NASA TECHNICAL REPORT

NASA TR R-332

c.1

NASA TR R-332

SIMULTANEOUS ESTIMATION OF THE STATE AND NOISE STATISTICS IN LINEAR DYNAMICAL SYSTEMS

by Paul D. Abramson, Jr.

Electronics Research Center
Cambridge, Mass.

NATIONAL AERONAUTICS AND SPACE ADMINISTRATION • WASHINGTON, D. C. • MARCH 1970

1. Report No. NASA TR R-332	2. Government Accession No.	3. Recipient's Catalog No.
4. Title and Subtitle Simultaneous Estimation of the State and Noise Statistics in Linear Dynamical Systems		5. Report Date March 1970 6. Performing Organization Code
7. Author(s) Paul D. Abramson, Jr.		8. Performing Organization Report No. C-88
9. Performing Organization Name and Address Electronics Research Center Cambridge, Massachusetts		10. Work Unit No. 125-17-05-24 11. Contract or Grant No.
12. Sponsoring Agency Name and Address National Aeronautics and Space Administration Washington, D. C. 20546		13. Type of Report and Period Covered Technical Report 14. Sponsoring Agency Code
15. Supplementary Notes Submitted by author to Massachusetts Institute of Technology as thesis for Doctor of Science degree (May 10, 1968)		

16. Abstract

An optimal procedure for estimating the state of a linear dynamical system when the statistics of the measurement and process noise are poorly known is developed. The criterion of maximum likelihood is used to obtain an optimal estimate of the state and noise statistics. These estimates are shown to be asymptotically unbiased, efficient, and unique, with the estimation error normally distributed with a known covariance. The resulting equations for the estimates cannot be solved recursively, but an iterative procedure for their solution is presented. Several approximate solutions are presented which reduce the necessary computations in finding the estimates. Some of the approximate solutions allow a real time estimation of the state and noise statistics.

Closely related to the estimation problem is the subject of hypothesis testing. Several criteria are developed for testing hypotheses concerning the values of the noise statistics that are used in the computation of the appropriate filter gains in a linear Kalman type state estimator. If the observed measurements are not consistent with the assumptions about the noise statistics, then estimation of the noise statistics should be undertaken using either optimal or suboptimal procedures.

Numerical results of a digital computer simulation of the optimal and suboptimal solutions of the estimation problem are presented for a simple but realistic example.

17. Key Words •Statistics (state and noise) •Optimal Procedure •Linear Dynamical System	18. Distribution Statement Unclassified-Unlimited

19. Security Classif. (of this report)	20. Security Classif. (of this page)	21. No. of Pages	22. Price *
Unclassified	Unclassified	343	$3.00

* For sale by the Clearinghouse for Federal Scientific and Technical Information
Springfield, Virginia 22151

SIMULTANEOUS ESTIMATION OF THE STATE AND NOISE STATISTICS IN LINEAR DYNAMICAL SYSTEMS*

By Paul D. Abramson, Jr.
Electronics Research Center

SUMMARY

An optimal procedure for estimating the state of a linear dynamical system when the statistics of the measurement and process noise are poorly known is developed. The criterion of maximum likelihood is used to obtain an optimal estimate of the state and noise statistics. These estimates are shown to be asymptotically unbiased, efficient, and unique, with the estimation error normally distributed with a known covariance. The resulting equations for the estimates cannot be solved recursively, but an iterative procedure for their solution is presented. Several approximate solutions are presented which reduce the necessary computations in finding the estimates. Some of the approximate solutions allow a real time estimation of the state and noise statistics.

Closely related to the estimation problem is the subject of hypothesis testing. Several criteria are developed for testing hypotheses concerning the values of the noise statistics that are used in the computation of the appropriate filter gains in a linear Kalman type state estimator. If the observed measurements are not consistent with the assumptions about the noise statistics, then estimation of the noise statistics should be undertaken using either optimal or suboptimal procedures.

Numerical results of a digital computer simulation of the optimal and suboptimal solutions of the estimation problem are presented for a simple but realistic example.

*Submitted to the Department of Aeronautics and Astronautics, Massachusetts Institute of Technology, on May 10, 1968, in partial fulfillment of the requirements for the degree of Doctor of Science.

TABLE OF CONTENTS

LIST OF ILLUSTRATIONS

LIST OF TABLES

Chapter 1

INTRODUCTION

1.1 Statement and Discussion of the Problem

Optimal estimation has received considerable attention in recent years in fields such as space navigation, statistical communication theory, and many others that often require the estimation of certain variables that are either not directly measurable or are being measured with instruments that are not sufficiently accurate for an adequate deterministic solution. In essence the procedures aim at reducing the effects of random disturbances associated with these "imperfect' instruments.

In many situations, the estimation procedure consists of no more than averaging repeated measurements of the "same" quantity made with the same or different instruments. In this way, the random errors made in each measurement might "average out," resulting in a higher confidence in the value of the quantity being measured than would be the case if only a single measurement was taken. In this type of operation, the improved confidence in the estimate depends upon the fact that the "same" quantity being measured is truly time invariant.

