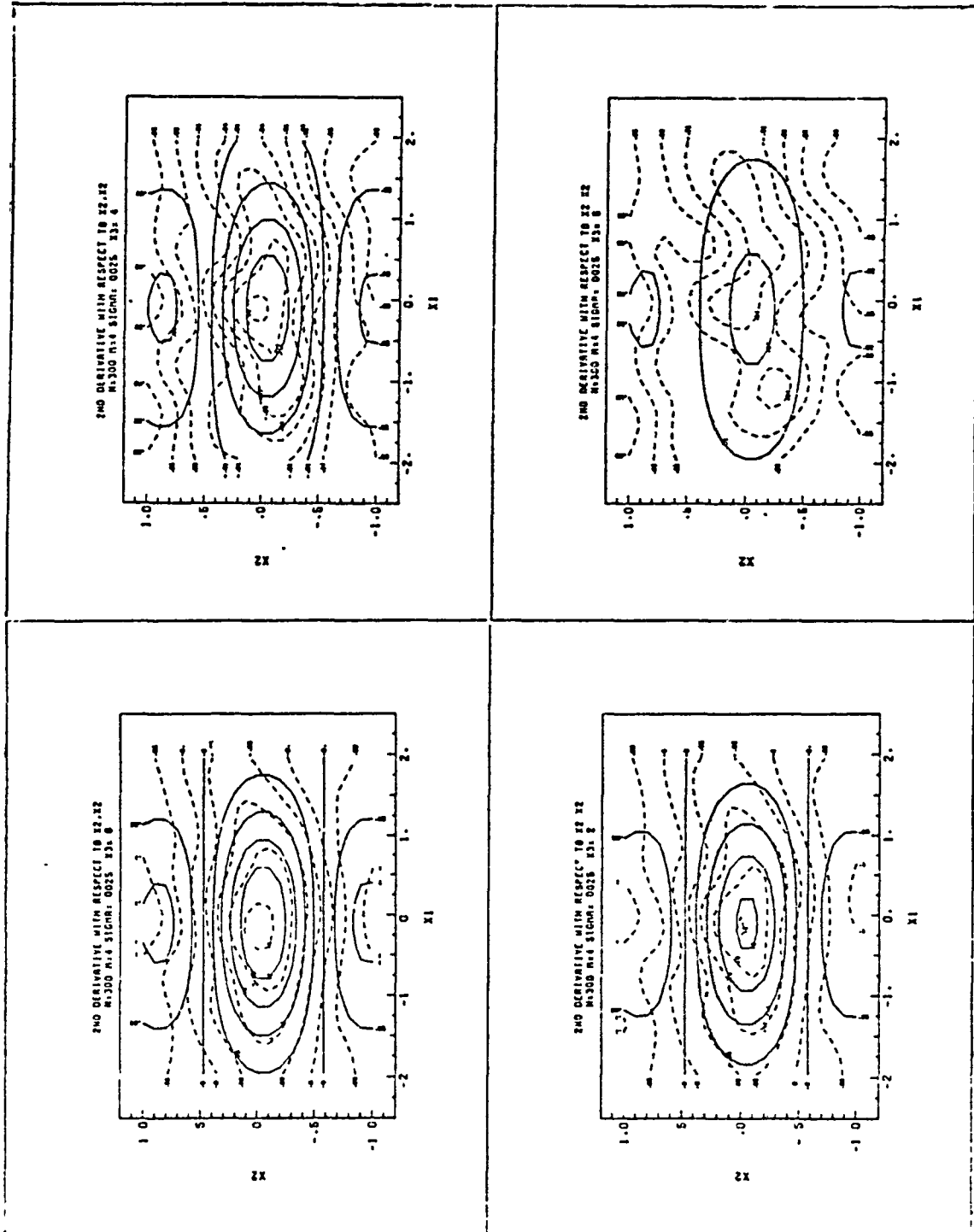


Figure 7.6c: Example 3--solid line is d^2f/dx_2^2 , dashed line is d^2g/dx_2^2 .

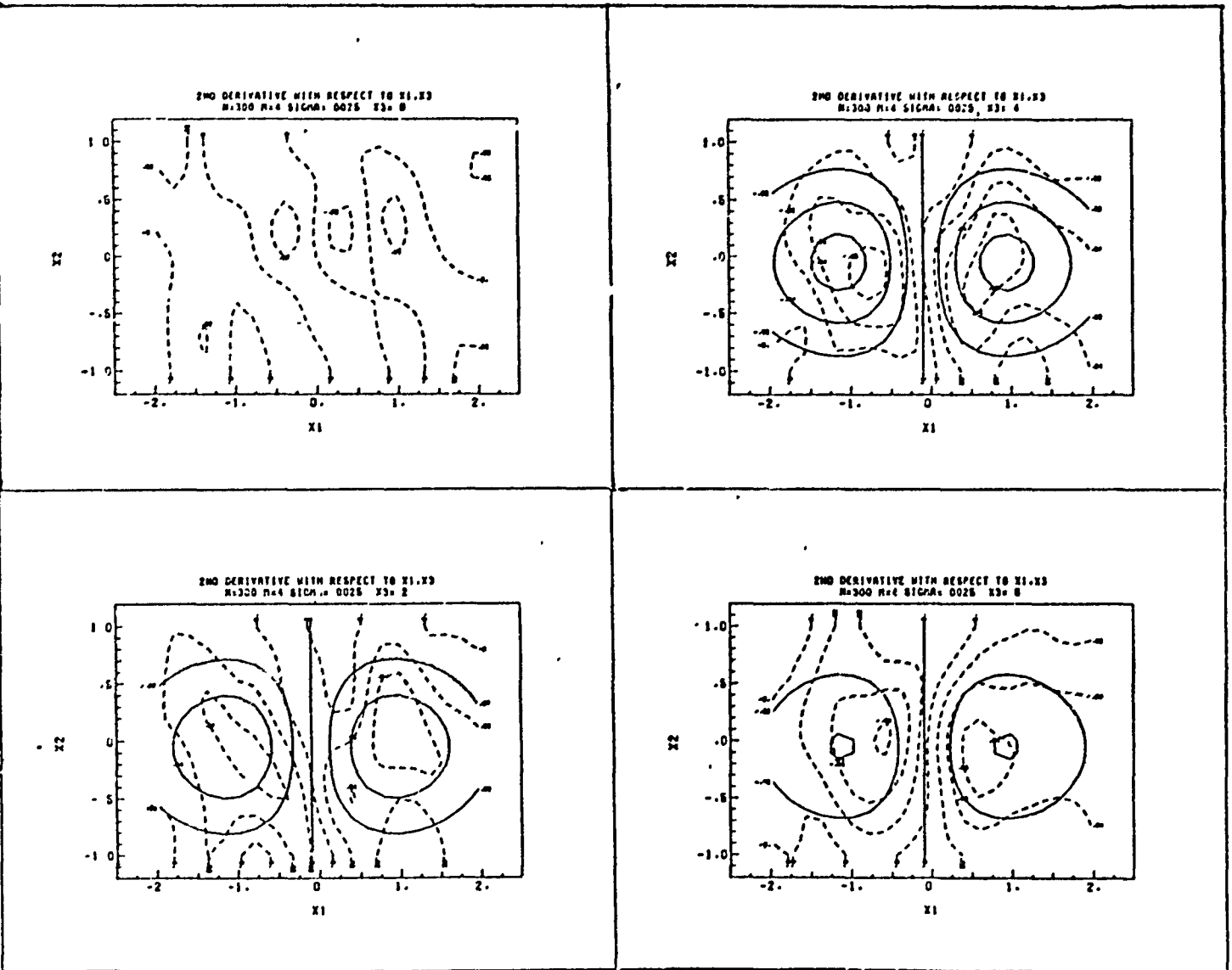


Figure 7.6d: Example 3--solid line is d^2f/dx_1dx_2 , dashed line is d^2g/dx_1dx_2 .

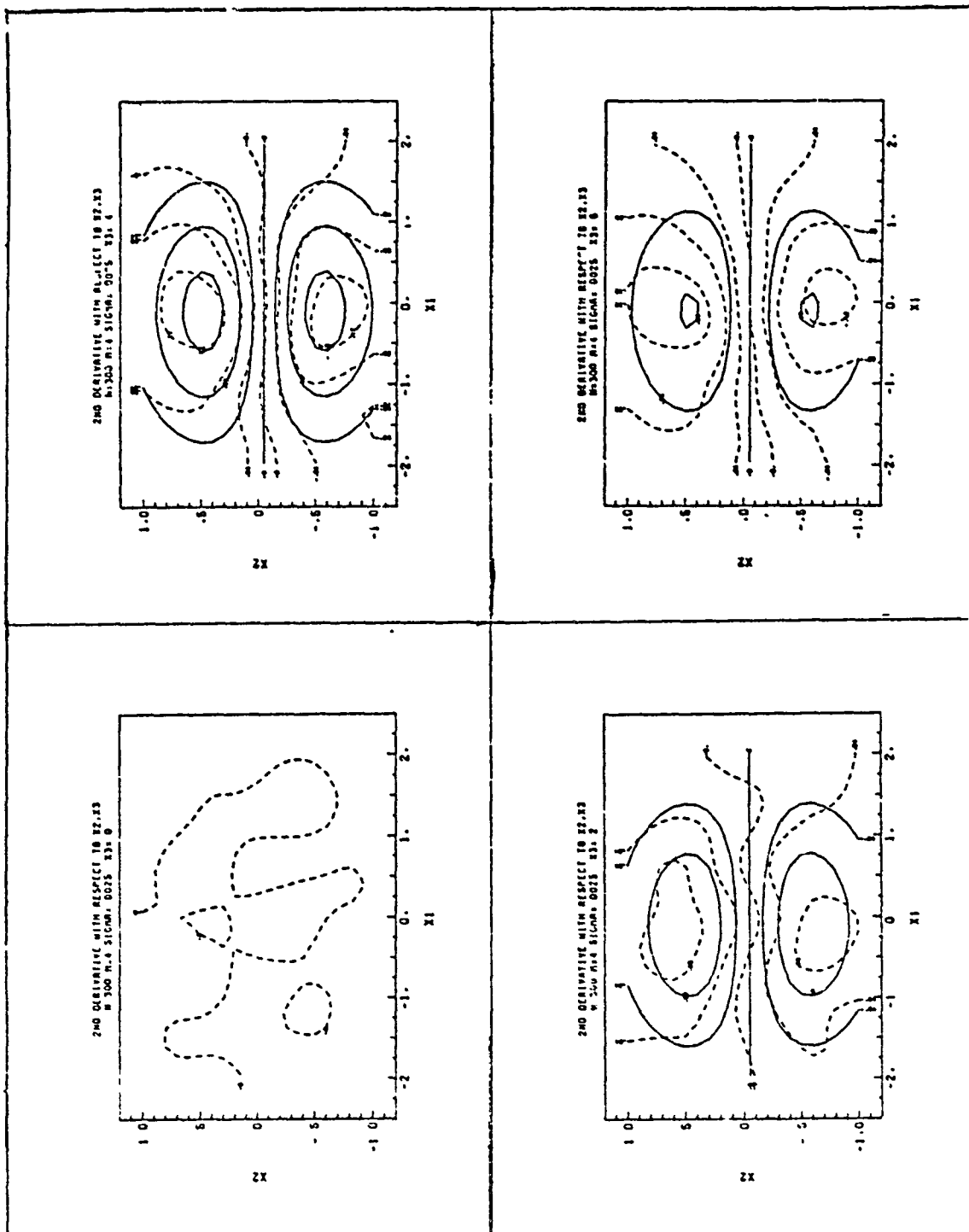


Figure 7.6e: Example 3--solid line is d^2f/dx_2dx_3 , dashed line is d^2g/dx_2dx_3 .

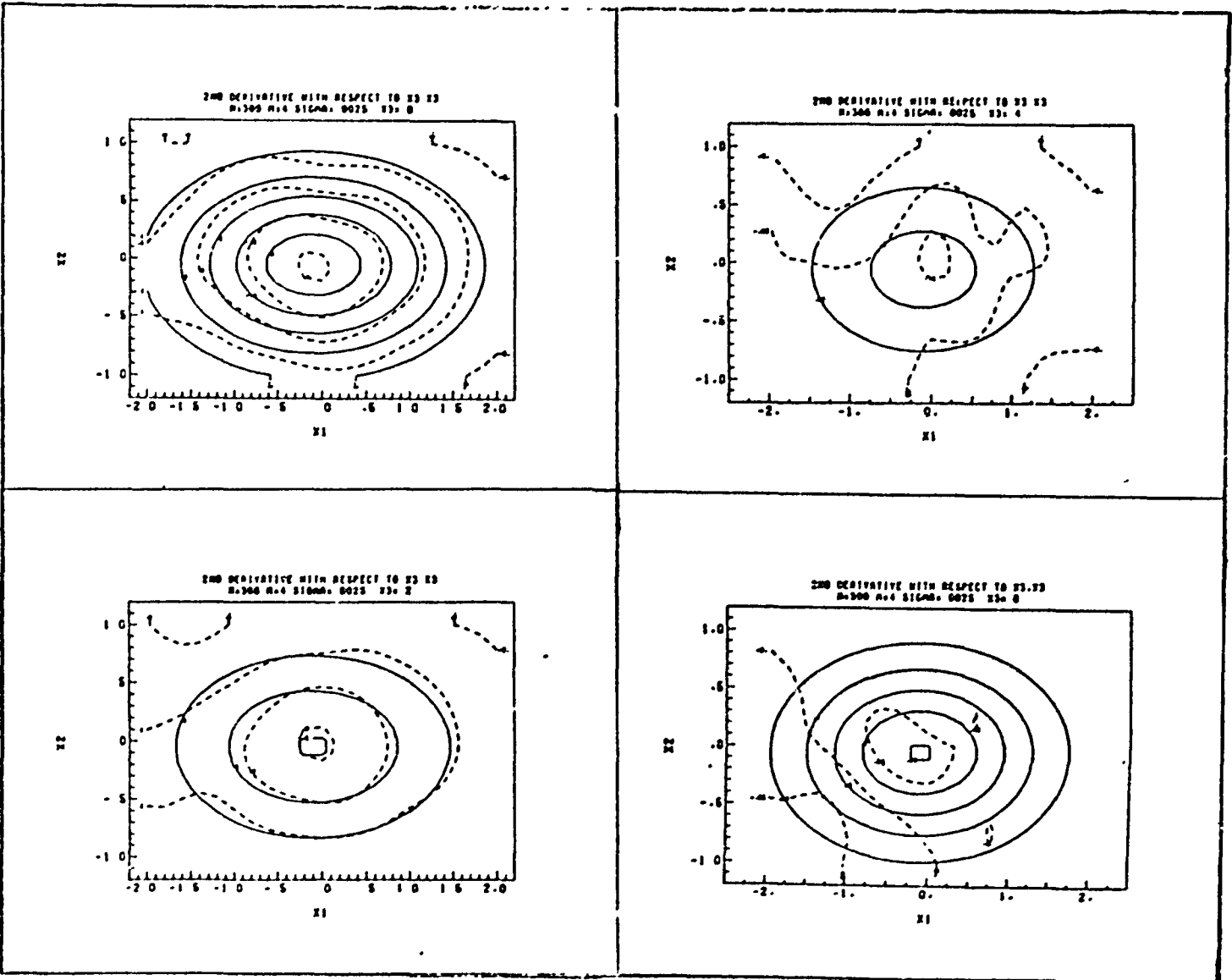


Figure 7.0f: Example 3--solid line is d^2f/dx_3^2 , dashed line is d^2g/dx_3^2 .

construct data sets z_{ks} , let each element of z_{ks} equal the corresponding element of z except for the k^{th} . The k^{th} element is set equal to $f(t_k) - s\sigma$, $\sigma = .0025$. Construct z_{ls} analogously except that the l^{th} element becomes $f(t_l) + s\sigma$.

With the data sets z_{ks} and z_{ls} probability plots in Figures 7.7 and 7.8 were obtained with MINITAB, Ryan, Joiner and Ryan (1976). The probability plot is constructed by ordering the residuals r_i from smallest to largest and plotting them against their corresponding normal scores. The i^{th} smallest normal score as used by MINITAB is the $(i-3/8)/300.25$ percentage point of the normal or Gaussian distribution. If the error distribution that is postulated in the model is the correct one, then the probability plot should be nearly linear. In the data sets constructed here the error distribution is not correct because the k^{th} or l^{th} point is biased and contains no random component.

The numbers in Figures 7.7 and 7.8 indicate how many points are plotted at that spot on the graph. An asterisk indicates one point and a plus sign indicates that more than 9 points are overlapping. In Figures 7.7b, c and d the outlier is identified as the point which is separate from the points which form the line. As the assumption of unbiasedness is more strongly violated it shows up more obviously in the plot.

Figures 7.8a-d demonstrate that this outlier detection scheme is not invincible and should be used in conjunction with other diagnostic checks. The point t_l has very high leverage because it is on the boundary of the data region. In linear regression this is analogous to the points at the extremes

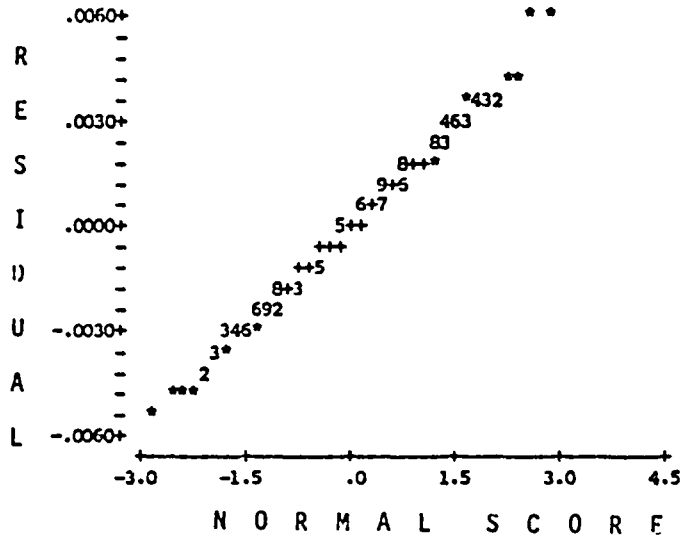


Figure 7.7a: Residuals vs. normal scores for one outlier, $f(t_k) - 0\sigma$, at t_k .

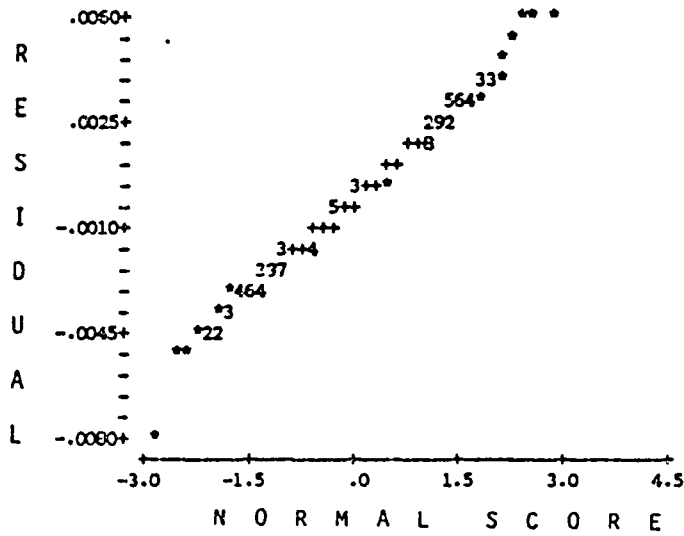


Figure 7.7b: Residuals vs. normal scores for one outlier, $f(t_k) - 6\sigma$, at t_k .