In more complex situations, the quantity being measured might change from one measurement time to another. Suppose it is known that the voltage across an electrical network decreases exponentially with time. A simple average of repeated measurements of the voltage made at different times would lead to an erroneous estimate. However, if the time constant associated with the exponential decay is known, then each measured voltage can be related to the voltage at any specified time. These computed voltages can then be averaged to obtain an estimate of the voltage at the specified time.

The examples illustrated above represent the most simple case of estimation in which each measurement carries the same weight so that simple linear averaging of the measurements is performed to obtain the estimate. However, if each measurement has associated with it a different confidence, usually characterized by the variance of the measurement error, then a more complicated estimation scheme must be employed which takes into account the differing accuracies of the measurements. Typical examples of this situation are: 1) when two or more different types of instruments are used to measure the same quantity, or 2) in the case of the previous example when there is some random characteristic in the exponential function of the voltage being measured. This leads to a reduction in the confidence in relating measurements made at some time distant from the specified time.

Operational or computational procedures involving a consideration of the variances of the various noises in the problem represent the first degree of sophistication in estimation. Various formulations have been advanced which characterize the statistical nature of the problem in some orderly pattern. There are two widely used techniques for optimal estimation when the time variation of the quantity being measured can be described by a linear differential equation and when the measurements are linearly related to the quantity being estimated. The initial significant work on this problem was by Wiener (Ref. 36) who developed the condition to be satisfied for optimal estimation in the least mean-squared-error sense. This condition is generally referred to as the Wiener-Hopf integral equation. He also developed the solution for the case of a time invariant system with stationary noise processes. This work and further extensions and modifications by others are known as Wiener filters.

In the Wiener filter, the measurement information is acknowledged to have a signal and a noise component. The filter, which is usually implemented as a linear analog filter, is designed so that the noise component of the measurement is more heavily attenuated than the signal component, thus allowing extraction of as much information from the measurement as is possible. However, non-time stationary, transient, or multiple input-output problems are difficult to solve by the Wiener approach.

Kalman (Ref. 16) treated the estimation problem from a different point of view and formulated the equivalent of the Wiener-Hopf integral equation as a vector-matrix differential equation in state space. He developed the solution for a linear system with normally distributed noises as a set of vector-matrix difference equations which are commonly termed the "Kalman filter." Information about the dynamics of the process being measured, statistics of the disturbances involved, and a priori knowledge of the quantities being estimated are included in the formulation of the problem.

In the Kalman filter, the estimation proceeds from any chosen starting time and is well suited for situations dominated by a transient mode, such as the launching of a space vehicle. In the steady state, the Kalman filter can be shown to be equivalent to a Wiener filter and thus can be considered as a more general formulation of the estimation problem. Further advantages of the Kalman filter are that the computations are performed recursively, in the time domain, and are readily applicable to nonstationary and multiple input-output systems. In the standard formulation of the Kalman estimation procedure, allowance is made for a variation of the noise variances with respect to time. However, this knowledge is assumed to be known prior to the actual filter operation. In an operational situation, the time varying filter gains can be precomputed and stored in the filter to be used in conjunction with the measurement information to obtain the optimal estimate. As an estimation procedure of the first degree of sophistication, i.e., with the consideration of the noise variances, this is indeed a very powerful and generally applicable procedure.

Kalman filtering can be thought of as a method of combining in an optimal fashion all information up to and including the latest measurement to provide an estimate at that time. The proper weighting to apply to the new measurement is determined by the relative "quality" of the new information as compared to the information contained in the estimate before the latest measurement. Poor measurements will receive less weight than good ones. If there is noise driving the system between measurement times, the filter will weight the extrapolated value of the old estimate less than if there were no noise. This is because noise introduces an uncertainty in the state of the system between measurement times. Consequently the estimate will depend less upon old estimates and more upon new measurements. The appropriate measures of the "quality" of the old estimate and the new measurement are respectively the covariance of the old estimation error and the covariance of the new measurement error.

These important points can be clarified by considering the following simple example. Let x_n represent the scalar state of a system at time "n." If the system can be described by a linear differential equation, then the state at time "n" can be related to the state at time "n-1" by the difference equation

$$x_n = \Phi(n,n-1)\ x_{n-1} + \Gamma_n\ w_n$$

$\Phi(n,n-1)$ is the state transition matrix and extrapolates the state from time n-1 to time n if the effects of w_n are ignored. Γ_n is the "forcing function matrix" and w_n is the state "driving noise" which is assumed to be a zero mean uncorrelated normally distributed noise with variance Q_n. Let $\hat{x}_{n|n-1}$ represent the estimate of x_n obtained after processing n-1 measurements and let $P_{n|n-1}$ represent the variance of the estimation error after n-1 measurements.

The measurement at time n is given by

$$z_n = H_n x_n + v_n$$

where v_n is additive noise representing the error in the measurement and H_n is the "observation matrix" which relates the measurement to the state. In this example, z_n is a scalar and $H_n = 1$. It is assumed that v_n is a zero mean uncorrelated normally distributed noise with variance R_n.

The scalar Kalman filter equation for incorporating this new measurement into the state estimate is given by

$$\hat{x}_{n|n} = \frac{R_n}{R_n + P_{n|n-1}} \hat{x}_{n|n-1} + \frac{P_{n|n-1}}{R_n + P_{n|n-1}} z_n$$

The variance of the estimation error after incorporation of the new measurement is given by

$$P_{n|n} = \frac{R_n}{R_n + P_{n|n-1}} P_{n|n-1}$$

If the state estimate $\hat{x}_{n|n-1}$ is very good compared with the information contained in z_n, then

$$P_{n|n-1} << R_n$$

and thus $$\hat{x}_{n|n} \simeq \hat{x}_{n|n-1}$$

and $$P_{n|n} \simeq P_{n|n-1}$$

In this case, the measurement datum is effectively rejected because it is so noisy that it is virtually useless. Since no new information has been added, the variance of the estimation error remains the same after the measurement.

In the other extreme case, suppose $\hat{x}_{n|n-1}$ is of very poor quality compared with the information contained in z_n. Then:

$$P_{n|n-1} >> R_n$$

and thus $$\hat{x}_{n|n} \simeq z_n$$

and $$P_{n|n} \simeq R_n$$

In this case, the estimate $\hat{x}_{n|n-1}$ is effectively rejected and the estimate $\hat{x}_{n|n}$ is based upon the single measurement z_n. In all cases falling between these two extremes, the estimate $\hat{x}_{n|n}$ is a linear combination of the old estimate $\hat{x}_{n|n-1}$ and the new measurement z_n.

Before computing the proper weighting factors given above, the variance of the state estimation error before the measurement at time n must be found. This can be done by studying how the actual state changes between time n-1 and time n and how the state estimate changes in this same time interval. Let $\hat{x}_{n-1|n-1}$ be the estimate of the state x_{n-1} after the measurement at time n-1. Since w_n is a zero mean independent random variable, the best estimate of the state at time n based upon the n-1 measurements is given by

$$\hat{x}_{n|n-1} = \Phi(n,n-1)\ \hat{x}_{n-1|n-1}$$

If $P_{n-1|n-1}$ represents the covariance of the estimation error at time n-1, it can be seen that

$$P_{n|n-1} = \Phi^2(n,n-1)\ P_{n-1|n-1} + \Gamma_n^2\ Q_n$$

A large driving noise variance will cause a large increase in the mean squared error in the estimate when it is extrapolated from one measurement time to the next.

The filter equations given above are for the case of a scalar state and measurement. In Chapter 2, the more general case of a vector state and measurement is treated. However, even in more complicated situations, the same interpretation can be applied to the operation of the filter. The primary purpose of the filter is to compute and apply the proper weighting factors so that the new measurement information can be incorporated with an old estimate of the state to provide a combined and improved state estimate.

Precise knowledge of the measurement and driving noise statistics is of fundamental importance in the operation of a Kalman filter. However, in any operational situation, the statistics of the noises that are used in the filter are in fact only estimates or predictions of the statistics of the noises that will actually be encountered. In some cases these estimates might be quite accurate, but in other cases they may be sufficiently in error to adversely affect the filter. One effect of this can be a large discrepency between the state estimation error covariance matrix as computed within the filter and the "actual" state estimation error covariance. If there is a difference between the computed and actual covariance of the old state estimate, the filter can make an error in computing the weighting for a new measurement. This subject is treated fully in Chapter 2 but it can be understood by considering the following example.

Suppose that it is assumed that there is no noise driving the state when in fact driving noise is present. Then the computed covariance of the state estimation error will generally be smaller than the actual estimation error covariance. This is because the driving noise introduces an error in extrapolating the state estimate from one measurement time to the next which is not accounted for in the computed state estimation error covariance matrix. The filter "thinks" it is doing a better job of estimating the state than is actually the case. If the filter thinks the old state estimate is much better than it actually is, it may assign little weight to new measurement information and thus effectively discard this new information. Of course, this is exactly the wrong thing to do. The old state estimate may be of very poor quality so that the new measurement information should be weighted quite heavily. However, in its ignorance, the filter fails to do this and as a result the actual estimation error may become very large while the filter "thinks" it is doing a good job of estimating the state.

A similar problem can arise in the case of vector measurements. If the relative quality of the different measurements is not well known, then more weight might be given to a measurement taken with an inaccurate instrument than to a very accurate one. This would lead to a greater estimation error than would be the case if the relative accuracy of the different measurements was known and the proper weighting assigned to each.

A priori estimates of the statistics of the noises can be obtained in several ways. They may be no more than educated guesses as to what noise environment may actually exist. It is often very difficult to predict with accuracy the operating conditions of a complicated and interrelated system, especially in research and development applications when little may be known before an experiment is conducted.

Another technique for obtaining the statistics of the noises is the analysis of previous experiments. These experiments may have been conducted in an operational environment or in the controlled environment of a laboratory. In either case, it is rarely possible to have complete confidence in the estimates of the noise statistics due to the necessarily finite number of experiments that can be performed and possible problems associated with the inability to isolate and distinguish the various effects of the different noises. And there is still a question as to whether the environment will remain constant between the time these estimates of the statistics are obtained and when the estimates are subsequently used in the Kalman filter.