ORIGINAL PAGE IS
OF POOR QUALITY

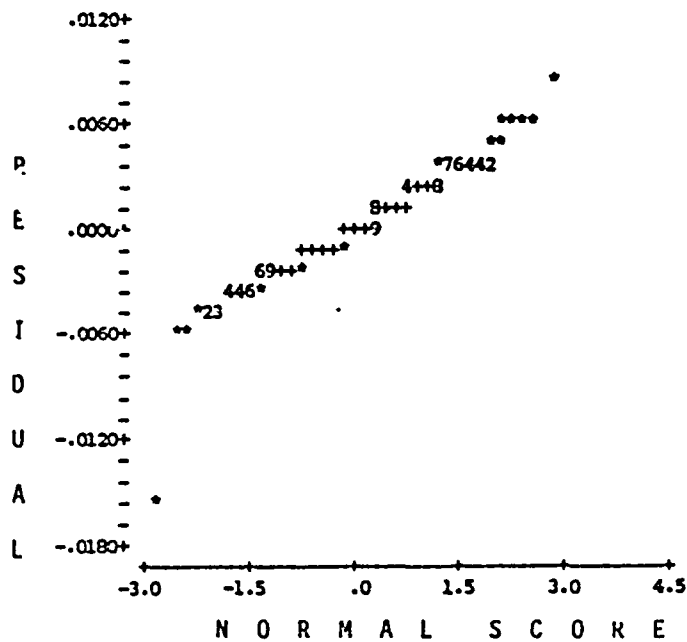


Figure 7.7c: Residuals vs. normal scores for one outlier, $f(t_k) - 10\sigma$, at t_k .

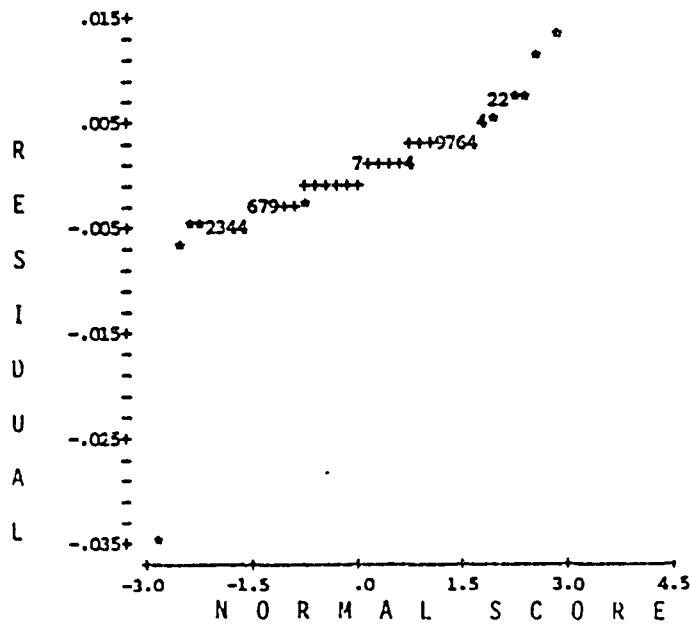


Figure 7.7d: Residuals vs. normal scores for one outlier, $f(t_k) - 20\sigma$, at t_k .

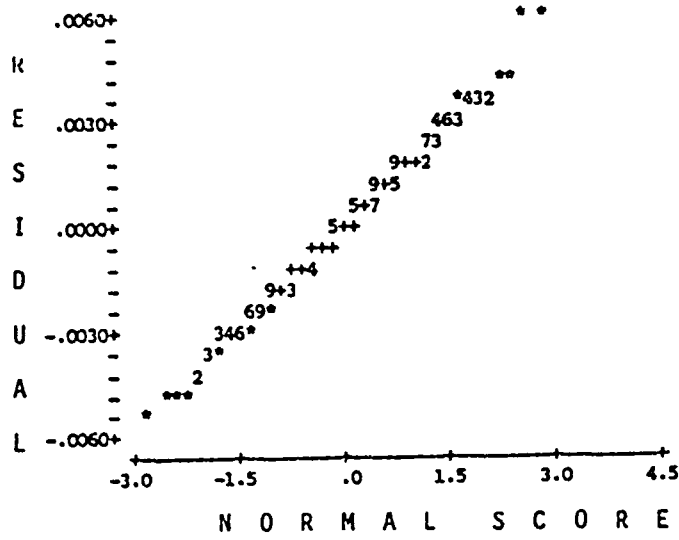


Figure 7.8a: Residuals vs. normal scores for one outlier, $f(t_1) + 0\sigma$, at t_1 .

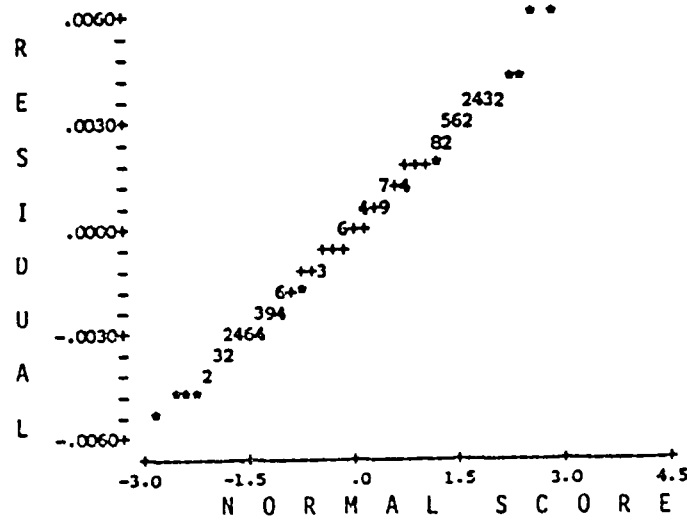


Figure 7.8b: Residuals vs. normal scores for one outlier, $f(t_1) + 6\sigma$, at t_1 .

ORIGINAL PAGE IS
OF POOR QUALITY.

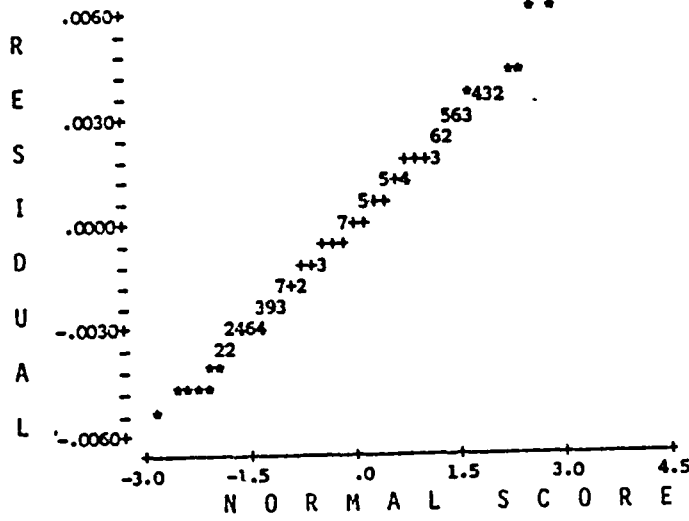


Figure 7.8c: Residuals vs. normal scores for one outlier, $f(t_1) + 10\sigma$, at t_1 .

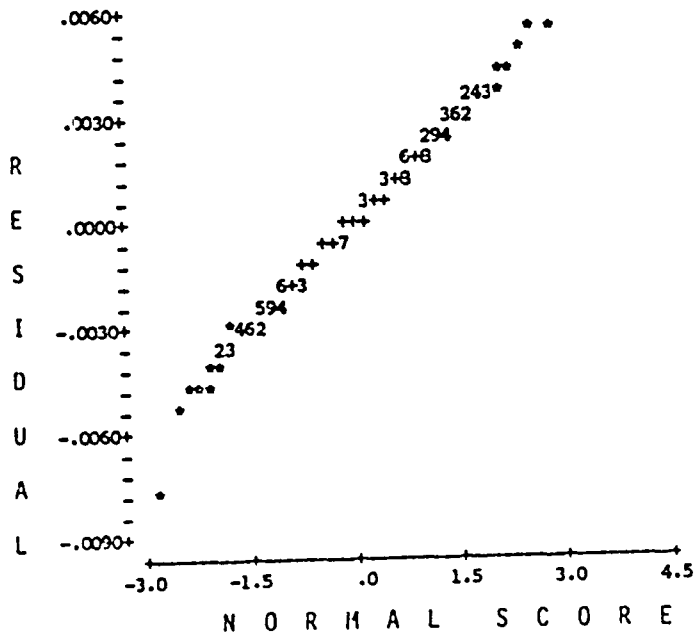


Figure 7.8d: Residuals vs. normal scores for one outlier, $f(t_1) + 20\sigma$, at t_1 .

of the independent variable range which also have high leverage. Because of this the residual at t_j is not large and does not show up in the probability plots of Figures 7.8a-d. The leverage at t_j is so large that it causes another point, the one in the lower left, in Figure 7.8d to appear as discrepant. The probability plot provides a technique to check model assumptions. However, as demonstrated here, this technique should be used in conjunction with other diagnostic checks and with a good understanding of the pitfalls which may be encountered.

Another diagnostic check which may be employed here is to plot the residuals, r_j , against the distance from t_j to t_1 . This is analogous to plotting the residuals against the independent variable in simple linear regression. If a nonrandom pattern is observed, such as serial correlation, then we have evidence that some model assumption is being violated. In practice, t_1 is unknown and hence it may be necessary to do all possible plots, $l=1, \dots, N$.

If a scaling D_σ had been used then the scaled residuals $D_\sigma^{-1}r$ would be plotted instead of r .

The procedure described here is a diagnostic method by which some of the model assumptions may be checked. Irregularly spaced multidimensional "noisy" data easily mask outliers. This technique provides a means which may detect these discrepant observations. It is presented here in the hope that it becomes a routine method to check for model violations in an analysis which uses LSS's.

The three dimensional results presented here are new and quite promising. A quantitative measurement of the goodness of fit of the estimated spline and its derivatives to the true function is given in Wendelberger (1981). Further Monte Carlo experiments will be performed in 3 and more dimensions.

8. Running the program

To evaluate an LSS at any point, $t \in R^d$ involves the execution of two computer programs. The first of these, called MAIN, produces the coefficients of the spline. The second, called EVALUATE, produces the spline, $g_{N,m,\lambda}(t)$. If $2m-2k-d$ is positive EVALUATE may also be used to produce the first ($k=1$) or second ($k=2$) derivative of $g_{N,m,\lambda}$. Depending upon the particular problem at hand the user specifies different options to be exercised by the program. These options will be explained card by card below. Card 1 will be abbreviated C1 and the commands are summarized in Table 8.1 with an example runstream given in Table 8.3.

C1 is used to specify whether or not the coefficient arrays c and d and the matrices X and P used to reconstruct the spline are written to unit 13. X contains the values of the independent variables and P contains the exponents of the polynomials in (2.5), where P is rigorously defined.

To accomplish storing the spline in unit 13 C1 should have SS13 in columns 1 through 4. If EVALUATE is not going to be run then the contents of unit 13 will be unused. In this case C1 should be DONT.

Someone other than the casual user may require other arrays and matrices which are also written to unit 13. See subroutine WRT13 in Wendelberger (1981) for details on the arrays and matrices which are written to unit 13. C2, to be described in the next paragraph, writes into unit 14. See subroutines AWRT14 and BWRT14 to determine the specific values which are written to unit 14.

TABLE 8.1

Input for MAIN

CARD	POSSIBLE VALUES	FORMAT
1	SS13, DONT	A4
2	SM14, UM14, DONT	A4
3	SR15, SP15, VL15, DONT	A4
4	MGCV, USEL	A4
4+	(λ)(Insert if C2 is USEL.)	E15.8
5	VARI, STAN, SAME (Omit if C2 is UM14.)	A4
6	(d,N,m) (Omit if C2 is UM14.)	3I5
7	Format of cards C8+1,...,C8+N.	18A4
8+1	(z_1, t_1^T, σ_1 or σ_1^2)	
.	(z_i) (If C2 is not UM14.)	
.	(z_i, t_i^T, σ_i or σ_i^2) (If C5 is STAN or VARI.)	
.	Format is provided on C7.	(See C7)
8+N	(z_N, t_N^T, σ_N or σ_N^2)	
9	YES, NO	A4

C2 provides the ability to store certain matrices in unit 14 by using SM14 in columns 1 through 4. The storage of these matrices makes it unnecessary to perform the bulk of the computations if a second analysis is to be performed. However, only the dependent variables may be changed for such a subsequent analysis. The relative variances or standard deviations must be identical to the run which used SM14 on C2.

UM14 in the first four columns of C2 provides for use of the matrices which have previously been stored in unit 14. If the value of C2 is DONT then the matrices are neither stored nor used.

C3 provides a means to retrieve certain information during the execution of MAIN and to store this information in unit 15. The first four columns of C3 must be SR15, SP15, VL15 or DONT. If C3 is SR15 the residuals $r_i = (z - g_{N,m,\lambda}(t))$ are stored in unit 15 with the format (G24.18). If C3 is SP15 the ordinate and abscissa for each point of the plot of the GCVF as given in the output are stored. First the number (n) of pairs is stored in I5 format followed by the ordered pairs $(i, \ln(V(1)^{a1+b}))$, where i is an index number $i=1, \dots, n$ and \ln is the natural logarithm; the format used is (I3,G24.18). If C3 is VL15 then b_i/N , $i=1, \dots, N-M$ with format (G24.18) followed by w with with the same format are stored. If none of the above are to be stored then C3 should be DONT.

The value of MGCV on C4 causes the GCVF to be minimized to determine λ^* . If the user wants to supply a value of λ then the value of C4 should be USEL. In that case C4+ is used. C4+ should contain the value of λ in (E15.8) format to be stored in a single precision variable. If C4 is MGCV then C4+ should not be included in the input stream.

C5 is not used if the value of C2 is UM14. Otherwise C5 is used to input relative variances or relative standard deviations or neither of these for the errors of the dependent variable. If the relative variances are to be read then C5 should be VARI; if the relative standard deviations are read then C5 is STAN; and if neither is read then C5 is SAME. The value SAME is equivalent to that of entering all 1's as the relative variances. However, if SAME is used then the program circumvents both multiplication and division by 1 since D_{σ} is simply the identity matrix.

C6 is not used if C2 is UM14. Otherwise C6 reads in the number of independent variables (= dimension), the number of observations N and the value of m to be used. The format used is (3I5).

C7 contains the format to be used to read in the data values. The format should require at most 72 spaces including the left- and right-most parentheses.

The data follow in C8+1 through C8+N. The data should be real Fortran variables, each data line should contain, in order, the dependent variable, the independent variable(s) and the relative variance or standard deviation if used. If C2 is UM14 then C8+1 through C8+N should contain only the dependent variables. They should be given in the identical sequence as the dependent and independent variable(s) were when C2 had the value SM14.

The last card to be read is C9. It should contain one of the values YES or NO. If YES then experimental confidence intervals are provided along with degrees of freedom and an estimate of the variance (Wahba, (1981)). If NO then these values are neither computed nor printed.

To evaluate the spline ($k = 0$), first derivative ($k = 1$) or second derivative ($k = 2$) the program EVALUATE is used. Previous to running EVALUATE the program MAIN must have been run with C1 writing the coefficients to unit 13 (C1 must have been SS13). EVALUATE will then read the matrices from unit 13 and calculate the spline, its first derivative or second derivative. The k th derivative ($k = 1$ or $k = 2$) will be calculated only if $2m-2k-d$ is greater than 0. A description of the input stream for EVALUATE is given in Table 8.2 with a sample runstream given in Table 8.3.

C1 contains two integer values in (2I5) format. The first integer, N', specifies the number of points $t \in R^d$ at which the function is to be evaluated. The second integer should be one of 0, 1 or 2 depending upon whether the spline, first or second derivative, respectively, is to be calculated.

The second card contains the format to be used to read in the N' points. The format should require at most 72 spaces, including the left- and right-most parentheses. The independent variables are read line by line in the same sequence as that which was used to calculate the coefficients.

C3 must be either SV15 or DONT. To store the values in unit 15, C3 should be SV15. This causes the values followed by the corresponding independent variable(s) to be written to unit 15. If C3 is DONT then the values are not written to unit 15.

C3+ is used only if C3 is SV15. Then C3+ should have the format which is to be used to write the calculated value(s) followed by the independent variable(s) into unit 15. This format may have at most 72 spaces including both the left- and right-most parentheses.

TABLE 8.2

Input for EVALUATION

CARD	POSSIBLE VALUES	FORMAT
1	(N', k)	2I5
2	Format to read $C_{4+1}, \dots, C_{4+N'}$.	18A4
3	SV15,DONT	A4
3+	Format for 15 (Omit if C2 is DONT.)	18A4
4+1		
.	Independent variable points	
.	of evaluation, t^T .	(See C2)
.	Format is provided in C2.	
4+N'		

(-6)

TABLE 8.3

Sample Runstreams	Comments
<pre>@XQT SMOOTH*SPLINE.MAIN SS13 SM14 DONT MGCV SAME 1 24 2 (F3.10,33X,F4.0) @ADD DATA. YES</pre>	<pre>Implements the MAIN program. Stores the spline coefficients in unit 13. Stores matrices in unit 14. Doesn't store other values. Minimize the GCVF to determine λ^*. The relative variances are all the same. One dimension, 24 observations, m=2. Format of the input data. Inserts data from Table 3.1 in runstream. Provide confidence intervals.</pre>
<pre>@XQT SMOOTH*SPLINE.EVALUATE 200 0 (36X,F8.4) SV15 (2E15.8) @ADD PLOTDATA.</pre>	<pre>Implements the EVALUATION program. At 200 points evaluate the spline. Format of the independent variables. Store the spline and independent variable values in unit 15. Format of above. Inserts abscissa points to be used for plotting.</pre>
<pre>@XQT SMOOTH*SPLINE.MAIN SS13 UM14 DONT USEL .00016E00 (F3.0) @ADD DATA. YES</pre>	<pre>Implements the MAIN program. Stores the spline coefficients in unit 13. Uses the matrices stored in 14 by MAIN above. Doesn't store other values. Use the following value of λ. Value of λ to be used. Format of the dependent variables. Inserts data from Table 3.1. Provides confidence intervals.</pre>
<pre>@XQT SMOOTH*SPLINE.EVALUATE 200 0 (36X,F8.4) SV15 (2E15.8) @ADD PLOTDATA.</pre>	<pre>Implements the evaluation program. At 200 points evaluate the spline. Format of the independent variable. Store the spline and independent variable in 15. Format of above. Inserts abscissa points to be used for plotting.</pre>

C4+1 through C4+N' contain the independent variable(s) at which the function is to be evaluated. These should be in the format given on C2. The independent variable(s) should be in the same sequence as used to obtain the coefficients with the program MAIN.

The programs MAIN and EVALUATE are written in ASCII FORTRAN Level 9R1 and are running on the UNIVAC 1100/80 computer at the University of Wisconsin. All calculations are performed in double precision.

The subroutines used by the programs MAIN and EVALUATE are named: AWRT14, BWRT14, CALC, CALD, CALRES, CHECKQ, COLOFK, CONINT, DATAR, DERIV1, DERIV2, E, ED1, ED2, GETASI, GETBM, GETR, GETRDE, GETTHM, GRAPHV, MAKEB, MAKETS, MINVL1, MINVL2, MQRDC, PRINT, PRNTLM, RCHECK, READ13, SPLINE, SVDB, VARDF, VLHELP, VOFL, WHATDO, WRT13, AND WRT15. GRAPHV, MINVL1 and MINVL2 are modeled after similar subroutines of the one dimensional smoothing spline program written by Fleisher (1979) and running at the Madison Academic Computing Center (MACC). A description of the program structure is given in Wendelberger (1981).

The following LINPACK subroutines are also used by the program MAIN: DAXPY, DCOPY, DDOT, DNRM2, DQRDC, DQRSL, DROT, DROTG, DSCAL, DSVDC, DSWAP and DTRSL. The code for these routines is not included here. It may be found in the LINPACK USERS' GUIDE by Dongarra, Bunch, Moler and Stewart (1979). One modification is made in the LINPACK subroutine DSVDC: the parameter MAXIT is increased from 30 to 60. This parameter sets the maximum number of iterations to be performed in the algorithm to determine the singular values and vectors of B before termination due to nonconvergence. Increasing MAXIT to 60 is

necessary because with large N , say $N > 140$, 30 iterations may not be large enough for some problems. An example with $N=150$ failed because $MAXIT=30$ was too small. However, with $MAXIT=60$ example 3 with $N=300$ was successfully run. In fact $MAXIT=60$ has proved ample for all examples tried to date. The version of the program described here uses the singular value decomposition to obtain the spectral decomposition of B . A new modified version uses the EISPACK (Smith, et al., (1976)) routines DTRED2 and DTQL2 to accomplish this task at a much reduced cost and at no loss in accuracy. This is because the singular value decomposition does not make use of the symmetry of B . The EISPACK routines do make use of the symmetry of B and thus the cost of the decomposition is roughly cut in half.

ACKNOWLEDGMENTS

The author wishes to express thanks to Grace Wahba for introducing this problem and providing guidance and encouragement, to Gene Golub who inspired this algorithm and to the attendees of both the First and Second SIAM Summer Research Conference on Numerical and Statistical Analysis from whose comments this work has benefited.

Thanks go to Alison Pollack for her comments on earlier drafts of this report and to the students of Statistics 840, fall semester 1980, who endured the earlier versions of the program.

This research was sponsored by the National Aeronautics and Space Administration under NASA Grant No. NAG5-128 and by the Office of Naval Research under Contract No. N00014-77-C-0675.

ORIGINAL PAGE IS
OF POOR QUALITY

REFERENCES

- Chatterjee, S. and Price, B., 1977: Regression analysis by example. John Wiley and Sons, 10-18.
- Craven, P. and G. Wahba, 1979: Smoothing noisy data with spline functions: Estimating the correct degree of smoothing by the method of generalized cross-validation. Numer. Math., 31, 377-403.
- Dongarra, J. J., J. R. Bunch, C. B. Moler, and G. W. Stewart, 1979: Linpack users' guide. Society for Industrial and Applied Mathematics, Philadelphia, Pennsylvania.
- Duchon, Jean, 1976: Interpolation des fonctions de deux variables suivant le principe de la flexion des plaques minces. R.A.I.O. Analyse numerique, Vol. 10, No. 12, 5-12.
- Fleisher, J., 1979: Spline smoothing routines. Reference manual for the 1110. Academic Computing Center, University of Wisconsin, Madison.
- Franke, R., 1979: A critical comparison of some methods for interpolation of scattered data, Naval Postgraduate School Report No. NPS-53-79-003, March, 1979.
- Golub, G. H. and C. Reinch, 1970: Singular value decomposition and least squares solutions. Numer. Math. 14, 403-420.
- Harder, R. L., and R. N. Desmarais, 1972: Interpolation using surface splines. J. Aircraft, Vol. 9, No. 2, 189-191.
- Lucas, H. A., 1978: Estimation of smoothing parameters to smooth noisy data and confidence regions for the underlying function. Ph.D. thesis, Dept. of Statistics, University of Wisconsin, 136 pp.
- MACC, 1978: Random Number Routines. Reference Manual for the 1110. Madison Academic Computing Center, University of Wisconsin-Madison.
- Mosteller, F. and J. W. Tukey, 1968: Data analysis including statistics, Handbook of Social Psychology, Vol. 2, Reading, MA, Addison-Wesley, 90-203.
- Prenter, P., 1975: Splines and variational methods, John Wiley and Sons, Inc.
- Reinsch, G., 1967: Smoothing by spline functions. Numer. Math., 10, 177-183.
- Ryan, T. A., B. L. Joiner and B. F. Ryan, 1976: MINITAB student handbook, Duxbury Press.
- Schoenberg, I. J., 1964: Spline functions and the problem of graduation. Proc. of Nat. Acad. of Sciences. 52, No. 4, 947-950.

- Smith, B.T., J. M. Boyle, J. J. Dongarra, B. S. Garbow, Y. Ikebe, V. C. Klema, and C. B. Moler, 1976: Matrix Eigensystem Routines--EISPACK Guide, Vol. 6. Lecture notes in Computer Science, Springer-Verlag, 551 pp.
- Wahba, G., 1979: How to smooth curves and surfaces with splines and cross-validation, Tech Report #555, Dept. of Statistics, University of Wisconsin.
- Wahba, G., 1980: Ill posed problems: Numerical and statistical methods for mildly, moderately and severely ill posed problems with noisy data. Tech Report #595, Dept. of Statistics, University of Wisconsin.
- Wahba, G., 1981: Bayesian confidence intervals for the cross validated smoothing spline. Tech. Report No. 645, Dept. of Statistics, University of Wisconsin.
- Wahba, G. and J. Wendelberger, 1980: Some new mathematical methods for variational objective analysis using splines and cross validation. Mon. Wea. Rev., Vol. 108, 8, 1122-1143.
- Wahba, G. and S. Wold, 1975: A completely automatic French curve: Fitting spline functions by cross-validation. Communications in Statistics, 4, 1-17.
- Wendelberger, J. G., 1981: Smoothing noisy data using multivariate splines and generalized cross-validation. Ph.D. thesis to appear, Dept. of Statistics, University of Wisconsin.

Unclassified

474

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

REPORT DOCUMENTATION PAGE		READ INSTRUCTIONS BEFORE COMPLETING FORM
1 REPORT NUMBER Technical Report No. 648	2 GOVT ACCESSION NO.	3 RECIPIENT'S CATALOG NUMBER
4 TITLE (and Subtitle) THE COMPUTATION OF LAPLACIAN SMOOTHING SPLINES WITH EXAMPLES		5. TYPE OF REPORT & PERIOD COVERED Scientific Interim
		6. PERFORMING ORG. REPORT NUMBER
7 AUTHOR(s) James Wendelberger		8 CONTRACT OR GRANT NUMBER(s) No. N00014-77-C-0675
9 PERFORMING ORGANIZATION NAME AND ADDRESS Department of Statistics, Univ. of Wisconsin 1210 W. Dayton St. Madison, WI 53706		10 PROGRAM ELEMENT PROJECT, TASK AREA & WORK UNIT NUMBERS
11 CONTROLLING OFFICE NAME AND ADDRESS Office of Naval Research 800 Quincy St. Arlington, VA		12. REPORT DATE September 1981
		13 NUMBER OF PAGES 68
14 MONITORING AGENCY NAME & ADDRESS (if different from Controlling Office)		15 SECURITY CLASS. (of this report)
		15a. DECLASSIFICATION/DOWNGRADING SCHEDULE
16 DISTRIBUTION STATEMENT (of this Report) Distribution of this report is unlimited.		
17 DISTRIBUTION STATEMENT (of the abstract entered in Block 20, if different from Report)		
18 SUPPLEMENTARY NOTES		
19 KEY WORDS (Continue on reverse side if necessary and identify by block number) Spline smoothing; Computation of splines; multivariate derivative estimation; Spline confidence intervals; Spline diagnostic checks.		
20 ABSTRACT (Continue on reverse side if necessary and identify by block number) Laplacian Smoothing Splines (LSS) are presented as generalizations of graduation, cubic and thin plate splines. The method of generalized cross validation (GCV) to choose the smoothing parameter is described. GCV is used in the algorithm for the computation of LSS's. An outline of a computer program which implements this algorithm is presented along with a description of the use of the program. Examples in one, two and three dimensions demonstrate how to obtain estimates of function values with confidence intervals and estimates of first and second derivatives. Probability plots are used as a diagnostic tool to check for model		

DD FORM 1 JAN 73 1473

EDITION OF 1 NOV 68 IS OBSOLETE
S/N 0102-LF-014-6601

Unclassified

Inadequacy.

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

ORIGINAL PAGE IS
OF POOR QUALITY

OMIT TO
475 END

Reprinted from MONTHLY WEATHER REVIEW, Vol 108, No 8, August 1980
American Meteorological Society
Printed in U S A

**Some New Mathematical Methods for Variational Objective Analysis
Using Splines and Cross Validation**

GRACE WAHBA AND JAMES WENDELBERGER

ORIGINAL PAGE IS
OF POOR QUALITY

Some New Mathematical Methods for Variational Objective Analysis Using Splines and Cross Validation

GRACE WAHBA¹ AND JAMES WENDELBERGER²

Department of Statistics, University of Wisconsin, Madison 53706

(Manuscript received 12 November 1979, in final form 9 April 1980)

ABSTRACT

Let $\Phi(x, y, p, t)$ be a meteorological field of interest (say, height, temperature, or a component of the wind field, etc.). We suppose that data $\{\Phi\}$, concerning the field of the form $\Phi_i = L_i\Phi + \epsilon_i$, are given, where each L_i is an arbitrary continuous linear functional and ϵ_i is a measurement error. The data Φ_i may be the result of theory, direct measurements, remote soundings, or a combination of these. We develop a new mathematical formalism exploiting the method of Generalized Cross Validation (GCV), and some recently developed optimization results, for analyzing this data. The analyzed field $\Phi_{\lambda, m}$ is the solution to the minimization problem: Find Φ in a suitable space of functions to minimize

$$N^{-1} \sum_{i=1}^N (L_i\Phi - \Phi_i)^2 + \lambda J_m(\Phi) \quad (1)$$

where

$$J_m(\Phi) = \sum_{\alpha_1, \alpha_2, \dots, \alpha_m} \frac{m!}{\alpha_1! \alpha_2! \dots \alpha_m!} \left| \left| \left| \left(\frac{\partial^m \Phi}{\partial x^{\alpha_1} \partial y^{\alpha_2} \dots \partial p^{\alpha_m}} \right)^2 dx dy dp dt \right. \right. \right|$$

Functions of $d = 1, 2$, or 3 of the four variables x, y, p, t are also considered. The approach can be used to analyze temperature fields from radiosonde measured temperatures and satellite radiance measurements *simultaneously* to incorporate the geostrophic wind approximation and other information. In a test of the method (for $d = 2$) simulated 500 mb height data were obtained at discrete points corresponding to the U.S. radiosonde network, by using an analytic representation of a 500 mb wave and superimposing realistic random errors. The analytic representation was reconvolved on a fine grid with what appears to be impressive results. An explicit representation for the minimizer of (1) is found and used as the basis for a direct (as opposed to iterative) numerical algorithm, which is accurate and efficient for N , somewhat less than the high-speed storage capacity of the computer. The results extend those of Sasaki and others in several directions. In particular, no starting guesses and no preliminary interpolation of the data is required, and it is not necessary to solve a boundary value problem or even assume boundary conditions to obtain a solution. Different types of data can be combined in a natural way. Prior climatologically estimated covariances are not used. This method may be thought of as a very general form of low-pass filter. The parameter λ controls the half-power point of the implied data filter, while m controls the rate of roll-off of the power spectrum of the analyzed field. From another point of view, λ and m play the roles of the most important free parameters in an (implicit) prior covariance. The correct choice of the parameter λ and to some extent m is important. These parameters are estimated from the data being analyzed by the GCV method. This method estimates λ and m for which the implied data filter has maximum internal predictive capability. This capability is assessed by the GCV method by implicitly leaving out one data point at a time and determining how well the missing datum can be predicted from the remaining data. The numerical algorithm given provides for the efficient calculation of the optimum λ and m .

1. Introduction

Sasaki (1960) introduced the idea of numerical variational analysis for the objective analysis of meteorological fields. In the most general form of variational analysis considered here we seek a func-

tion $\Phi(x, y, p, t)$ of four variables representing a meteorological field of interest (say, height, temperature, or a component of the wind field) as a function of ground projection coordinates (x, y) , the vertical coordinate p , and time t . This function should be suitably close to the height, temperature, or wind field as measured at a finite set of positions, pressures, and times, it should reflect known behavior of such fields, and it should be "smooth" in some sense.

For an example of known behavior, if we fix p at

¹ Research supported by the Office of Naval Research under Contract N00014-77-C-0675.

² Research supported by the National Science Foundation under Grant ATM75-23223.

500 mb, then Φ is the 500 mb geopotential height. Letting $\Phi = \Phi(x, y, p, t)$, then the sum of the tendency and horizontal advection

$$\partial\Phi/\partial t + c_x(\partial\Phi/\partial x) + c_y(\partial\Phi/\partial y)$$

should be small, where c_x and c_y are the x and y components of the wind velocity. Sasaki and others have incorporated weak (i.e., approximate) and strong (i.e., exact) constraints involving the tendency, the advection, the geostrophic wind, balance, horizontal momentum, adiabatic energy, and the hydrostatic and continuity equations (Sasaki, 1971; Lewis, 1972; Lewis and Grayson, 1972; Achtemeier, 1975).

Using the sum of the tendency and advection as a weak constraint, Sasaki (1971) suggests finding Φ to minimize

$$J(\Phi) = \iiint_R \left\{ \left[\hat{a}(\Phi - \hat{\Phi})^2 + a \left[\left(\frac{\partial\Phi}{\partial t} + c_x \frac{\partial\Phi}{\partial x} + c_y \frac{\partial\Phi}{\partial y} \right)^2 \right] + \left[a_t \left(\frac{\partial\Phi}{\partial t} \right)^2 + a_x \left(\frac{\partial\Phi}{\partial x} \right)^2 + a_y \left(\frac{\partial\Phi}{\partial y} \right)^2 \right] \right\} dt dx dy, \quad (1.1)$$

where \hat{a} , a , a_t , a_x , a_y are smoothing parameters to be determined, $\hat{\Phi}$ is the observed height field data, c_x and c_y are the (observed) components of wind velocity, and R is the spatial and temporal region of interest. The first term represents the desire that Φ be close to the data, the second that the sum of the tendency and horizontal advection is small and the third, that the function be smooth in x , y and t .

Since Φ , c_x and c_y are only measured at a (relatively sparse) set of irregularly spaced points, Sasaki assumed that the data have been interpolated to a grid sufficiently fine for numerical analytic purposes. After some simplifying assumptions, the Euler equation for the minimizer of (1.1) was obtained by Sasaki (1971) and the minimizer is found to satisfy an elliptic partial differential equation with some boundary conditions. Various authors using this and other constraints (see, e.g., Lewis and Grayson, 1972) have chosen values for the smoothing parameters, and solved the resulting Euler equations numerically to obtain an objectively analyzed field.

In this paper we develop a general mathematical formalism basically embodying Sasaki's approach with the following five modifications:

1) It is not necessary to first interpolate the data to a grid to obtain Φ ; raw data is used directly.

2) The problem of providing or enforcing boundary data is eliminated.

3) The main unknown smoothing parameters are estimated from the data to be analyzed, rather than from historical data or by guesswork.

4) The method provides a technique whereby raw indirect data, such as satellite radiance data, can be combined with direct data such as balloon temperature data in a single analysis procedure. This can be done without preconverting the radiance data to temperatures.

5) Discretization is the last step rather than the first, so this source of error does not propagate through the analysis. This can be important (see Nitta and Hovermale, 1969).

The method to be described avoids the problem of solving partial differential equations numerically. However, it has its own challenging numerical problems which we have been able to solve simply using existing packages for medium sized (but not large) data sets.

To introduce our general method, we begin with the simplest nontrivial example. We fix time as well as pressure and suppose that $\Phi = \Phi(x, y)$ is the 500 mb height at (x, y) at time $t = 0$. We ignore the tendency and advection (second term) in (1.1) and suppose observations $\hat{\Phi}(x_i, y_i) = \hat{\Phi}_i$, $i = 1, 2, \dots, N$, of the 500 mb height at the N stations with coordinates (x_i, y_i) , $i = 1, 2, \dots, N$, are given. We want to obtain a function Φ which is smooth and such that $\Phi(x_i, y_i) \approx \hat{\Phi}_i$, $i = 1, 2, \dots, N$. Consider the minimization of

$$N^{-1} \sum_{i=1}^N [\Phi(x_i, y_i) - \hat{\Phi}_i]^2 + \lambda J_1(\Phi), \quad (1.2)$$

where

$$J_1(\Phi) = \iint \left[\left(\frac{\partial\Phi}{\partial x} \right)^2 + \left(\frac{\partial\Phi}{\partial y} \right)^2 \right] dx dy, \quad (1.3)$$

and λ is given.

If one attempts to minimize (1.2) by, for example, writing the Euler equation one finds that the solution involves a Green's function for the Laplacian operator Δ , $\Delta\Phi = \partial^2\Phi/\partial x^2 + \partial^2\Phi/\partial y^2$, and, unfortunately, this Green's function is not bounded. Sasaki (1971) observes a similar phenomenon [see paragraph which includes Eq. (32)] but ignores it. For this and other reasons to be discussed, we seek to find the minimizer (in a suitable space of functions) of

$$N^{-1} \sum_{i=1}^N [\Phi(x_i, y_i) - \hat{\Phi}_i]^2 + \lambda J_m(\Phi) \quad m = 2, 3, \dots \quad (1.4)$$

where

$$J_2(\Phi) = \iint \left[\left(\frac{\partial^2 \Phi}{\partial x^2} \right)^2 + 2 \left(\frac{\partial^2 \Phi}{\partial x \partial y} \right)^2 + \left(\frac{\partial^2 \Phi}{\partial y^2} \right)^2 \right] dx dy \quad (1.5)$$

or, more generally,

$$J_m(\Phi) = \iint \sum_{\nu=0}^m \binom{m}{\nu} \left(\frac{\partial^m \Phi}{\partial x^\nu \partial y^{m-\nu}} \right)^2 dx dy, \quad m = 2, 3, \dots \quad (1.6)$$

If $J_m(\Phi)$ is small, then Φ will be smooth

We have deliberately omitted any mention of the domain of integration. If the domain of integration in (1.5) and (1.6) is taken as a bounded region R then it can be shown that the minimizer of (1.4) satisfies

$$\Delta^m \Phi = 0, \quad (x, y) \neq (x_i, y_i), \quad i = 1, 2, \dots, N,$$

where Δ is the Laplacian, i.e.,

$$\Delta \Phi = \frac{\partial^2 \Phi}{\partial x^2} + \frac{\partial^2 \Phi}{\partial y^2},$$

and it satisfies the natural (Neumann) boundary conditions. This result in a similar problem appears in Dyn and Wahba (1979). We avoid the necessity of solving a boundary-value problem by letting the domain of integration be $-\infty \leq x, y \leq \infty$. The boundary conditions are shifted to ∞ . The solution will be defined for $-\infty < x, y < \infty$. However, we will only compute it on R and, of course, it will only have meaning if there are data points not too far from the boundary. We are also assuming here that the world is flat in R , although the entire analysis that we do here can be done on the sphere [for the theory, see Wahba (1979c)].