Thus in many situations, the assumption that the a priori estimates of the statistics of the measurement and driving noises are good estimates may not be justified. The primary objective of this work is to develop an optimal estimator of the state that remains optimal when the statistics of these noises are not precisely known a priori. In the process of estimating the state under these conditions, optimal estimates of the measurement and driving noise statistics are also obtained.

In developing optimal estimators for the state and noise statistics, it is not assumed that the statistics of the noises are known precisely a priori. Instead, it is assumed that the uncertainty in knowledge of these statistics has a particular distribution about some a priori value. This is completely analogous to the usual assumption made in Kalman filtering that the initial state of the system is not known precisely, but rather the uncertainty in knowledge of the state can be described by a suitable probability density function. In both cases, it is assumed that the distribution of the uncertainty is known a priori. This represents the second degree of sophistication is estimation procedures. It reduces by one level the necessary specification of the values of the noise statistics. Instead of having to specify their exact values, all that need be specified is the possible distribution these values might have. In fact, it will subsequently be shown that the exact shape of this distribution is relatively unimportant when a large number of measurements has been taken.

The above discussion can be clarified by considering the following simple example. It will be shown that the a priori estimates of the noise statistics can be improved at the same time that the state is being estimated. All measurements contain some information about the noises as well as the state, whether these measurements are taken in the laboratory or in an operational environment. So a procedure can be devised to utilize this information about the noises actually encountered to improve our knowledge of the noise statistics.

Suppose the state that is to be estimated is a time invariant scalar and the measurements of the state are given by

$$z_n = x + v_n$$

where x is the constant state and v_n is a zero mean independent normally distributed measurement noise with time invariant variance R. If a single measurement is taken, the optimal estimate of the state x is given by

$$\hat{x}_{1|1} = z_1$$

and the variance of the state estimation error is given by

$$P_{1|1} = R$$

If repeated measurements are performed, it is easy to show that the optimal estimate of the state after the n^{th} measurement is given by

$$\hat{x}_{n|n} = \frac{1}{n} \sum_{j=1}^{n} z_j$$

and the variance of the state estimation error is

$$P_{n|n} = \frac{1}{n} R$$

Thus increasing the number of measurements decreases the variance of the estimation error by the factor (1/n). Note that in this simple example, the measurement noise variance is not needed to define the optimal estimate of the state. This is a consequence of the fact that if the actual measurement noise variance is assumed to be time invariant, and if there is no a priori information about the state, then all measurements are given the same weight, regardless of the actual value of R. However, the variance of the state estimation does depend upon the actual value of R as given above. In more complicated situations, such as vector measurements or when there is noise driving the state, the optimal state estimate does depend upon the relative sizes of the noise covariances involved. But in this case, only the variance of the state estimation error depends upon R.

If the value of R is unknown, its value can be estimated from the measurements themselves. In the above case, when the true state is time invariant, an estimate of R can be defined by

$$\hat{R}_n \triangleq \frac{1}{n-1} \sum_{j=1}^{n} (z_j - \hat{x}_{n|n})^2$$

It is easy to show that such an estimate is an unbiased estimate of the noise variance. The expected value of $\hat{R}_n$ is given by

$$\varepsilon(\hat{R}_n) = \frac{1}{n-1} \sum_{j=1}^{n} \varepsilon[(z_j - \hat{x}_{n|n})^2]$$

where ε( ) represents an average over the ensemble of all possible measurement noises with covariance R. It can be seen that

$$z_j - \hat{x}_{n|n} = v_j - \tilde{x}_{n|n}$$

where $$\tilde{x}_{n|n} = \hat{x}_{n|n} - x$$

and $$\tilde{x}_{n|n} = \frac{1}{n} \sum_{k=1}^{n} v_k$$

So $$(z_j - \hat{x}_{n|n})^2 = v_j^2 + \frac{1}{n} 2 \sum_{k=1}^{n} \sum_{s=1}^{n} v_k v_s - \frac{2}{n} \sum_{k=1}^{n} v_j v_k$$

and $$\varepsilon[(z_j - \hat{x}_{n|n})^2] = R + \frac{1}{n} R - \frac{2}{n} R = \frac{n-1}{n} R$$

In obtaining the above expression, use was made of the independence of the measurement noises at different times. Then

$$\varepsilon(\hat{R}_n) = \frac{1}{n-1} \sum_{j=1}^{n} (\frac{n-1}{n} R) = R$$

It can be shown that the variance of the R estimation error is given by

$$\varepsilon[(\hat{R}_n - R)^2] = \frac{2 R^2}{n-1}$$

Thus as the number of measurements increases, the variance of the noise variance estimation error becomes small and $\hat{R}_n$ becomes an arbitrarily good estimate of the actual measurement noise variance.