The solution, which we call $\Phi_{\lambda, m, \lambda}$, to the problem is as follows. Find Φ in a suitable space X to minimize

$$N^{-1} \sum_{i=1}^N [\Phi(x_i, y_i) - \phi_i]^2 + \lambda \iint_{-\infty}^{\infty} \sum_{\nu=0}^m \binom{m}{\nu} \left(\frac{\partial^m \Phi}{\partial x^\nu \partial y^{m-\nu}} \right)^2 dx dy \quad (1.7)$$

This was obtained by Duchon (1976a) and further studied by Meinguet (1978, 1979) and Wahba (1979a,b). It is known as a 'thin plate spline' and is a natural generalization to two dimensions of the one-dimensional smoothing polynomial spline (Reinsch, 1967).

We will give an explicit computable formula for $\Phi_{\lambda, m, \lambda}$ later. Problems in assigning boundary values are eliminated, and no preliminary analysis of the raw data is used.

$\Phi_{\lambda, m, \lambda}$ may be considered as the result of applying a low-pass filter to the data. In frequency space it can be shown that λ controls the half-power point of the filter and m the steepness of the roll-off [see Wagner (1971), Craven and Wahba (1979) and Wahba (1978a)]. In one dimension the filter function $f(\nu)$ as a function of wavenumber ν looks like $f(\nu) = 1/(1 + \lambda \nu^{2m})$. We choose λ and m from the data by the GCV (generalized cross-validation) method (Craven and Wahba, 1979; Golub *et al.*, 1979) which proceeds as follows. The criteria for a good choice of λ and m is taken to be the ability to predict the value of the field where data are withheld.

To estimate this predictive ability from the data let $\Phi_{\lambda, m, \lambda}^{(k)}$ be the function which is the minimizer of (1.7) with the k th data point omitted. If λ and m are good choices, then on the average $\Phi_{\lambda, m, \lambda}^{(k)}(x_k, y_k) - \phi_k$ should be small and we measure this by the ordinary cross-validation function

$$V_m^0(\lambda) \equiv N^{-1} \sum_{k=1}^N [\Phi_{\lambda, m, \lambda}^{(k)}(x_k, y_k) - \phi_k]^2 \quad (1.8)$$

This expression is difficult to compute, furthermore, effects of unequal spacing of data points are not suitably accounted for. For these and other technical reasons recounted in Craven and Wahba (1979) and Golub *et al.* (1979), one should measure the ability of $\Phi_{\lambda, m, \lambda}$ to predict missing data by the generalized cross-validation function (GCVF)

$$V_m(\lambda) = N^{-1} \sum_{k=1}^N [\Phi_{\lambda, m, \lambda}^{(k)}(x_k, y_k) - \phi_k]^2 u_k(m, \lambda), \quad (1.9)$$

where the $u_k(m, \lambda)$ are certain weights which have been given in Craven and Wahba (1979) and Golub *et al.* (1979). $V_m(\lambda)$ turns out to have a collapsed representation which is relatively easy to compute. For each $m = 2, 3, 4, \dots$, up to some preset maximum, $V_m(\lambda)$ is computed as a function of λ and the value $\lambda(m)$ of λ minimizing $V_m(\lambda)$ is determined. Then m is selected by comparing $V_m(\lambda(m))$ over m . A computer implementation of this example has been made and applied to data simulated from a mathematical model for a 500 mb height field. The results are presented in Section 4.

We next generalize this approach to allow the imposition of weak constraints. Continuing with $p = 500$ mb $t = 0$ we consider as an example the geostrophic wind approximation

$$u_g \approx -f^{-1} \partial \Phi / \partial y, \quad v_g \approx f^{-1} \partial \Phi / \partial x,$$

where Φ is the 500 mb height, u_g and v_g are eastward and northward components of the geostrophic wind, and f is the Coriolis parameter. If the eastward and northward components of the wind are measured at each station, one can seek Φ to minimize

$$N^{-1} \sum_{i=1}^n \sigma_1^{-2} [\Phi(x_i, y_i) - \phi_i]^2 + N^{-1} \sum_{i=1}^n \sigma_2^{-2} \left(\frac{\partial \Phi}{\partial y} \Big|_{x_i, y_i} + f \tilde{u}_i \right)^2 + N^{-1} \sum_{i=1}^n \sigma_3^{-2} \left(\frac{\partial \Phi}{\partial x} \Big|_{x_i, y_i} - f \tilde{v}_i \right)^2 + \lambda J_m(\Phi), \quad (1.10)$$

where $N = 3n$, Φ is the measured 500 mb height and \tilde{u}_i, \tilde{v}_i are the observed wind components at station i . σ_1^2 is a weight which is, ideally, the mean-square error in the measured height field. σ_2^2 is the sum of the mean-square error in the measured eastward component of the wind and the mean-square error in the geostrophic approximation to the true eastward wind. σ_3^2 has the corresponding meaning for the northward component of the wind.

For $m \geq 3$ an explicit formula for the minimizer $\Phi_{\lambda, m, \lambda}$ of (1.10) will be given.

Since we are going to choose λ from the data, it is only necessary that σ_1^2/σ_2^2 and σ_1^2/σ_3^2 are known reasonably well. Assuming all mean-square errors are known, it has been suggested by Reinsch (1967) and others to choose λ so that the first three terms in (1.10) with Φ replaced by $\Phi_{\lambda, m, \lambda}$ sum to 1. However, it has been shown (see Wahba, 1975; Craven and Wahba, 1979) that this will lead systematically to undersmoothing.

The idea of the generalized cross-validation function extends to the choice of λ and m in the minimizer of (1.10) and we can obtain the GCVF $V_m(\lambda)$ which can be minimized to estimate good values of m and λ .

In this example where σ_1^2, σ_2^2 and σ_3^2 may be different the minimizer of the GCVF estimates λ and m which best predict missing data points, inversely weighted by the appropriate σ_i^2 .

We next turn to the analysis of a temperature field using both direct (balloon) and remote (satellite radiance) data. We assume that all data are measured at $t = 0$ and that $\Phi(x, y, p)$ represents the temperature. The data consist of direct measurement of the temperature from station i at pressure p_i , and indirect satellite measurements of radiances $I_i(\nu)$ at frequency ν and subsatellite point (x_i, y_i) . In the simplest case (cloudless, looking down), after some linearization and approximations, a known function $r_i(\nu)$ of the measured radiance $I_i(\nu)$ can be related to the temperature Φ by

$$r_i(\nu) = \int_0^{p_i} K(\nu, p) \Phi(x_i, y_i, p) dp, \quad (1.11)$$

* In the process of linearizing to obtain (1.11), first order is used. One could obtain this first order field by analyzing the balloon data alone by leaving out the radiance data in what follows [second term in (1.12)].

where $K(\nu, p)$ is known for each frequency $\nu = \nu_1, \dots, \nu_n$ (see Fritz *et al.*, 1972).

Thus we seek Φ to minimize

$$N^{-1} \sum_{i,k} \sigma_k^{-2} [\Phi(x_i, y_i, p_k) - \phi_{i,k}]^2 + N^{-1} \sum_{i,\nu} \sigma_\nu^{-2} \times \left[\int_0^{p_i} K(\nu, p) \Phi(x_i, y_i, p) dp - r_i(\nu) \right]^2 + \lambda J_m(\Phi) \quad (1.12)$$

where N is the total number of observations and

$$J_m(\Phi) = \sum_{\alpha_1 + \alpha_2 + \alpha_3 = m} \left(\frac{m!}{\alpha_1! \alpha_2! \alpha_3!} \right) \times \iiint_{-\infty}^{\infty} \left(\frac{\partial^m \Phi}{\partial x^{\alpha_1} \partial y^{\alpha_2} \partial p^{\alpha_3}} \right)^2 dx dy dp. \quad (1.13)$$

We will give an explicit formula for the minimizer $\Phi_{\lambda, m, \lambda}$ of (1.12) and the GCVF $V_m(\lambda)$ for this problem for $m \geq 2$. In theory, there is no difficulty in adding weak temperature constraints, or in carrying out the analysis in three space variables and one time variable with direct data, indirect data and weak constraints (A finite number of strong constraints can be added, too, and we briefly indicate how.) In practice the method has computational limits. The computation of $\Phi_{\lambda, m, \lambda}$ requires the solution of a linear system of dimension close to the number N of "data" and "weak constraint" terms. The computation of the GCVF required the solution of an eigenvalue problem of size N . We are obtaining very good results with N up to as large as 140 with present methods on the Univac 1110 at the University of Wisconsin, Madison, but improved algorithms will have to be developed to go beyond this point on this size machine. There is reason to believe that this can be done. Some algorithms handling four times as many points in certain special cases have been developed by Paihua (1978). Other numerical methods suitable for large data sets are suggested in Wahba (1980a,b). $\Phi_{\lambda, m, \lambda}$ is found in terms of coefficients of certain basis functions, so that the bulk of the numerical work is only done once for each set of data $\Phi_{\lambda, m, \lambda}$, and in certain cases its derivatives, can be evaluated on a fine grid essentially for free.

We briefly mention the relationship of this work to some other approaches in the literature. Fritsch (1971) discusses a related form of two-dimensional spline objective analysis. Wagner (1971) analyzed some of Sasaki's variational objective analysis methods from the point of view of their properties as low-pass filters and experimented with the parameter which controls the half power point of the filter (here λ), with the equivalent of our $m = 2$. The Fields of Information Blending developed by M.

Holl and associates also has the capability of blending different types of data. In our notation Holl's approach is to minimize a discrete approximation to the Φ which minimizes

$$\sum_{i=1}^N \sigma_i^{-2} (L_i \Phi - \phi_i)^2$$

[see Holl (1976), Eq. (5)] The data are assembled on a regular discrete grid and Φ is computed only on the same grid. Derivatives are replaced by finite differences. Some of the smoothing is effected by the fact that there are more terms in the sum above than there are grid points on which Φ is to be computed and, in addition, the system of equations to be solved to obtain the minimizer is solved approximately by iterative techniques, where the choice of weighting parameters and number of iterations will have a filtering effect (see, also, Wahba, 1980a, Section 8).

The discussion would not be complete without noting that, in general, variational objective analysis methods involving a quadratic non-negative definite penalty term like $\lambda J_m(\Phi)$ are intimately related to certain forms of (Gandin) optimum objective analysis methods. We illustrate this remark by a simple discretized example. Consider a vector of variables of interest $x = (x_1, x_2, \dots, x_n)'$. Suppose $z_i = x_i + e_i$ is observed for $i = 1, \dots, n$, where the e_i are supposed to be zero mean independent Gaussian random variables with variance σ^2 . Suppose that the x_i have a prior Gaussian distribution with $E x_i = 0$ and $E x_i x_j = \sigma_{ij}$, where E is mathematical expectation. Letting Σ be the $n \times n$ matrix with ij th entry σ_{ij} , then the conditional expectation \hat{x} of x given the data $z = (z_1, \dots, z_n)'$ is

$$\hat{x} = \Sigma(\Sigma + \sigma^2 I)^{-1} z,$$

where I is the $n \times n$ identity matrix. However, it is also true that \hat{x} given above is the solution to the minimization problem find x to minimize

$$n^{-1} \sum_{i=1}^n (x_i - z_i)^2 + \lambda J(x),$$

where $J(x) = x' \Sigma^{-1} x$ and $\lambda = \sigma^2/n$. Returning to functions $\Phi(x, y)$ for example, there is a prior covariance on $\Phi(x, y)$ such that $\Phi_{\lambda, m, \lambda}$, the minimizer of (1.7) has the property that $\Phi_{\lambda, m, \lambda}(x, y)$ is the conditional expectation of $\Phi(x, y)$ given the data $z_i = \Phi(x_i, y_i) + e_i$, where the e_i are independent zero mean Gaussian (error) random variables with common variance σ^2 . The theory behind this remark can be found in Kunsch and Wahba (1970, 1971) and Wahba (1978b, 1979c). The choice of m controls the rate of decay of the power spectrum of the signal with wavenumber, equivalently the shape of the low-pass filter in the frequency domain. Thiebaux (1980) discusses the relationship of m to prior covariances in some related but slightly different

examples. Details for low-pass filtering on the sphere by variational methods may be found in Wahba (1979c, Section 4.3).

In Section 2 we provide the solution to a general minimization problem of which all the previously mentioned problems are special cases. In Section 3 we describe the GCVF which allows the estimation of λ and m from the data being analyzed. In Section 4, results of a Monte Carlo test of the method is given, using realistic simulated 500 mb height data where the "true" field is known. Numerical methods used are somewhat nonstandard and are described in some detail in the Appendices.

Analysis of the height field via minimization of (1.4) is an isotropic method. Thiebaux (1977) has provided some evidence that an improved analysis may be obtained using methods which have different north-south and east-west scales. This feature may be incorporated here by making a change of scale $x \rightarrow kx$ and $y \rightarrow ky$. A good scale parameter k may be estimated by GCV simultaneously with λ and m . Some very preliminary numerical results with actual reported 500 mb height data from the U.S. rawinsonde network suggests that the $k = 1$ (i.e., isotropic) analysis can be improved upon by estimating k (see Wendelberger 1981). We do not discuss anisotropic methods further here.

Kreiss (1979a,b) notes that for successful numerical solution of certain differential equations related to numerical weather forecasting, it is desirable to have initial conditions that have certain continuity properties. We conjecture that the methods suggested here can be used to provide these initial conditions.

2. Solution of a general minimization problem.

In this section we give a solution to a general minimization problem of which the minimization problems of Eqs. (1.7), (1.10) and (1.12) are special cases.

Our results hold in any number of dimensions, where most meteorological problems of interest will involve $d = 2, 3$ or 4 . The $d = 1$ case results in the familiar polynomial smoothing spline (see Reinsch, 1967). We will say a function u of d variables x_1, x_2, \dots, x_d is "smooth" if $J_m(u)$, defined by

$$J_m(u) = \sum_{\alpha_1 + \dots + \alpha_d = m} \frac{m!}{\alpha_1! \alpha_2! \dots \alpha_d!} \times \int \dots \int \left(\frac{\partial^m u}{\partial x_1^{\alpha_1} \dots \partial x_d^{\alpha_d}} \right)^2 dx_1 \dots dx_d,$$

is small.

We seek to find a u which is simultaneously compatible with the data z_1, z_2, \dots, z_n , and is appropriately smooth. The data are assumed to be

$$z_i = L_i u + e_i,$$

where the L_i are (any) continuous linear functionals of u and the e_i are measurement errors. A rigorous definition of a continuous linear functional as being used here is given in Appendix A, but we note the most useful ones here. Let $t^* = (x_1^*, \dots, x_d^*)$ be a fixed point in d dimensions. Then

$$Lu = u(t^*)$$

is a continuous linear functional for each fixed t^* provided

$$2m - d > 0$$

and

$$Lu = \frac{\partial^k u}{\partial x_1^{\alpha_1} \dots \partial x_d^{\alpha_d}} \Big|_{x_1=x_1^*, \dots, x_d=x_d^*}$$

with $\alpha_1 + \dots + \alpha_d = k$ is also a continuous linear functional for each fixed t^* , provided

$$2m - 2k - d > 0$$

This allows incorporation of the winds as estimates of the gradient of the pressure field via the geostrophic approximation L of the form

$$Lu = \int_{\Omega} \left[K(x_1, \dots, x_d) u(x_1, \dots, x_d) dx_1 \dots dx_d \right]$$

is a continuous linear functional if, for example, Ω is a bounded set and

$$\int_{\Omega} |K(x_1, \dots, x_d)| dx_1 \dots dx_d < \infty.$$

This allows merging of radiance data with direct temperature data in the objective analysis of temperature fields. We remark that $Lu = u(t^*)$ is not a continuous linear functional if $m = 1, d = 2$, and this leads to the difficulties mentioned previously in regard to the minimization of (1.2).

We suppose that the e_i are independent zero mean errors with $Ee_i^2 = \sigma_i^2$. We seek to find u in a suitable (Hilbert) space of functions (defined in Appendix A) to minimize

$$N^{-1} \sum_{i=1}^N (Lu_i - z_i)^2 \sigma_i^{-2} + \lambda J_m(u) \quad (2.1)$$

In this Section λ and m are fixed. In Section 3 we show how to choose λ and m . We will give an explicit formula for the u which minimizes (2.1) for general L . The special cases (1.7), (1.10) and (1.12) and others of interest can then be deduced. Computational algorithms are discussed in the Appendices and a numerical test of the method on simulated 500 mb height data is given in Section 4.

The minimizer u of (2.1) is expressible in terms of polynomials of total degree less than m and the fundamental solutions of the iterated Laplacian. Before stating the result, we define some

notation. In d -dimensional space there are

$$M = \binom{d+m-1}{d}$$

polynomials of total degree less than or equal to $m-1$. We let $\{\phi_i\}_{i=1}^M$ be these M polynomials. For example, if $d = 2, m = 3$, then $M = 6$ and

$$\left. \begin{aligned} \phi_1(x_1, x_2) &= 1, & \phi_2(x_1, x_2) &= x_1 \\ \phi_3(x_1, x_2) &= x_2 \\ \phi_4(x_1, x_2) &= x_1^2, & \phi_5(x_1, x_2) &= x_1 x_2 \\ \phi_6(x_1, x_2) &= x_2^2 \end{aligned} \right\} \quad (2.2)$$

Observe that $J_m(\phi_i) = 0, i = 1, 2, \dots, M$, so that polynomials of total degree $\leq m-1$ are considered infinitely smooth by this method. We define the Laplacian Δ by

$$\Delta u = \sum_{k=1}^d \frac{\partial^2 u}{\partial x_k^2}$$

If u and all its derivatives up to order $m-1$ are continuous and are zero at infinity, then by integration by parts, one has

$$J_m(u) = \int_{R^d} \int u \Delta^m u dx_1 \dots dx_d$$

Thus, the iterated Laplacian Δ^m would play a role in an Euler equation approach for the solution of the variational problem (2.1), although we do not use that method to obtain the solution.

Letting $s = (x_1, \dots, x_d), t = (x_1, \dots, x_d)$ and

$$|s - t| = \left[\sum_{i=1}^d (x_i - y_i)^2 \right]^{1/2},$$

then the fundamental solution of the iterated Laplacian is given by $E_m(s, t)$ defined by

$$E_m(s, t) = E(|s - t|), \quad (2.3)$$

where

$$E(r) = \begin{cases} \theta_m r^{2m-d} \ln r, & d \text{ even,} \\ \theta_m = \frac{(-1)^{d/2+1+m}}{2^{2m-1} \pi^{d/2} (m-1)! (m-d/2)!} \\ \theta_m r^{2m-d}, & d \text{ odd, } \theta_m = \frac{\Gamma(d/2 - m)}{2^{2m} \pi^{d/2} (m-1)!} \end{cases}$$

[$E_m(s, t)$ has the property $\Delta_m^m E_m(s, t) = \delta(s - t)$, where the subscript (s) indicates that Δ^m is applied to E_m considered a function of s , and δ is the delta function; see Schwartz (1966)]. We can now state the result.

Let L_1, L_2, \dots, L_N be N linearly independent continuous linear functionals and suppose

$$L_k \sum_{i=1}^N a_i \phi_i = 0, \quad k = 1, 2, \dots, N. \quad (2.4)$$

implies that all the a_i are 0.

Then the solution to the problem Find $u_{\lambda, m, \lambda}$ to minimize

$$N^{-1} \sum_{j=1}^N \left(\frac{L_j u - z_j}{\sigma_j} \right)^2 + \lambda J_m(u) \quad (2.5)$$

is unique and has the representation

$$u_{\lambda, m, \lambda}(t) = \sum_{j=1}^N c_j \xi_j(t) + \sum_{\nu=1}^M d_\nu \phi_\nu(t), \quad (2.6)$$

where

$$\xi_j(t) = L_{j(s)} E_m(t, s), \quad j = 1, 2, \dots, N, \quad (2.7)$$

and $L_{j(s)}$ means the linear functional L_j applied to what follows considered as a function of s . The coefficients $c = (c_1, \dots, c_N)'$ and $d = (d_1, \dots, d_M)'$ are determined by

$$(K + N\lambda D_\sigma^{-2})c + Td = z, \quad (2.8)$$

$$T'c = 0, \quad (2.9)$$

where K is the $N \times N$ symmetric matrix with jk th entry

$$L_{j(s)} L_{k(s)} E_m(s, t), \quad (2.10)$$

T is the $N \times M$ matrix with $j\nu$ th entry

$$L_j \phi_\nu, \quad (2.11)$$

and D_σ is the $N \times N$ diagonal matrix with jj th entry σ_j . An outline of the derivation is given in Appendix A.

EXAMPLES

The simplest example is when the bounded linear functionals are all evaluation functionals $L_j u = u(t_j)$, $j = 1, 2, \dots, N$. For condition (2.4) to be satisfied it is necessary that the N points t_1, \dots, t_N do not lie in a hyperplane of dimension $d - 1$ or less. For example, if $d = 2$, then we need

$$N \geq \binom{m+1}{2}$$

and the N points must not fall on a straight line. Then

$$L_{j(s)} E_m(s, t) = E_m(t_j, t) = \xi_j, \quad (2.12)$$

$$L_{j(s)} L_{k(s)} E_m(s, t) = E_m(t_j, t_k), \quad (2.13)$$

$$L_j \phi_\nu = \phi_\nu(t_j) \quad (2.14)$$

If

$$Lu = \int_a^b K(x_3) u(x_1^*, x_2^*, x_3) dx_3,$$

then

$$\xi(x_1, x_2, x_3) = \int_a^b K(y_3) E_m(x_1^*, x_2^*, y_3, x_1, x_2, x_3) dy_3,$$

etc. In general, ξ of this form may not be known explicitly, and then a quadrature approximation may be necessary. An appropriate quadrature approximation for a similar problem can be found in Dyn and Wahba (1979). An example of ξ , when $L_j u$ involves derivatives is given in Appendix B.

We make the important observation that estimates of derivatives of u up to order l may be obtained by differentiating $u_{\lambda, m, \lambda}$ analytically, provided $2m - 2l - d > 0$.

We remark that in the more familiar Hilbert spaces of functions for which it is only assumed that

$$\left[\int \dots \int u^2(x_1, \dots, x_d) dx_1, \dots, dx_d \right]^{1/2} < \infty,$$

the evaluation functionals $L_k u = u(t_k)$ are not continuous linear functionals.

3. The generalized cross-validation (GCV) method for choosing λ and m

We describe the generalized cross-validation (GCV) method for choosing λ and m . We emphasize that λ and m are the "tuning parameters" of this method (every objective analysis technique has tuning parameters!) and one of the novel features being reported here is the ability to estimate good values of λ and m automatically from the data being analyzed. Frequently, this task is performed by trial and error. We remark that GCV also can be used with other methods but we do not pursue this point here.

To describe the GCV method, we first define the "ordinary" cross-validation function $V_m^0(\lambda)$. Let $u_{\lambda, m, \lambda}^{(k)}$ be the minimizer of

$$N^{-1} \sum_{j \neq k} \left(\frac{L_j u - z_j}{\sigma_j} \right)^2 + \lambda J_m(u),$$

i.e., the k th data point has been left out. Then,

$$L_k u_{\lambda, m, \lambda}^{(k)} - z_k \quad (3.1)$$

is the difference between the k th data point and an estimate of the k th data point from the remaining data when m and λ are used. If m and λ are a good choice the quantities in (3.1) should be small on the average and this can be measured by

$$V_m^0(\lambda) = N^{-1} \sum_{k=1}^N (L_k u_{\lambda, m, \lambda}^{(k)} - z_k)^2 \sigma_k^{-2}. \quad (3.2)$$

The general idea is that one would choose λ and m to minimize (3.2). It turns out that (3.2) is very difficult to compute. Furthermore, it is shown in Craven and Wahba (1979, hereafter CW), Golub *et al.* (1979, hereafter GEFW) and Wahba (1977) that from a theoretical point of view it is better to choose λ and m to minimize a certain weighted version $V_m^w(\lambda)$ of $V_m^0(\lambda)$ defined by

ORIGINAL PAGE IS
OF POOR QUALITY

$$V_m(\lambda) = N^{-1} \sum_{k=1}^N [L_k u_{\lambda, m, \lambda}^{(k)} - z_k]^2 \sigma_k^{-2} w_k(m, \lambda), \quad (3.3)$$

where $w_k(m, \lambda)$ are weights defined by $w_k(m, \lambda)$

$$= [1 - a_{kk}(m, \lambda)]^2 / [1 - N^{-1} \sum_{i=1}^N a_{ii}(m, \lambda)]^2, \quad (3.4)$$

where the $a_{kk}(m, \lambda)$ satisfy

$$\frac{\partial}{\partial z_k} L_k u_{\lambda, m, \lambda} = a_{kk}(m, \lambda). \quad (3.5)$$

If the a_{kk} were all the same the weights would be 1. We first review the results from CW and GHW, on the optimality properties of the GCV estimates $\hat{\lambda}$ and \hat{m} which are obtained as the minimizers of (3.3). Then we provide a simplified expression for $V_m(\lambda)$ which is amenable to computation, and in fact is much easier to compute than $V_m^0(\lambda)$.

The optimality properties of λ and m are based on assuming that

$$z_j = L_j \mu + \epsilon_j, \quad j = 1, 2, \dots, N,$$

where μ is the "true" field and ϵ_j is an error which is assumed to have mean zero and mean-square σ_j^2 .

* In fact, here and elsewhere it is only necessary that the σ_j are relatively correct, since one can multiply all the σ_j by an arbitrary constant which then gets absorbed in λ .

We define an error function when m and λ are used as

$$R_m(\lambda) = EN^{-1} \sum_{j=1}^N (L_j \mu - L_j u_{\lambda, m, \lambda})^2 \sigma_j^{-2}, \quad (3.6)$$

where the E means expected value. $R_m(\lambda)$ is not computable, of course, since μ is not known. However, it is shown in GHW and CW that under rather general circumstances the λ and m which minimize $V_m(\lambda)$ are good estimates of the λ and m which minimize $R_m(\lambda)$, and $EV_m(\lambda) \approx R_m(\lambda) + \alpha$ constant, for λ near the minimizer of $R_m(\lambda)$.

We now give a different, but equivalent, expression for $V_m(\lambda)$ of (3.3) which is suitable for efficient numerical evaluation. First, it can be shown by the same reasoning as in CW, Lemma 3.2, that

$$L_k u_{\lambda, m, \lambda}^{(k)} - z_k \equiv (L_k u_{\lambda, m, \lambda} - z_k) / [1 - a_{kk}(m, \lambda)] \quad (3.7)$$

and, substituting this into (3.3) gives an alternative expression for $V_m(\lambda)$ of (3.3), viz.,

$$V_m(\lambda) = N^{-1} \sum_{k=1}^N (L_k u_{\lambda, m, \lambda} - z_k)^2 \sigma_k^{-2} \times (1 - N^{-1} \sum_{i=1}^N a_{ii})^{-1}, \quad (3.8)$$

where $a_{ii} = a_{ii}(m, \lambda)$. Letting $A_m(\lambda)$ be the $N \times N$ matrix defined by

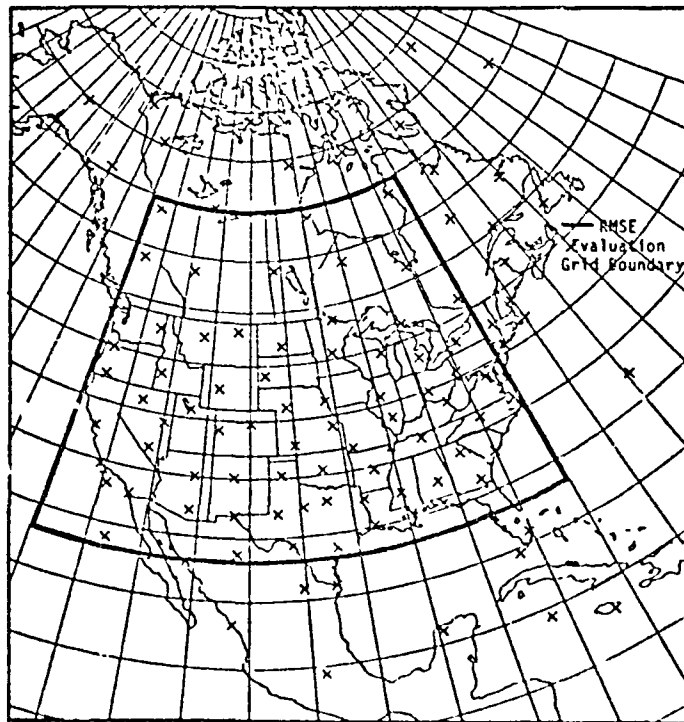


FIG. 1. Location of model radiosonde stations and boundary of grid used for evaluation of the analysis. (The station at San Juan, Puerto Rico, is not shown.)

$$\begin{pmatrix} L_{1j} \\ L_{2j} \\ \vdots \\ L_{Nj} \end{pmatrix} = A_m(\lambda)z, \quad z = (z_1, \dots, z_N)',$$

then the expression (3.8) for $V_m(\lambda)$ can be written in the equivalent form

$$V_m(\lambda) = \frac{N^{-1} \|D_{\sigma}^{-1} [I - A_m(\lambda)]z\|^2}{\|N^{-1} \text{Trace}[I - A_m(\lambda)]\|^2}, \quad (3.9)$$

where D_{σ} is the diagonal matrix with μ th entry σ_{μ} , the trace of a matrix is the sum of its diagonal entries, and $\|\cdot\|$ is the Euclidean norm. An explicit formula for $I - A_m(\lambda)$ in terms of the matrices K , T and D is given in Appendix C. In Appendix C we describe a numerical algorithm for computing $V_m(\lambda)$ and finding the minimizing λ for each m , as well as computing the coefficients c and d . This algorithm was successfully implemented for the special case $d = 2$, $L_{\mu j} = u(t_j)$, $\sigma_j^2 = \sigma^2$, $m = 2, 3, 4, 5$ or 6 and

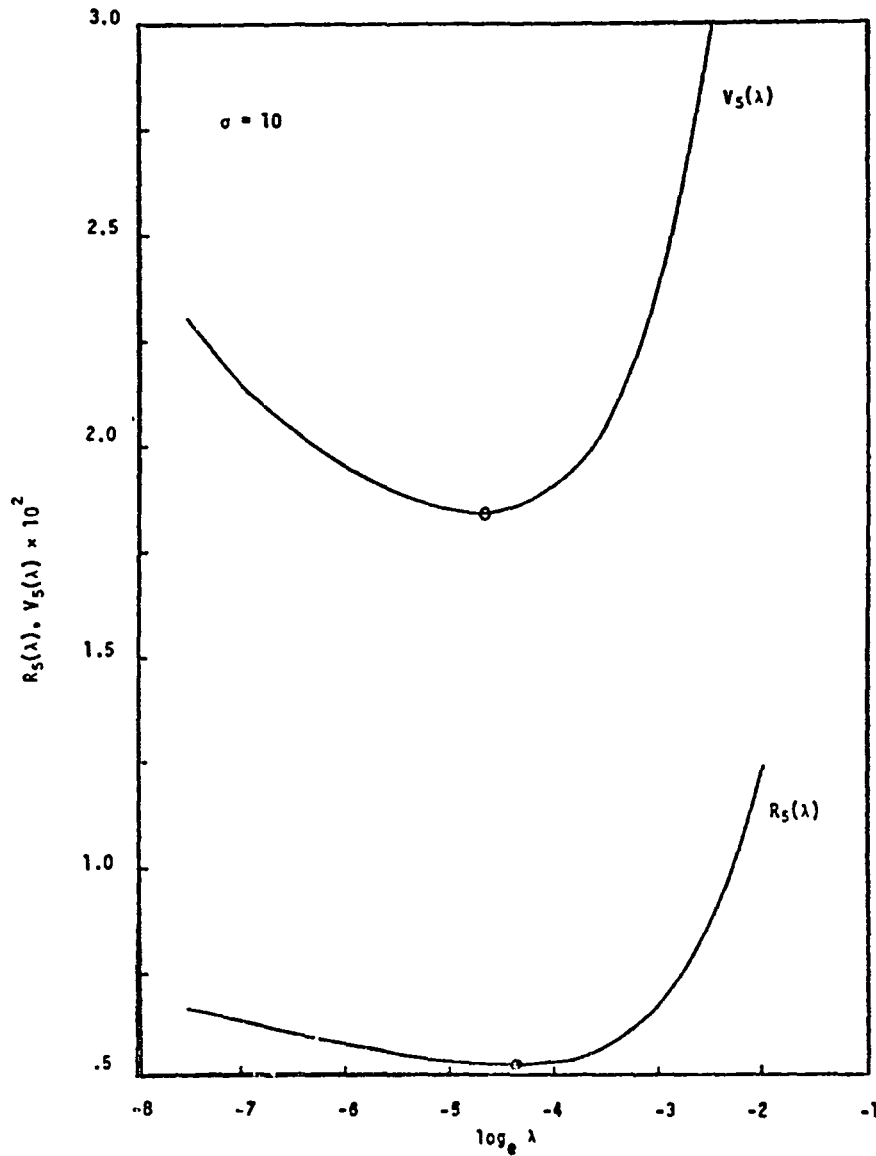


FIG 2 $R_m(\lambda)$ and $V_m(\lambda)$ $m = 5$ Example 1

$N = 120$, to give the numerical results in the next section

4. Numerical experiments

We have programmed and tested the method for analyzing data from simulated 500 mb height fields using simulated data at $N = 114$ North American radiosonde station locations. The simulated data were obtained from a mathematical model of 500 mb height fields used by Dr. Thomas Koehler of the Department of Meteorology at the University of

Wisconsin that was based on an earlier model developed by Sanders (1971). The location of the 114 stations is given in Fig. 1. The equations generating the field are given in Appendix B. Discussion of the rationale behind the model appears in Koehler's (1979) thesis. Contour maps of the model fields appear below together with contour maps of the analyzed fields determined from the simulated data. Data were simulated by computing the true 500 mb height at station i by calling Koehler's program and adding a simulated measurement error. The simu-

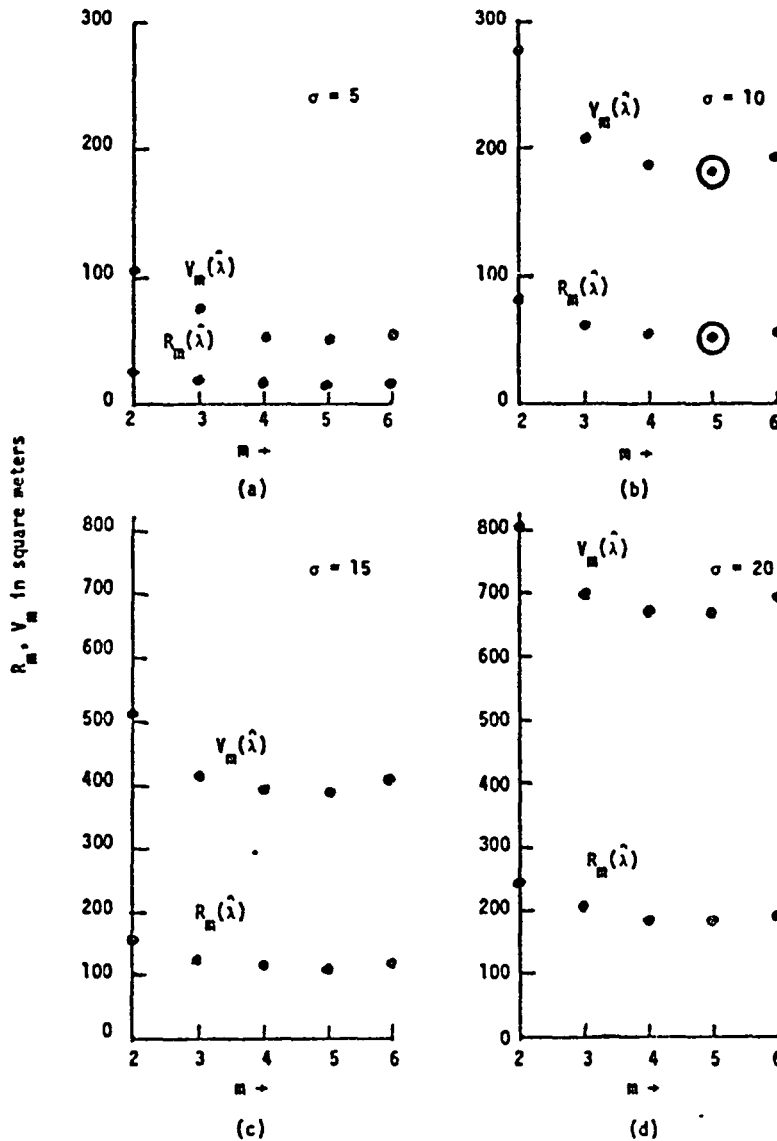
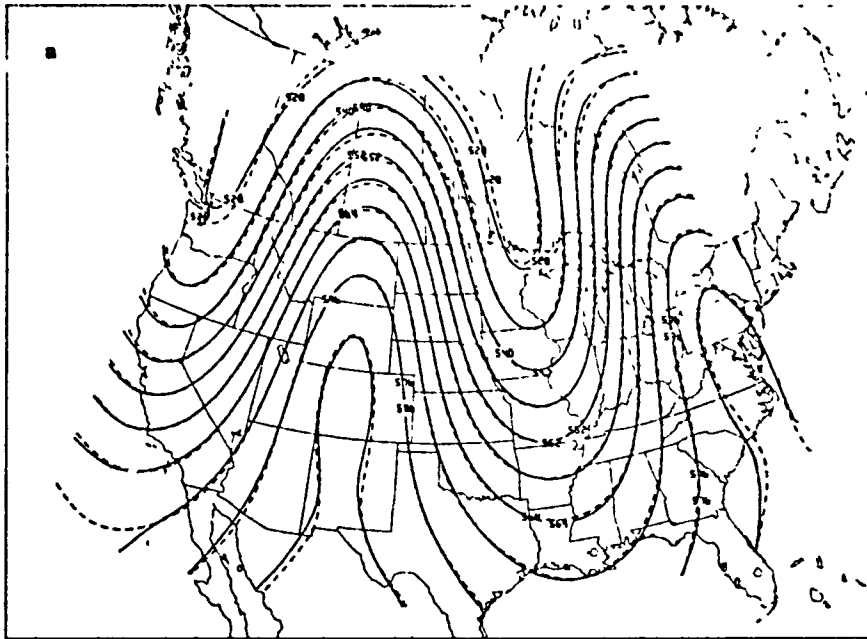
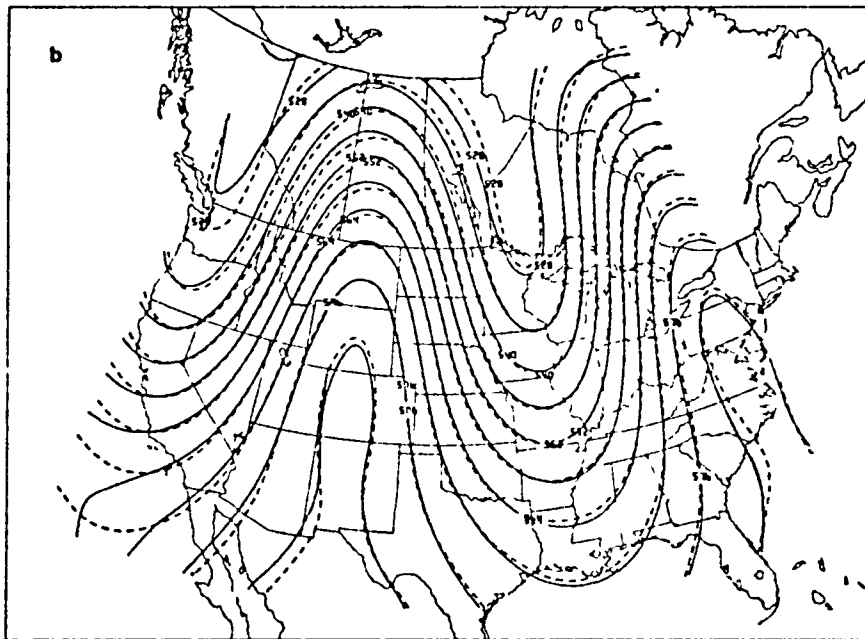