With an estimate of R, an estimate of the state estimation error variance can be obtained.

$$\hat{P}_{n|n} = \frac{1}{n} \hat{R}_n$$

As was mentioned before, in most cases some estimate of the measurement noise variance is available before the above measurements are taken. Suppose an estimate of R is obtained from a series of measurements and it differs from an a priori value obtained by some other means. Now the question is which value more accurately represents the variance of the measurement noise, the a priori value or the value obtained from the measurements. The concepts of relative weighting discussed in connection with Kalman state estimation offer a solution to this problem.

There is usually some measure of accuracy associated with the a priori estimate of R. This measure is often the variance of possible deviations of the actual value of R about the a priori estimate. If it is felt that the a priori estimate is highly accurate, the variance about the true value would be small. Conversely, if it is felt that the a priori estimate of R is highly inaccurate, the variance about the true value of R would be large.

A combined estimate of the measurement noise variance can by defined by

$$\hat{R}_n^c = \frac{\sigma_{R_n}^2}{\sigma_{R_n}^2 + \sigma_{R_o}^2} \hat{R}_o + \frac{\sigma_{R_o}^2}{\sigma_{R_n}^2 + \sigma_{R_o}^2} \hat{R}_n$$

where $\hat{R}_o$ is the a priori estimate of R, $\hat{R}_n$ is the estimate obtained from the measurements, $\sigma_{R_o}^2$ is the variance of the true value of R about the a priori estimate, and $\sigma_{R_n}^2$ is the variance of the true value about the estimate $\hat{R}_n$. $\sigma_{R_n}^2$ is given by

$$\sigma_{R_n}^2 = \varepsilon[(\hat{R}_n - R)^2] = \frac{2R^2}{n-1}$$

In order to compute $\sigma_{R_n}^2$, the true value of R must be known. However, for moderately large n, the approximation can be made

$$\sigma_{R_n}^2 \simeq \frac{2\ \hat{R}_n^2}{n-1}$$

By analogy to the state estimation problem, a measure of the variance of the combined estimate of the measurement noise variance is given by

$$\sigma_{R_n^c}^2 = \frac{\sigma_{R_n}^2}{\sigma_{R_n}^2 + \sigma_{R_o}^2} \sigma_{R_o}^2$$

If the a priori estimate $\hat{R}_o$ is of high accuracy compared with $\hat{R}_n$, then

$$\sigma^2_{R_o} << \sigma^2_{R_n}$$

and thus $$\hat{R}^c_n \simeq \hat{R}_o$$

and $$\sigma^2_{R^c_n} \simeq \sigma^2_{R_o}$$

If the a priori estimate is of low accuracy compared with $\hat{R}_n$, then

$$\sigma^2_{R_o} >> \sigma^2_{R_n}$$

and thus $$\hat{R}^c_n \simeq \hat{R}_n$$

and $$\sigma^2_{R^c_n} \simeq \sigma^2_{R_n}$$

In all cases falling between these two extremes, the estimate $\hat{R}^c_n$ is a linear combination of the a priori estimate and the estimate obtained from the measurements.

Of course, the situation is not always as simple as in the previous example. The state may be a time-varying vector with additive driving noise. The measurements may be vectors indicating that several measurement devices of possibly differing accuracies are used to measure the state at any time. In such cases, the problem is simultaneously estimating the state and the noise covariances becomes much more complicated.

The resulting equations for optimal estimates of the state and noise statistics are generally coupled nonlinear equations that must be solved by some numerical procedure. But the essence of the problem is the same. From the information contained in the measurements taken in an operational environment, improvements can be made in the estimates of not only the state but also the statistics of the measurement and driving noises. The performance of the state estimator in such a situation can be improved compared with the estimator that uses incorrect values of the noise statistics in computing the appropriate filter gains.

1.2 Historical Background

Optimal state estimation when the statistics of the measurement and driving noises are poorly known is but one class of problems within the more general area of state estimation in the presence of "modeling errors." In the formulation of the Kalman filter, it is assumed that the dynamics of the system can be accurately modeled as a set of linear differential or difference equations with precisely known coefficients. This is reflected in the value of the state transition matrix that is used to extrapolate the state estimate from one measurement time to the next. In fact, the modeling of the system might involve approximations. The number of state variables that are necessary to accurately model the system might be so great that the number of computations needed to estimate all of the variables becomes prohibitively large. Often the number of computations can be

... by including only the most significant state variables in the filter model. This will reduce the complexity of the filter but can introduce additional errors in the estimation of the reduced number of state variables.

It may not be possible to model the system dynamics by any set, no matter how large, of linear differential equations. The motion of the state might be described by a set of nonlinear differential equations which can only be approximated by a set of linear differential equations describing the motion of the system about some nominal path. This too can introduce errors in the state estimation that are not accounted for within the model.

There are other sources of modeling error. The elements of the state transition matrices used within the filter may not be accurately known. The actual measurements may be a nonlinear function of the state although it was assumed in the derivation of the filter equations that the measurements are a linear function of the state. These nonlinearities may not be highly significant but they can cause additional state estimation errors.

All of these "modeling errors," including inaccurately known noise statistics, can result in a degradation of the Kalman filter performance.

Many authors have studied the problem of optimal estiation and control of a linear plant whose parameters may ot be accurately known. A comprehensive list of references ı this subject would be prohibitively long. For this ason, the only works cited here are those that have some bearing on the problem of optimal state estimation in the presence of modeling errors.

Spang (Ref. 34) has studied the problem of optimal control of a linear plant with unknown coefficients under the assumptions that there is no measurement noise and the statistics of the noise driving the state are precisely known. He also assumes that the uncertainty in knowledge of the coefficients describing the plant have some distribution of values that can be represented by a probability density function of coefficient values. The optimal control signal which minimizes a quadratic error measure is obtained by finding the conditional mean of the system tracking error, conditioned upon the actual measurements of the system but averaged over the distribution of all possible plant coefficient values. In this way, the error is minimized over the ensemble of all possible trials with systems whose parameters vary in a fashion described by the assigned probability density function. No attempt is made to estimate the actual plant coefficients. Although Spang is concerned primarily with optimal control, several of the concepts he develops have direct application to optimal state estimation when the parameters of the system are unknown.

Drenick (Ref. 8) has also studied this problem. He also assumes that the uncertainty in the parameters of a linear plant can be described by a probability density function whose first two moments are known. His optimal control signal minimizes the conditional mean squared tracking error and is a function of the measurements on the system and the first two moments of the parameter distributions. However, using his procedure, there is no way to estimate the values of the unknown system parameters except in a very restricted set of problems.

Magill (Ref. 21) takes an interesting and rather unique approach to the problem of optimal state estimation when certain statistical parameters of the problem are unknown. These parameters, called the parameter vector, are assumed to come from a finite set of values that are known a priori. The optimal estimator is composed of a set of Kalman type state estimators, with each filter using one of the finite number of parameter vectors to compute the proper measurement gains. The outputs of the filters are weighted and added, with the weighting of each filter output being determined by the conditional probability that the parameter vector being used in that filter is the true parameter vector. These conditional probabilities are functions of the measurements and are obtained by relatively simple but nonlinear calculations.

The following works are primarily concerned with obtaining relatively simple and easy-to-use procedures rather than finding an "optimal" solution to the problem. The approaches to the problem are quite different but there is one common feature. This feature is the real time examination of measurement residuals to determine if a Kalman type state estimator is performing as predicted. The measurement residual is defined as the difference between an actual measurement and the predicted measurement, this prediction being based upon the predicted state at the time of the measurement. If the measurement at time n is given by

$$z_n = H_n x_n + v_n$$

and $\hat{x}_{n|n-1}$ is the estimate of the state x_n before the measurement z_n, then the measurement residual is defined by

$$\Delta z_n = z_n - H_n \hat{x}_{n|n-1}$$

If there are no modeling errors, it is easy to show that Δz_n is a zero mean random variable with covariance

$$\varepsilon(\Delta z_n \ \Delta z_n^T) = R_n + H_n P_{n|n-1} H_n^T$$

where R_n is the covariance of the measurement noise v_n, and $P_{n|n-1}$ is the covariance of the state estimation error before the nth measurement.

Jazwinski (Ref. 15) has suggested introducing into the model of the dynamics of the system a zero mean random driving noise which in some sense can account for the effect of any modeling error. However, the covariance of this noise is not known a priori since it is not known what modeling errors are actually present. Jazwinski proposes a simple and reportedly effective procedure for determining how much "driving noise" to introduce into the model based upon an examination of a single residual at a time. If the squared residual is much larger than predicted by the filter, the computed covariance of the old state estimate is artificially increased at that time so that the new measurement is weighted more heavily than would be the case if no adjustment was made. In this way possible divergence problems in the filter are minimized because as soon as the residuals become large, indicating that there is an error in the model, the measurements are weighted heavily. This tends to reduce the estimation error to a level consistent with that predicted by the filter.

No attempt is made to estimate the value of the covariance of the added driving noise since in fact it does not exist. It was included to account for any unknown modeling errors. Even if the covariance is estimated, such an estimate would have little statistical significance since it would be based upon an examination of a single measurement residual. So Jazwinski's procedure should be viewed as an attempt to reduce the effect of modeling errors on the filter operation rather than an attempt to improve our knowledge of the model.

Dennis (Ref. 5) addresses himself to a more complicated problem, that of estimating the effects of errors in modeling the dynamics of the system as well as estimating the covariances of the measurement and driving noises. Only his procedure for estimating the statistics of the noises is of interest here.

Dennis develops expressions for a real time estimator of the measurement and driving noise covariances. The estimates are subsequently used in the computation of the appropriate weighting gains in a Kalman state estimator. Dennis' solution for the estimation of the noise statistics is suboptimal in the sense that no optimality criterion is used in defining these estimates. The expressions were obtained by an examination of the characteristics of quadratic functions of certain measurement residuals. From this examination a reasonable, if not optimal, estimator is postulated. However, in many useful applications there are several problems associated with the use of this estimator. It is not always possible to estimate all of the unknown elements of the measurement and driving noise covariance matrices. Depending upon the dimension and nature of the measurement, some or all of the elements of the driving noise covariance may not be observable, and as a result, a singular situation is created. There are also certain situations when the estimators may be biased and result in estimates that do not converge to the true values of the noise covariances as the number of measurements becomes large. Dennis does not develop expressions for the evaluation of the quality of the noise covariance estimates. Such measures of quality would be needed if it is desired to incorporate the estimates obtained from the measurements with some a priori estimates to obtain a combined estimate based upon a priori knowledge of the noise covariances and the information contained in the measurements.