FIG. 7. $V_m(\hat{\lambda})$ and $R_m(\hat{\lambda})$.

FIG 4a The model (dashed line) and analyzed (solid line) fields with $\sigma = 5$

lated measurement error was obtained by calling the pseudo random number generator RAENBR in the University of Wisconsin Academic Computing Center library. This program obtains a pseudo random

normally distributed number with mean 0 and standard deviation 1 and multiplies this number by a constant which is given here as the standard deviation of the measurement error. This procedure re-

FIG 4b As in Fig 4a except with $\sigma = 10$

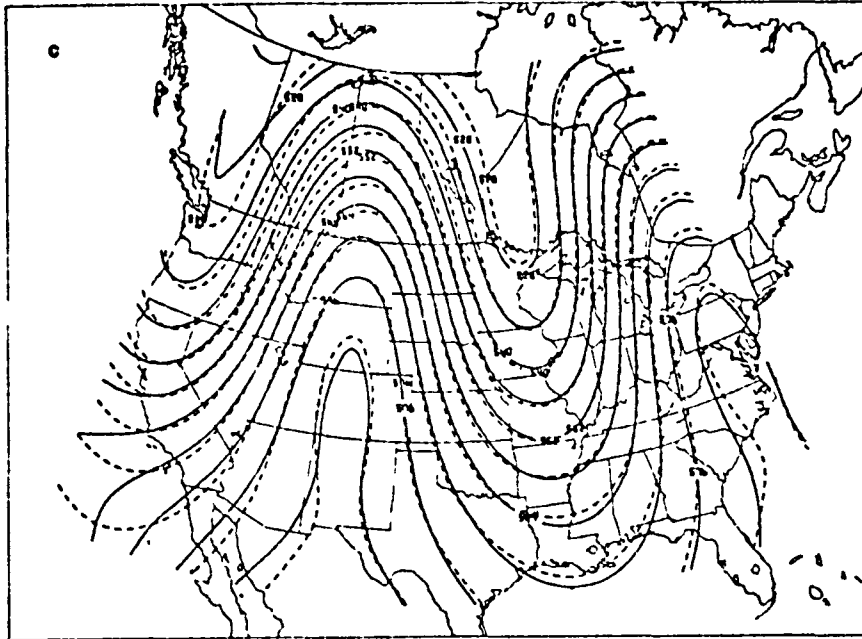


FIG 4c As in Fig 4a except with $\sigma = 15$

sulted in a set of 114 simulated measured 500 mb heights which were then used to obtain an analyzed field. This is the simulated data vector z . To recapitulate the formulas for obtaining the analyzed

field, we go back to Section 2, with $d = 2, N = 114$, $L_i u = u(t_i)$, where $t_i = (x_i, y_i)$, the coordinates of the i th station. We have considered $m = 2, 3, 4, 5$ and 6. The analyzed field is given by $u_{\lambda, m, \lambda}$ of Eq

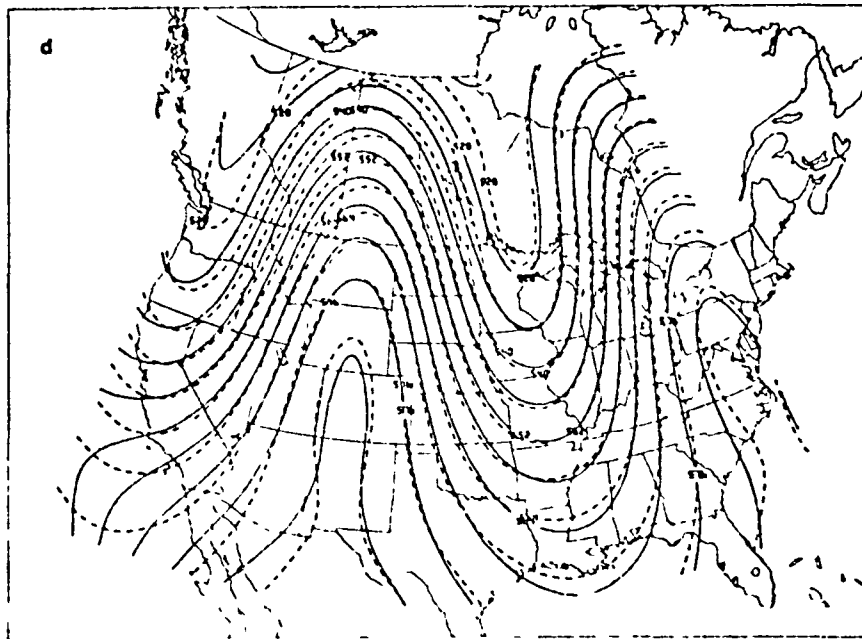


FIG 4d As in Fig 4a except with $\sigma = 20$

(2.6), where ξ_i is defined by (2.12), ϕ_i is defined by (2.2) for $m = 3$ and analogous formulas for other m . K is defined by (2.10) and (2.13), and I is defined by (2.11) and (2.14). For each m , $V_m(\lambda)$ is defined by (3.9), where D_r is taken as the identity matrix since all measurement errors are assumed to have the same standard deviation. $V_m(\lambda)$ is computed as in Appendix C but since D_r is the identity matrix, then $\bar{T} = T$. The earth was assumed "flat" and latitude and longitude coordinates were treated as (x, y) for the analysis of the field and then converted back to latitude and longitude in the contour maps given below. To minimize round-off errors x and y were rescaled to be roughly of magnitude one in absolute value for the calculations.

In the first series of experiments we considered one field (to be called Example 1) and considered $\sigma = 5, 10, 15$ and 20 m. For each data set (i.e., value of σ) we let $m = 2, 3, 4, 5$ and 6 . Let us first examine the choice of λ . In the first example discussed here, $\sigma = 10$ and $m = 5$ ($m = 5$ was the "estimated" m for this case, more about that next.) Fig. 2 gives a plot of $V_5(\lambda)$ vs λ and $R_5(\lambda)$. Here $R_m(\lambda)$ is defined as

$$R_m(\lambda) = N^{-1} \sum_{i=1}^N [u_{i,m,\lambda}(t_i) - u(t_i)]^2,$$

where $u(t_i)$ is the "true", i.e., model 500 mb height field at station i . Theoretically, $V_m(\lambda)$ should "track" $R_m(\lambda)$ near the minimum of $R_m(\lambda)$ (see Craven and

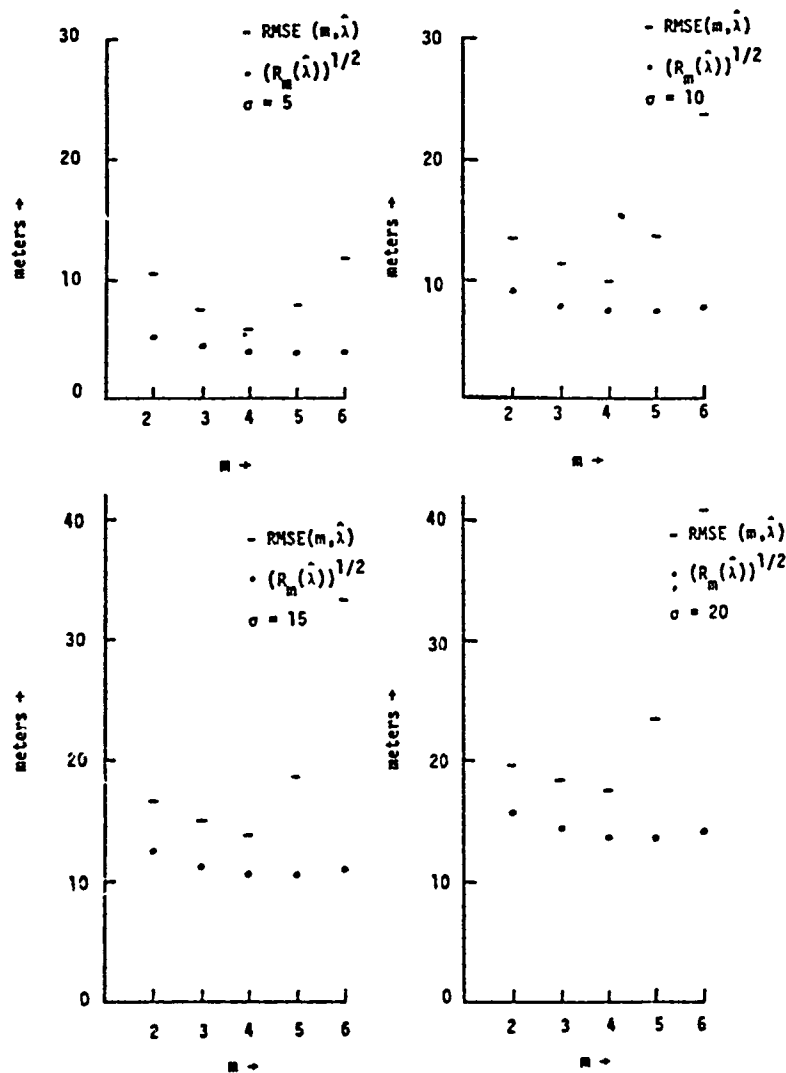


FIG. 5. $RMSE(m, \hat{\lambda})$ and $[R_m(\hat{\lambda})]^{1/2}$ vs m for $\sigma = 5, 10, 15$ and 20 .

Wahba, 1979, Golub *et al.*, 1979) In practice $R_m(\lambda)$ is not known but in this example which is fairly typical, it can be seen that the minimizer, call it $\hat{\lambda}$, of $V_m(\lambda)$ is a very good estimate of the minimizer of $R_m(\lambda)$. In fact the 'inefficiency' $R_m(\hat{\lambda})/\min_{\lambda} R_m(\lambda) = 1.005$.

Fig 3 illustrates how m is chosen from the data and how good this choice is. To study variability of the method with m and σ , the same set of 114 pseudo-random numbers has been used in each of the $20 = 5 \times 4$ analyses behind Fig 3. The pseudo-random number for station i was multiplied by $\sigma = 5, 10, 15$ and 20 in turn to get four data sets.

Fig 3a plots V_m at the minimizing value $\hat{\lambda}$ for $m = 2, 3, 4, 5$ and 6 for the first data set ($\sigma = 5$). The minimizing value $\hat{\lambda}$ will be different in each case. According to Fig 3a the choice of $m = 5$ would be made from the data. For comparison $R_m(\hat{\lambda})$ is also given. Figs 3b, 3c and 3d give the same plots for the other three data sets with $\sigma = 10, 15$ and 20 . It is seen that the choice $m = 5$ would be made from the data in each case. In general, $R_m(\lambda)$ is very close to $\min_{\lambda} R_m(\lambda)$ and these plots suggest that choosing m to minimize $V_m(\hat{\lambda})$ will result in a good choice of m . However $R_m(\hat{\lambda})$ for $m = 4$ and $m = 6$ is only slightly larger than $R_m(\hat{\lambda})$. The two points corresponding to the $m = 5, \sigma = 10$ case of Fig 2 are circled in Fig 3b. Figs 4a-4d give the model and analyzed field for $m = 5$ with the estimated λ for each σ tried. The model field contours (dashed lines) are the same in each figure. The analyzed field contours are solid lines. The contours are labeled in tens of meters.

From the data behind Fig 3 one can establish that $[R_m(\hat{\lambda})]^{1/2}$ is between 0.6σ and 0.8σ . Thus the measurement noise is being filtered out to give a better estimate overall, of the station 500 mb height than the measured heights!

In practice, of course, we want the analyzed field to be a good estimate of the true field over a whole region, not just at the points where it is measured. To determine how well this goal is being met the RMSE (root-mean-square error) of the analyzed field over a 17×26 grid covering the region outlined over North America with a solid line in Fig 1 was computed. This RMSE is defined as follows.

$$\text{RMSE} = \text{RMSE}(m, \hat{\lambda})$$

$$= \left\{ \frac{1}{17 \times 26} \sum_{i=1}^{17} \sum_{j=1}^{26} [u_{i,j}(m, \hat{\lambda}(\theta_i, \phi_j)) - u(\theta_i, \phi_j)]^2 \right\}^{1/2}$$

where $\hat{\lambda}$ is the estimated λ for each m . The RMSE is, of course, an overall measure of how well an entire field can be estimated over a region from the 114 data points.

Fig 5 gives plots of $\text{RMSE}(m, \hat{\lambda})$ for the four values of σ tried. $\text{RMSE}(m, \hat{\lambda})$ is generally greater than $[R_m(\hat{\lambda})]^{1/2}$. For comparison $[R_m(\hat{\lambda})]^{1/2}$ is also plotted. The excess of $\text{RMSE}(m, \hat{\lambda})$ over $[R_m(\hat{\lambda})]^{1/2}$ reflects the inability of the method to interpolate between data points.

It can be seen from Fig 5 that by the RMSE criteria an m somewhat smaller than 5 would give slightly better results in these examples. To what extent this result on a model field carries over to real

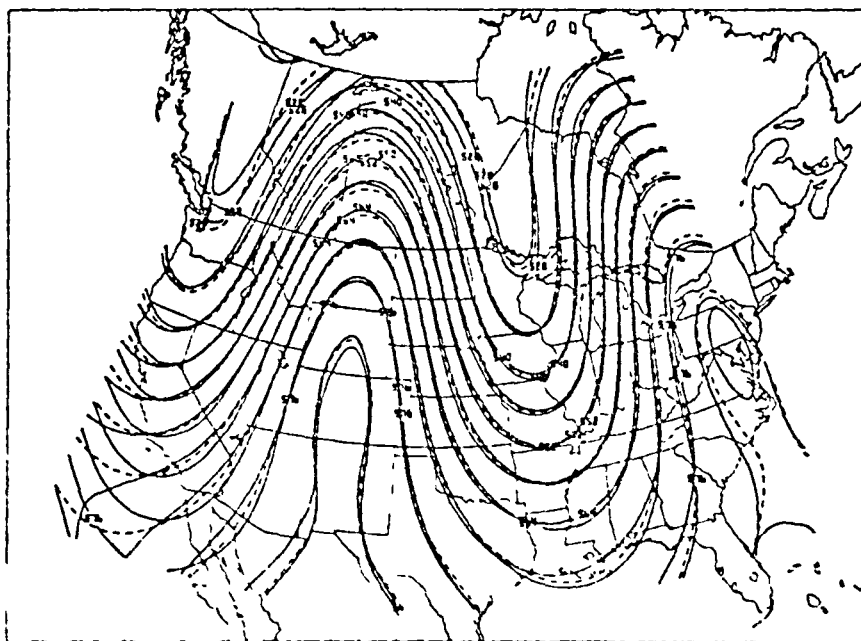


FIG. 6. The model field (dashed line) and two analyzed fields (solid lines) with replicated data.

fields is really a question of how closely the model represents the real world with respect to the feature being tested

To get a feel for the variability of the analysis with actual variation in the measurement errors, Exam-

ple 1 above with $\sigma = 10$ was replicated beginning with a new set of random numbers. $V_p(\lambda)$ was computed from the data and $m = 5$ was again chosen from the data.