Shellenbarger (Ref. 31) is exclusively concerned with estimating the values of the measurement and driving noise covariances so that the proper gains can be computed for estimating the state. His technique is aimed at finding an approximate solution to this problem and consequently his estimator for these parameters is suboptimal. He bases his estimates of the noise covariance parameters upon an examination of a single measurement residual at a time. If the measurement is of small dimension compared with the number of covariance parameters being estimated, there is no unique solution for all of the noise covariance parameters. In addition to this, there is also a question of a possible bias in the noise covariance estimator.

The work of Smith (Ref. 33) is even more restricted in that he attempts to estimate only the measurement noise covariance, assuming that the dynamical model of the state and the covariance of the driving noise are known precisely. His work results in a suboptimal estimator for the state and measurement noise covariance. Here too there is a question of a possible bias in the noise covariance estimator.

Because of the relevance of noise covariance estimation to this work, a short review of the procedures of Dennis, Shellenbarger, and Smith is included in Chapter 4. Although their procedures are suboptimal and there are problems associated with implementing their estimators in certain cases, it is felt that there are some situations when these estimators provide an adequate solution to the problem of inaccurately known noise statistics. Their procedures are much simpler that the optimal procedures developed in Chapter 3 and provide some insight into the variety of techniques that are available for an approximate solution to the problem.

1.3 Summary of Thesis

As was previously mentioned, the primary objective of this thesis is the development of an optimal estimator of the state and statistics of the measurement and driving noises. However, several other related subjects are also treated.

In Chapter 2, it will be shown that a biased or correlated measurement or driving noise can be estimated using a linear recursive filter identical in form to the usual Kalman filter for estimating the state. This is a consequence of the fact that such a biased or correlated noise is observable in terms of a linear function of the measurements. It will also be shown that an error in the values of the measurement and driving noise covariances used to compute Kalman filter gains does not produce an observable effect in a linear function of the measurements. Therefore, any estimator of these covariances is inherently a nonlinear estimator since a nonlinear function of the measurement is needed in the estimation loop. In the simple example given previously, it was shown that an estimator for the measurement noise variance is a quadratic function of the measurements.

Initially an attempt was made to formulate the problem of noise covariance estimation in terms of minimum variance estimation, but the nonlinearities in the problem immediately produced great analytical difficulties. This is one of the reasons why the criterion of maximum likelihood was chosen to define the optimal estimates of the state and the noise statistics. As the name might imply, maximum likelihood estimates are the most probable values of the state and statistics for a given set of measurements. The techniques of maximum likelihood result in complicated equations, but the theory of maximum likelihood estimators is sufficiently developed to allow a proper handling of the problem.

The points of maximum likelihood are found by setting the derivatives of a suitable likelihood function to zero and then solving the resulting equations for the unknown parameters. There is a likelihood equation associated with each parameter being estimated. When the noise covariance matrices are assumed to be time invariant, the solution of the likelihood equations for the optimal state estimate is just a Kalman type estimator which uses the optimal estimates of the noise covariances to compute the appropriate filter gains. Unfortunately, there is no general closed form solution of the likelihood equations for these optimal noise covariance estimates. However, an iterative procedure is proposed for the solution of the likelihood equations corresponding to the estimates of the noise covariances. These estimates are shown to be asymptotically unbiased, efficient, consistent, and unique, with the estimation error normally distributed with a known covariance.

In addition to the optimal solution discussed in Chapter 3, several suboptimal solutions of the problem are given in Chapter 4. These solutions can result in a major savings in the computational requirements but they do not have the wide range of applicability of the optimal solution.

Chapter 5 is devoted to a discussion of hypothesis testing. Hypothesis testing is closely related to the estimation problem. Certain criteria are developed for making decisions as to whether observed measurements are consistent with assumptions about the statistics of the measurement and driving noises. However, the tests themselves do not allow a determination of the reasons the measurements fail a particular hypothesis test, but rather indicate that there is some error in the model of the system and/or measurement. The tests can usually be conducted at less computational expense than a more complicated noise covariance estimation procedure, so they can be used to determine if such additional estimation should be conducted.

In Chapter 6, the numerical results of a computer simulation of the theoretical results are presented. The optimal and suboptimal estimators are simulated to study their performance in a simple but realistic situation. The techniques of hypothesis testing are also studied to find the power of certain tests in detecting errors in the values of the noise statistics used within a Kalman filter.