The estimated value $\hat{\lambda}$ in the second replicate was

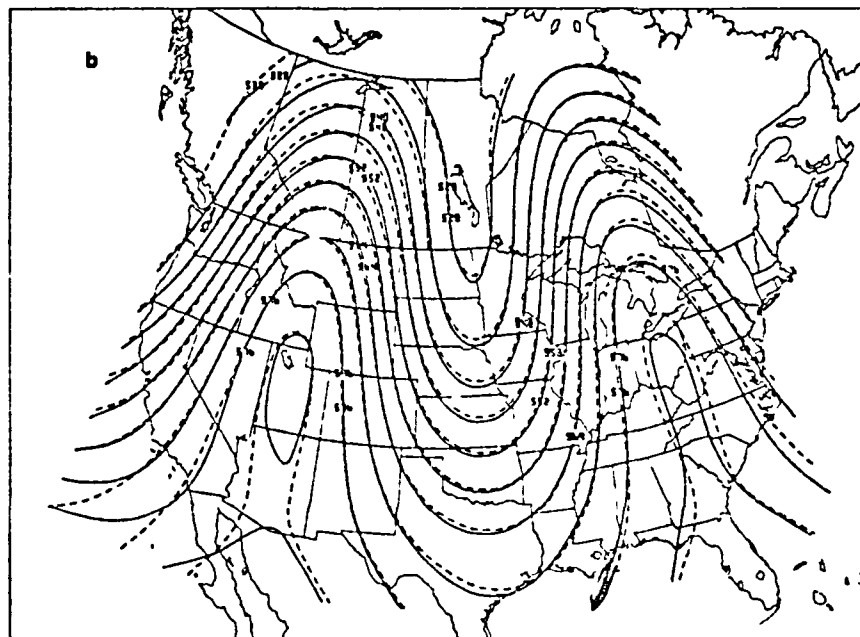
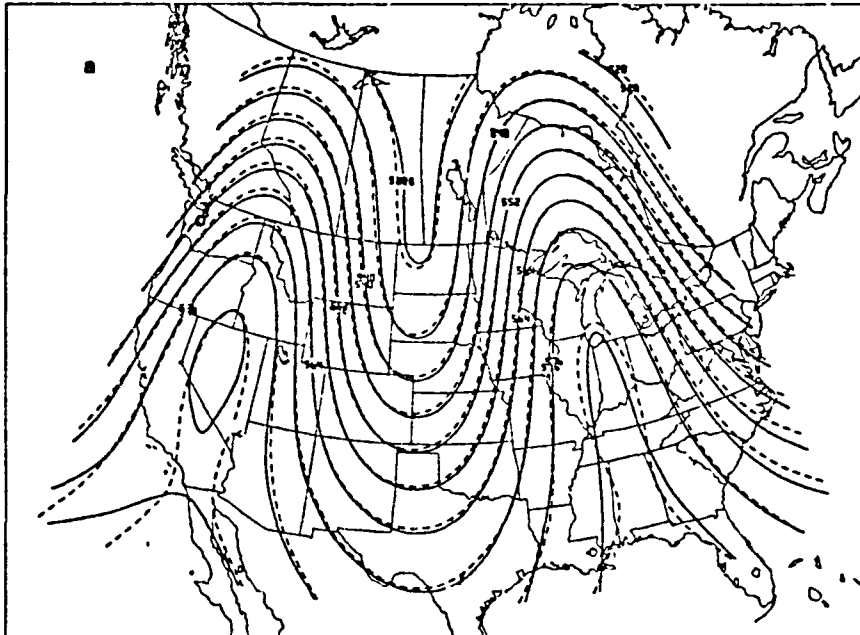


FIG 7 Four examples with $\sigma = 10$ (a) $ALON_0 = 105$, (b) $ALON_0 = 100$, (c) $ALON_0 = 95$, (d) $ALON_0 = 90$

very close to $\hat{\lambda}$ in the first replicate (Remember that the "model" field is identical in both cases.) However, while the RMSE was 13.69 in the first replicate, it was 17.13 in this one. The model and the two analyzed fields for this case appear in Fig. 6

Finally, we look at variations as the field varied. Three other fields, in addition to the first example, were generated by moving the field from west to east. The four fields are characterized by the parameter $ALON_0$ in the model in Example 1.

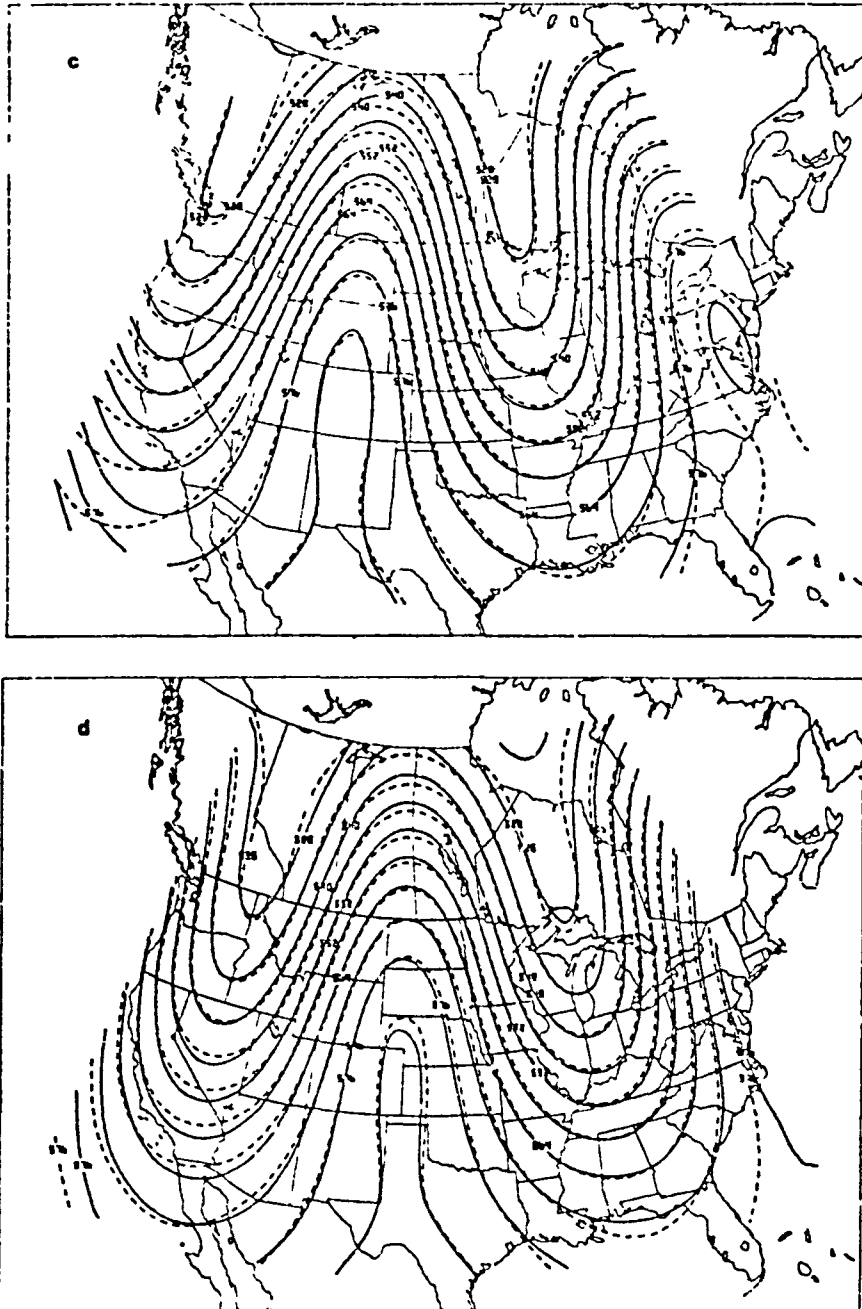


FIG 7 (Continued)

ALON₀ = 95, the other three cases are 90, 100 and 105. The second replicate with ALON₀ = 95 is used in this series, and the same set of 114 original random numbers used in the second replicate is used in the other three examples here. A set of data with $\sigma = 10$ was generated for each of these three new fields. The estimated values of m were

ALON ₀	\hat{m}
105	4
100	4
95	5 (already given)
90	6

Fig. 7 gives a plot of the true and analyzed field in each of these four cases. The RMSE values were

ALON ₀	RMSE($\hat{m}, \hat{\lambda}$)
105	8.40
100	10.80
95	17.13
90	13.08

We have used $\sigma = 10$ as a typical value here because the results in Thiébaux (1977, Table 1, first row, fifth column) suggest that the root-mean-square measurement error at Topeka is less than 10 m (assuming zero mean measurement errors). Note an earlier study (Air Weather Service, 1955) has estimated σ at ~ 20 m.

The question of whether in practice m and λ can be effectively chosen once and for all or should be estimated dynamically from the data has not been completely addressed here. This question can be addressed with "model" data only to the extent that the model represents the real world with respect to the phenomena being studied. Furthermore, if the criteria is minimum RMSE then this question cannot be answered with real data unless it is available on a fine grid. Predictive ability on the measurement grid can be studied in experiments philosophically like those of Thiébaux (1977), who omitted data from Topeka and then examined how well the Topeka data could be estimated from other data. We are presently doing this with both the isotropic and anisotropic method and preliminary results are very promising.

A few preliminary experiments we have carried out with a limited set of examples have resulted in effectively similar values of λ for fixed m . If m and λ can be fixed, then the cost of repetitive estimation of $u_{\lambda, m}$ from data from a given set of stations becomes very inexpensive.

Ultimately, whether or not m and λ should be estimated from the data or can safely be "fixed" at some prior value will have to be determined with respect to the ultimate use to which the analyzed field is put (e.g. if it is used in a forecast model, then one should determine whether dynamic estima-

tion of λ and m is cost effective in terms of better forecasts).

Acknowledgments. The authors wish to acknowledge a number of invaluable discussions with Professor Don Johnson and Drs. Tom Koehler and Tom Whittaker. This research was sponsored by the Atmospheric Sciences Section, National Science Foundation, under Grant ATM75-23223 and the Office of Naval Research under Contract N00014-77-C-0675.

APPENDIX A

Outline of the Derivation of Eq. (2.6)

The solution to the minimization problem of (2.5) will be found by the use of geometry in Hilbert space. By use of classical methods, it is possible to characterize the solution as the solution to a partial differential equation with delta functions and derivatives of delta functions on the right-hand side, but the present approach leads simply to algorithms which do not require the numerical solution to a partial differential equation. The reader not familiar with Hilbert spaces may find Akhiezer and Glazman (1961, pp. 1-21 and 30-35) provide the necessary definitions of Hilbert space norm and inner product. The Hilbert spaces we will use all possess a reproducing kernel which is used to construct the solution; these kernels will be described below. We wish to minimize

$$N^{-1} \sum_{i=1}^N (L_i u - z_i)^2 \sigma_i^{-2} + \lambda J_m(u) \quad (2.5)$$

in an appropriate Hilbert space X of functions for which $J_m(u)$ is finite. We first define a suitable inner product on X . Let s_1, s_2, \dots, s_M be a fixed set of M points in Euclidean d -space with the property that

$$\sum_{i=1}^M a_i \phi_i(s) = 0, \quad \text{for } s = s_1, \dots, s_M$$

implies that all a_i are 0. The particular choice of these points is unimportant as they will cancel out later. An inner product $\langle u, v \rangle$ is defined on X by

$$\langle u, v \rangle = \sum_{i=1}^M u(s_i)v(s_i) + \sum_{\alpha_1 + \alpha_2 = m} \frac{m!}{\alpha_1! \alpha_2!} \times \int_{R_d} \int \frac{\partial^{\alpha_1} u}{\partial x_1^{\alpha_1}} \frac{\partial^{\alpha_2} v}{\partial x_2^{\alpha_2}} dx_1 \dots dx_d \quad (A1)$$

It follows from (A1) that the norm $\|u\|$ on X is given by

¹ The rigorous definition of X is: X is the vector space of all the Schwartz distributions for which all the partial derivatives in the distributional sense of total order m are square integrable. [see Mengesha 1979 Eq. (4)]

AUGUST 1980

GRACE WAHBA AND JAMES WENDELBERGER

1139

$$\|u\|^2 = \sum_{j=1}^M u^2(s_j) + J_m(u) \quad (A2)$$

A (real) continuous linear functional L defined on functions u in X is a functional which assigns a real number to each u with the property

$$L(\alpha u_1 + \beta u_2) = \alpha L u_1 + \beta L u_2$$

for any u_1 and u_2 and furthermore there exists a constant C so that

$$|Lu| \leq C \|u\|, \quad \text{all } u \in X \quad (A3)$$

For the familiar space L_2 of functions with

$$\|u\|^2 = \int_{R_d} u^2(x_1, \dots, x_d) dx_1 \dots dx_d$$

$Lu = u(t^*)$ is not a continuous linear functional because (A3) cannot be satisfied. However, in all the spaces X that we will consider, $Lu = u(t^*)$ will be a continuous linear functional. By the Riesz representation theorem (Akhiezer and Glazman, 1961, p. 33),* if L is a continuous linear functional on functions in a Hilbert space X , then there is a function η_i in X , called the representer of L , such that

$$Lu = \langle \eta_i, u \rangle$$

Suppose these η_i were given. Then our minimization problem is as follows: Find u in X to minimize

$$N^{-1} \sum_{i=1}^N (\langle \eta_i, u \rangle - z_i)^2 \sigma_i^{-2} + \lambda J_m(u) \quad (A4)$$

We look at this problem from a geometric point of view. Any u in the Hilbert space X can be written as a linear combination of $\eta_1, \dots, \eta_N, \phi_1, \dots, \phi_M$ plus some function ρ which is perpendicular to each η_i and ϕ_ν , that is,

$$u = \sum_{i=1}^N c_i \eta_i + \sum_{\nu=1}^M d_\nu \phi_\nu + \rho \quad (A5)$$

for some coefficients $c = (c_1, \dots, c_N)'$, $d = (d_1, \dots, d_M)'$, where

$$\left. \begin{aligned} \langle \eta_i, \rho \rangle &= 0, \quad i = 1, 2, \dots, N \\ \langle \phi_\nu, \rho \rangle &= 0, \quad \nu = 1, 2, \dots, M \end{aligned} \right\} \quad (A6)$$

By substituting (A5) into (A4) and using (A6) repeatedly, one can show that for u of the form (A5) to minimize (A4), it is necessary that $\rho = 0$. By using Lemma 5.1 in Kimeldorf and Wahba (1971), it can also be established [assuming (2.4)] that the coefficients c_i must satisfy

$$\langle \phi_\nu, \sum_{i=1}^N c_i \eta_i \rangle = 0, \quad \nu = 1, 2, \dots, M. \quad (A7)$$

* Akhiezer and Glazman use "linear functional" for what we are calling "continuous linear functional".

which is equivalent to (2.9), i.e.,

$$T'c = 0, \quad (2.9)$$

since $\langle \eta_i, \phi_\nu \rangle = L_i \phi_\nu$. It remains to find the η_i and the coefficients $c = (c_1, \dots, c_N)'$ and $d = (d_1, \dots, d_M)'$. To find the η_i , we use the theory of reproducing kernels. [For more details concerning what follows, see Aronszajn (1950) and Kimeldorf and Wahba (1971).] A Hilbert space X is said to possess a reproducing kernel (rk) if, for each t^* in R^d , the functional $Lu = u(t^*)$ is a continuous linear functional. Then there exists a representer q_t in X such that

$$Lu = u(t^*) = \langle q_t, u \rangle.$$

We define the function $Q(s, t)$ of two (vector) variables s and t by

$$Q(s, t) = \langle q_s, q_t \rangle,$$

where Q is called the reproducing kernel for X . The basic property of the reproducing kernel is that given Q , one can find the representers of any continuous linear functionals. The η_i are given by

$$\eta_i(t) = L_{i(s)} Q(s, t), \quad (A8)$$

and, furthermore,

$$\langle \eta_i, \eta_j \rangle = L_{i(s)} L_{j(t)} Q(s, t), \quad (A9)$$

where, as before, the subscript (s) indicates that the functional L_i is to be applied to what follows considered as a function of s .

Using results in Duchon (1976a, 1976b) and Meinguet (1979), it is possible to deduce that the reproducing kernel $Q(s, t)$ for X with the inner product given by (A.1) is given by

$$Q(s, t) = K(s, t) + P(s, t), \quad (A10)$$

where

$$\begin{aligned} K(s, t) &= L_m(s, t) - \sum_{r=1}^M p_r(t) E_m(s_\nu, s) \\ &\quad - \sum_{\mu=1}^M p_\mu(s) E_m(t, s_\mu) \\ &\quad + \sum_{\mu=1}^M p_\mu(s) p_\mu(t) E_m(s_\mu, s_\mu), \end{aligned}$$

$$P(s, t) = \sum_{r=1}^M p_r(s) p_r(t)$$

$E_m(s, t)$ is as defined in (2.3), and p_1, \dots, p_M are the M polynomials of total degree less than m satisfying $p_i(s_\mu) = 1$, if $\mu = i$ and is equal to zero otherwise. To verify that $Q(s, t)$ is the reproducing kernel for X it is sufficient to check that $\langle q_t, u \rangle = u(t)$, where $q_t(s) = Q(s, t)$, and that q_t is in X . This can be done using Meinguet (1979).

Now, letting $\xi(t)$ be as in (2.7),

$$\xi_i(t) = L_{i(s)} E_m(s, t) \quad (2.7)$$

and using

$$\eta_i(t) = I_{(s)} Q(s, t),$$

and assuming that c_1, \dots, c_N satisfy (A7), it can be verified that

$$\sum_{i=1}^N c_i \eta_i(t) = \sum_{i=1}^N c_i \xi_i(t) - \sum_{i=1}^N \sum_{j=1}^M c_i \xi_i(s_j) p_j(t). \quad (A11)$$

Thus, since the double sum on the right is a polynomial, this establishes that the minimizer of (2.5) has the form

$$u_{N, m, \lambda}(t) = \sum_{i=1}^N c_i \xi_i(t) + \sum_{j=1}^M d_j \phi_j(t). \quad (2.6)$$

The coefficients c and d are obtained as follows: Since $\sum c_i \eta_i$, $\sum c_i \xi_i$ and $u_{N, m, \lambda}$ differ by polynomials,

$$J_m(\sum c_i \eta_i) = J_m(\sum c_i \xi_i) = J_m(u_{N, m, \lambda}).$$

By (A11), $\sum_{i=1}^N c_i \eta_i(s_\nu) = 0$, $\nu = 1, 2, \dots, M$, so by use of (A2) and (A9)

$$J_m(\sum_{i=1}^N c_i \eta_i) = \|\sum_{i=1}^N c_i \eta_i\|^2 = \sum_{i=1}^N \sum_{j=1}^N c_i c_j L_{i(s)} L_{j(s)} Q(s, t). \quad (A12)$$

Using (A7) and (A10) it can be shown that the right-hand side of (A12) is equal to

$$\sum_{i=1}^N \sum_{j=1}^N c_i c_j L_{i(s)} L_{j(s)} E_m(s, t) = c' K c.$$

Using

$$\begin{pmatrix} L_1 u_{N, m, \lambda} \\ \vdots \\ L_N u_{N, m, \lambda} \end{pmatrix} = Kc + Td,$$

Eq. (A4) is equal to

$$N^{-1}(Kc + Td - z)' D_\sigma^{-2} (Kc + Td - z) + \lambda c' K c. \quad (A13)$$

Minimization of this expression with respect to c and d gives the desired equations for c and d , i.e.,

$$(K + N\lambda D_\sigma^{-2})c + Td = z, \quad (2.8)$$

$$T'c = 0 \quad (2.9)$$

We close this Appendix with the observation that one can also enforce strong constraints by the same methods. Suppose one wishes to minimize

$$N^{-1} \sum_{k=1}^N (L_k u - z_k)^2 \sigma_k^{-2} + \lambda J_m(u), \quad (A14)$$

subject to

$$L_j u = z_j, \quad j = N_1 + 1, \dots, N. \quad (A15)$$

The minimizer of (A14) subject to (A15) is obtained by setting $\sigma_j^2 = 0$, $j = N_1 + 1, \dots, N$ in (2.8). However, the computational procedure given in Appendices C and D is not suitable for this case since the procedure involves division by σ_j^2 . For a computational procedure for this case see Wahba (1980b, Sec. 6.2).

APPENDIX B

Example of the Calculation of ξ_j , K and T for L_j Involving Differentiation

Let $d = 2$, $m = 3$

$$L_j u = \frac{\partial u}{\partial x_1} \Big|_{x_1=x_1^j},$$

$$L_k u = \frac{\partial u}{\partial x_1} \Big|_{x_1=x_1^k},$$

then

$$\begin{aligned} \xi_j(x_1, x_2) &= \frac{\partial}{\partial y_1} E_3(y_1, x_1^j, x_1, x_2) \Big|_{y_1=x_1^j} \\ &= \frac{\partial}{\partial y_1} \frac{\theta_2}{2} [(y_1 - x_1)^2 + (x_1^j - x_2)^2] \\ &\quad \times \ln[(y_1 - x_1)^2 + (x_1^j - x_2)^2] \Big|_{y_1=x_1^j} \\ &= \theta_2 \{ 2[(x_1^j - x_1)^2 + (x_1^j - x_2)^2] \\ &\quad \times (x_1^j - x_1) \ln[(x_1^j - x_1)^2 + (x_1^j - x_2)^2] \\ &\quad + [(x_1^j - x_1)^2 + (x_1^j - x_2)^2](x_1^j - x_1) \}. \\ L_{j(s)} L_{k(t)} E_m(s, t) &= \frac{\partial^2}{\partial y_1 \partial x_1} \frac{\theta_1}{2} [(y_1 - x_1)^2 + (x_1^j - x_2)^2] \\ &\quad \times \ln[(y_1 - x_1)^2 + (x_1^j - x_2)^2] \Big|_{y_1=x_1^j, x_1=x_1^k} \\ &= \frac{\partial}{\partial x_1} \xi_j(x_1, x_2^k) \Big|_{x_1=x_1^j}, \end{aligned}$$

etc

APPENDIX C

Calculation of $V_m(\lambda)$ and Its Minimizer, and of c and d

Calculation of c , d and $V_m(\lambda)$ are based on formulas (C1)-(C3) below. These formulas are derived in Appendix D

$$c = R(R'KR + N\lambda R'D_\sigma'R)^{-1}R'z \quad (C1)$$

$$d = (T'D_\sigma^{-2}T)^{-1}T'D_\sigma^{-2}(z - Kc) \quad (C2)$$

[c and d have originally been given in (2.8) and (2.9)]

$$I - A_m(\lambda) = N\lambda D_\sigma^{-2}R(R'KR + N\lambda R'D_\sigma'R)^{-1}R', \quad (C3)$$

where R is any $N \times N - M$ dimensional matrix of rank $N - M$ satisfying $R'T = 0_{N-M, M}$

It is shown in Appendix D that the $N - M \times N - M$ dimensional matrix B defined by $B = R'KR$ is always strictly positive definite (although K may not be). This allows some of the calculations below to proceed.

We now discuss a computational procedure which we have successfully implemented for the special case $d = 2$, $L_i u = u(t_i)$, $\sigma_i^2 = \sigma^2$, $m = 2, 3, 4, 5$ or 6 , and $N \leq 120$.

R can always be chosen so that $R'D_\sigma^2 R = I_{N-M}$ where I_{N-M} is the $N - M$ dimensional identity matrix. This is done numerically as follows. Let $\tilde{T} = D_\sigma^{-1} T$ and form the matrix $C = I - T(\tilde{T}'T)^{-1}\tilde{T}'$. This symmetric non-negative definite matrix is a projection matrix of rank $N - M$ satisfying $T'C = O_{M \times N}$, and so it has $N - M$ eigenvalues equal to 1 and M eigenvalues equal to 0. The $N - M$ eigenvectors $\tilde{r}_1, \tilde{r}_2, \dots, \tilde{r}_{N-M}$, say, corresponding to the ones have the property

$$\tilde{T}'\tilde{r}_j = 0, \quad j = 1, 2, \dots, N - M,$$

and the property that the $N \times N - M$ dimensional matrix \tilde{R} with columns $\tilde{r}_1, \dots, \tilde{r}_{N-M}$ satisfies $\tilde{R}'\tilde{R} = I$. The eigenvectors corresponding to the ones are not individually uniquely defined of course, any set will do. Let $r_j = D_\sigma^{-1}\tilde{r}_j$ and $R = D_\sigma^{-1}\tilde{R}$. Then $T'r_j = \tilde{T}'\tilde{r}_j = 0$, $j = 1, 2, \dots, N - M$ and $R'D_\sigma^2 R = \tilde{R}'\tilde{R} = I$. Thus R is the desired matrix. We successfully used EISPACK (Smith *et al.*, 1976) in double precision to deliver the $\{\tilde{r}_j\}$ given C , for N up to ~ 120 . Once R is determined, let the eigenvalue decomposition of $B = R'KR$ be

$$B = UD_B U',$$

where U is orthogonal and D_B is the diagonal matrix with diagonal entries the eigenvalues b_i , of B , $i = 1, 2, \dots, N - M$. U and D_B are again obtained by EISPACK. The b_i are theoretically all positive. Then c is readily computed from the identity

$$c = RU(D_B + N\lambda I)^{-1}U'R'z$$

and d is computed from

$$d = (\tilde{T}'\tilde{T})^{-1}\tilde{T}'D_\sigma^{-1}(z - Kc).$$

By (C3) we obtain

$$D_\sigma^{-1}(I - A_m)z = N\lambda D_\sigma c$$

and using (3.9) we have

$V_m(\lambda)$

$$\begin{aligned} &= \frac{N^{-1}(N\lambda)^2 \|D_\sigma R U (D_B + N\lambda I)^{-1} w\|^2}{N^{-2}(N\lambda)^2 [\text{Trace } D_\sigma^{-1} R (B + N\lambda R'D_\sigma^2 R)^{-1}]} \\ &= \frac{N^{-1} \sum_{i=1}^{N-M} \frac{w_i^2}{(b_i + N\lambda)}}{\left(N^{-1} \sum_{i=1}^{N-M} \frac{1}{b_i + N\lambda} \right)^2} \end{aligned}$$

where $w = (w_1, \dots, w_{N-M})' = U'R'z$. For fixed m , given the w_i^2 and the b_i , it is not hard to find the λ minimizing the right hand side of this expression by global search. It is convenient to work in units of $\log \lambda$.

APPENDIX D

Derivation of (C1), (C2) and (C3)

We obtain (C1) and (C2) from (2.8) and (2.9).

$$c = R(R'KR + N\lambda R'D_\sigma^2 R)^{-1}R'z, \quad (C1)$$

$$d = (T'D_\sigma^{-2}T)^{-1}T'D_\sigma^{-2}(z - Kc), \quad (C2)$$

$$(K + N\lambda D_\sigma^2)c + Td = z, \quad (2.8)$$

$$T'c = 0. \quad (2.9)$$

Here R is any $N \times N - M$ matrix of rank $N - M$ satisfying $R'T = 0$. Since $T'c = 0$ there exists a unique $N - M$ vector γ , say, with

$$c = R\gamma. \quad (D1)$$

Multiplying the left side of (2.8) by R' and substituting in (D1) gives

$$\left. \begin{aligned} R'(K + N\lambda D_\sigma^2)R\gamma &= R'z \\ \gamma &= [R'(K + N\lambda D_\sigma^2)R]^{-1}R'z \end{aligned} \right\} \quad (D2)$$

and multiplying the left side of (D2) by R gives (C1). To get (C2) we multiply the left side of (2.8) by $T'D_\sigma^{-2}$ to get

$$T'D_\sigma^{-2}Kc + T'D_\sigma^{-2}Td = T'D_\sigma^{-2}z. \quad (D3)$$

Finally we multiply the left side of (D3) by $(T'D_\sigma^{-2}T)^{-1}$ to obtain (C2).

To obtain (C3)

$I - A_m(\lambda)$

$$= N\lambda D_\sigma^2 R(R'KR + N\lambda R'D_\sigma^2 R)^{-1}R', \quad (C3)$$

it is necessary to know that

$$L_k \xi_j = L_{k(s)} L_{j(t)} E_m(t, s).$$

This is not hard to check from the definitions. Then one has

$$L_k u_{\nu, m, \lambda} = \sum_{j=1}^N c_j L_k \xi_j + \sum_{r=1}^M d_r L_k \phi_r,$$

$$k = 1, 2, \dots, N,$$

or

$$\begin{pmatrix} L_1 u_{\nu, m, \lambda} \\ L_2 u_{\nu, m, \lambda} \\ \vdots \\ L_N u_{\nu, m, \lambda} \end{pmatrix} = Kc + Td,$$

and by the definition of $A_m(\lambda)$, we have

$$Kc + Td = A_m(\lambda)z,$$

Thus,

$$\{I - A_n(\lambda)\}z = z - Kc - Td \quad (D4)$$

But from (2.8)

$$z - Kc - Td = N\lambda D_\sigma \bar{c}. \quad (D5)$$

Substituting (C1) into (D5) and the result into (D4) gives (C3).

We now give a brief argument why the $N - M \times N - M$ matrix $B = R'KR$ is always strictly positive definite. Let K_0 and R_0 be the special cases of K and R when $L_\lambda u = u(t_\lambda)$. Duchon (1976b) has shown in this case that $R_0'K_0R_0$ is always strictly positive definite for any $N \geq M$. By using the fact that all continuous linear functionals in a reproducing kernel Hilbert space are limits of sums of evaluation functionals, one can show the positive definiteness in general [see Dyn and Wahba (1979) for more details].

APPENDIX E

500 mb Height Model

As mentioned earlier, the height field used in the numerical experiments is the same as that used by T. Koehler. Koehler adopted the model of Sanders to represent meteorological phenomena of interest (in particular, we used pressure surfaces) over an area the size of North America. In his model the height z of any pressure surface p at longitude θ and latitude ϕ is defined as follows:

$$\begin{aligned} z(\theta, \phi, p) = & \bar{z} \cos\{(2\pi/L_\theta)(\theta_0 - \theta + \Delta\theta)\}G'(\phi) + \bar{z} \\ & + [T_m(1000)/\gamma]\{1 - (p/1000)^{R\gamma/g}\} \\ & - (R/g)\{\ln(1000/p) - (\alpha/2)[\ln(1000/p)]^2\} \\ & \times \{(a r/\sin\phi_0)(\cos\phi_0 - \cos\phi) \\ & + \hat{T} \cos\{(2\pi/L_\theta)(\theta_0 - \theta)\}G(\phi)\}, \end{aligned}$$

where

$$\begin{aligned} T_m(1000) &= 278 \text{ K} \\ \hat{T} &= 10 \text{ K} \\ \bar{z} &= 150 \text{ m} \\ \bar{z} &= 90 \text{ m} \\ R\gamma/g &= 0.0953 \\ a &= 0.9 \times 10^{-5} \text{ K m}^{-1} \\ r &= 6371 \text{ km} \\ \Delta\theta &= 9^\circ \\ L_\theta &= 30^\circ \\ \phi_0 &= 45^\circ \\ \alpha &= 0.621 \\ R &= 287.04 \text{ m}^2 \text{ s}^{-2} \text{ K}^{-1} \\ g &= 9.8 \text{ m s}^{-2} \\ p &= 500 \text{ mb} \\ \theta_0 &= \text{ALON}_0 \end{aligned}$$

Also

$$\begin{aligned} G(\phi) = & b \left[\frac{18}{\pi} (\phi - \phi_0) \right]^n + c \left[\frac{18}{\pi} (\phi - \phi_0) \right]^2 \\ & + d \left[\frac{18}{\pi} (\phi - \phi_0) \right]^2 + e, \end{aligned}$$

with

$$b = -1/60, \quad d = -10/60,$$

$$c = 11/60, \quad e = 1,$$

and

$$G'(\phi) = \int_{\phi_0}^{\phi} \frac{\sin\phi'}{\sin\phi_0} \frac{\partial G(\phi')}{\partial \phi'} d\phi'.$$

In the numerical experiments the parameter ALON_0 was varied taking the values 105, 100, 95 and 90. This parameter determined the longitude at which the wave "begins". Hence, by decreasing ALON_0 the wave "moves" from west to east. For the physical interpretation of the other constants and functional form of the model the reader is referred to Koehler (1979).

REFERENCES

- Achtemeier G L. 1975 On the initialization problem. A variational adjustment method. *Mon Wea Rev.* 103, 1089-1103.
- Air Weather Service. 1955 Accuracies of radiosonde data TR 105-133. Headquarters Air Weather Service Military Air Transport Service. USAF Washington DC 12 pp.
- Akhiezer N I and I M Gluzman. 1961 *Theory of Linear Operators in Hilbert Space* Vol 1 147 pp [Translated from the Russian by Merlind Nestell. Ungar New York].
- Aronszajn, N. 1950 Theory of reproducing kernels. *Trans Amer Math Soc.* 58, 337-404.
- Craven, P. and G Wahba. 1979 Smoothing noisy data with spline functions. estimating the correct degree of smoothing by the method of generalized cross-validation. *Numer Math.* 3, 377-403.
- Duchon, Jean. 1976a Interpolation des fonctions de deux variables suivant le principe de la flexion des plaques minces. *RAIRO Anal Numer.* 10, 5-12.
- . 1976b Fonctions spline du type Plaque mince en dimension 2, No. 231. *Seminaire d'Analyse Numerique Mathematiques Appliquees*. Universite Scientifique et Medicale de Grenoble.
- Dyn, N. and G Wahba. 1979 On the estimation of functions of several variables from aggregated data. MRC Tech Summary Rep No. 1974. Mathematics Research Center, University of Wisconsin, Madison. 34 pp.
- Fritsch, J M. 1971 Objective analysis of a two dimensional data field by the cubic spline technique. *Mon Wea Rev.* 99, 379-386.
- Fritz S, D Q Wark H E Fleming W L Smith H Jacobowitz, D T Hilleary and J C Alishouse. 1972 Temperature sounding from satellites. NOAA Tech Rep NES-59, 49 pp.
- Golub G, M Heath and G Wahba. 1979 Generalized cross-validation as a method for choosing a good ridge parameter. *Technometrics*, 21, 215-223.
- Holl M. 1976 The upper-air analysis capability FIB/UA introducing weighted spreading. Project M 213, Meteorology Int., Inc., Monterey CA 32 pp.
- Kimeldorf, G., and G Wahba. 1970 A correspondence between

AUGUST 1980

GRACE WAHBA AND JAMES WENDELBERGER

1143

- Bayesian estimation on stochastic processes and smoothing by splines *Ann Math Statist*, 41, 495-502
- , and —, 1971 Some results on Tchebycheffian spline functions *J Math Anal Appl*, 33, 82-95
- Koehler, T., 1979 A case study of height and temperature analysis derived from Nimbus 6 satellite soundings on a fine mesh model grid Ph.D. thesis, Dept. of Meteorology, University of Wisconsin-Madison, 186 pp
- Kreiss, H. O., 1979a Problems with different time scales for ordinary differential equations University of Uppsala Computer Sciences Department Tech Rep No 68
- , 1979b Problems with different time scales for partial differential equations University of Uppsala Computer Sciences Department Tech Rep No 75
- Lewis, J. M., 1972 An operational upper air analysis *Tellus*, 24, 514-530
- , and T. H. Grayson 1972 The adjustment of surface wind and pressure by Sasaki's variational matching technique *J Appl Meteor*, 11, 586-597
- Meinguet, Jean 1978 Multivariate interpolation at arbitrary points made simple Rep No 118 Institute de Mathematique Pure et Appliquee Universite Catholique de Louvain (to appear in *Z Agnew Math Phys*)
- , 1979 An intrinsic approach to multivariate spline interpolation at arbitrary points *Proceedings of the NATO Advanced Study Institute on Polynomial and Spline Approximation*, B. Sahney Ed., Calgary 1978 (in press)
- Nitta, T., and J. Høvermale 1969 A technique of objective analysis and initialization for the primitive forecast equations *Mon Wea Rev*, 97, 652-658
- Paihua Montes 1978 Quelques methodes numeriques pour le calcul de fonctions splines a une et plusieurs variables Thesis Analyse Numerique Universite Scientifique et Medicale de Grenoble 171 pp
- Reinsch, G., 1967 Smoothing by spline functions *Numer Math*, 10, 177-183
- Sanders, F., 1971 Analytic solutions of the nonlinear omega and vorticity equation for a structurally simple model of disturbance in the baroclinic westerlies *Mon Wea Rev*, 99, 393-407
- Sasaki, U., 1960 An objective analysis for determining initial conditions for the primitive equations Tech Rep 208 Department of Oceanography and Meteorology A&M College of Texas 555 pp
- , 1971 A theoretical interpretation of anisotropically weighted smoothing on the basis of numerical variational analysis *Mon Wea Rev*, 99, 698-708
- Smith, V. T., J. M. Boyle, B. S. Garbow, Y. Ikebe, V. C. Klema and C. B. Moler, 1976 *Matrix Eigensystem Routines-EISPACK Guide* Springer Verlag, 387 pp
- Thiebaux, H. J., 1977 Extending estimation accuracy with anisotropic interpolation *Mon Wea Rev*, 105, 691-699
- Wagner, K., 1971 Variational analysis using observational and low pass filtering constraints MS thesis, The University of Oklahoma, Norman, 39 pp
- Wahba, G., 1975 Smoothing noisy data with spline functions *Numer Math*, 25, 383-393
- , 1978a Automatic smoothing of the log spectral density University of Wisconsin-Madison, Tech Rep 536 To appear in *J Amer Statist Assoc*
- , 1978b Improper priors, spline smoothing and the problem of guarding against model errors in regression *J Roy Statist Soc*, B40, 364-372
- , 1979a How to smooth curves and surfaces with splines and cross validation *Proceedings of the 24th Design of Experiments Conference* US Army Research Office, Rep 79 2, 167-192
- , 1979b Convergence of 'thin plate' splines when the data are noisy *Smoothing Techniques for Curve Estimation*, T. Gasser and M. Rosenblatt, Eds., *Lecture Notes in Mathematics*, No 757 Springer-Verlag 232-245
- , 1979c Spline interpolation and smoothing on the sphere Dept. of Statistics Tech Rep No 584, University of Wisconsin, Madison, 78 pp.
- , 1980a Spline bases regularization and generalized cross validation for solving approximation problems with large quantities of noisy data Dept. of Statistics, Tech Rep No 597, University of Wisconsin-Madison 8 pp [To appear in *Proceedings of the International Conference on Approximation Theory in Honor of George Lorenz*, January 1980 Austin, TX Ward Cheney, Ed. Academic Press]
- , 1980b Ill-posed problems Numerical and statistical methods for mildly, moderately and severely ill-posed problems with noisy data Dept. of Statistics Tech Rep No 595, University of Wisconsin-Madison 69 pp [To appear in the *Proceedings of the International Conference on Ill-Posed Problems* Newark, DE October 1979, M. Z. Nashed, Ed., Academic Press]
- Wendelberger, J., 1980 Smoothing noisy data using multivariate splines and generalized cross validation Ph.D. thesis, Dept. of Statistics, University of Wisconsin

**END
DATE
FILMED**

MAR 24 1983

End of Document