Chapter 2

EXPECTATION OPERATORS AND MAXIMUM LIKELIHOOD ESTIMATION

2.1 Introduction

In this chapter two types of expectation operators are defined and maximum likelihood parameter estimation discussed. A precise understanding of the expectation operator notation is necessary for subsequent work, so important definitions and results are given here. The maximum likelihood equations are utilized to establish the notation and results of the familiar linear state estimation problem with and without the use of a priori information about the state. The question of unbiasedness and the covariance of the state estimate in the presence of inaccurately known noise statistics is also discussed. More general parameter estimation problems and a more detailed examination of the properties of maximum likelihood estimators are treated in Chapter 3.

2.2 Conditional and Unconditional Expectation Operators

Let x and y be random variables (possibly vector valued) with joint probability density function f(x,y) defined over the range $-\infty < x,y < \infty$. The conditional expectation, or mean, of x, conditioned upon the value of y is defined by

$$\varepsilon(x|y) \triangleq \int_{-\infty}^{\infty} x \, f(x|y) \, dx \qquad (2.2.1)$$

where f(x|y) is the conditional probability density function of x given y. Define

$$f(x) \triangleq \int_{-\infty}^{\infty} f(x,y)\ dy$$

$$f(y) \triangleq \int_{-\infty}^{\infty} f(x,y)\ dx$$

Applying Bayes' rule,

$$f(x|y) = \frac{f(x,y)}{f(y)} \qquad (2.2.2)$$

The unconditional expectation of x is defined by

$$\begin{aligned} E(x) &\triangleq \int_{-\infty}^{\infty} x\ f(x)\ dx \\ &= \int_{-\infty}^{\infty} x\left[\int_{-\infty}^{\infty} f(x,y)\ dy\right] dx \\ &= \int_{-\infty}^{\infty} x\left[\int_{-\infty}^{\infty} f(x|y)\ f(y)\ dy\right] dx \\ &= \int_{-\infty}^{\infty} \left[\int_{-\infty}^{\infty} x\ f(x|y)\ dx\right] f(y)\ dy \\ &= \int_{-\infty}^{\infty} \varepsilon(x|y)\ f(y)\ dy \qquad (2.2.3) \end{aligned}$$

The first expectation, ε(x|y), is the expected value of x if y were fixed at the conditioned value. It is found by averaging over all other random influences with a constant value of y. The second expectation, E(x), is the expected value of x which represents an average over the distribution of y as well as over all other random influences.

The conditional covariance of x is defined by

$$\mathrm{cov}(x|y) \triangleq \varepsilon((x - \varepsilon(x|y))(x - \varepsilon(x|y))^T|y)$$

$$= \varepsilon(x\ x^T|y) - \varepsilon(x|y)\ \varepsilon(x^T|y) \qquad (2.2.4)$$

The unconditional covariance of x is defined by

$$\mathrm{cov}(x) \triangleq E((x - E(x))(x - E(x))^T)$$

$$= E(x\ x^T) - E(x)\ E(x^T) \qquad (2.2.5)$$

But $E(\mathrm{cov}(x|y)) = E(x\ x^T) - E(\varepsilon(x|y)\varepsilon(x^T|y))$

$$= \mathrm{cov}(x) + E(x)\ E(x^T) - E(\varepsilon(x|y)\varepsilon(x^T|y))$$

and $\mathrm{cov}(\varepsilon(x|y)) = E(\varepsilon(x|y)\varepsilon(x^T|y)) - E(x)\ E(x^T)$

so $$\mathrm{cov}(x) = E(\mathrm{cov}(x|y)) + \mathrm{cov}(\varepsilon(x|y)) \qquad (2.2.6)$$

Thus the unconditional covariance can always be decomposed into the sum of two components: 1) the average conditional covariance and 2) the covariance of the conditional average.

The use of the conditional and unconditional expection operators in this work is somewhat unconventional because the random variables y may represent the parameters of the probability density function of x. It is not usual to think of the parameters of a probability density function as themselves being random variables. However, in situations where it is desired to estimate the values of these parameters on the basis of observed values of a random variable x, by considering y to be a random variable any a priori information about the value of y can be utilized coherently in forming an a posteriori estimate of the value of y. It may not seem legitimate to regard the value of y as itself being the outcome of a random experiment. Usually it is more natural to regard y simply as a fixed, though unknown, constant which appears as a parameter in the x distribution from which sample values are taken. However, if this approach is used, there is no way to utilize a priori information about y and accordingly the performance of the estimator would be degraded.

In the extreme case when no a priori information about y exists, then introduction of the concept of an initial distribution for y would be unjustified and of no practical use. In the other extreme case when it is assumed that the parameters are known precisely a priori, then the probability density function of y would reduce to impulses at the known values of the parameters. However, in such a situation, in the absence of any other random influences on y, there would be no need for the entire estimation process since it is assumed that the values of y are known. In all cases falling between these two extremes, by introduction of a realistic if not precisely correct density function for y, the realities of the situation can be more closely modeled than by considering that the parameters y are either exactly known or completely unknown a priori.

The above discussion can be illustrated by a simple example. Let x by a normal variable with mean m and variance s, with conditional probability density function $f(x|m,s)$. Furthermore let m and s be random variables with a joint probability density function $f(m,s)$. For simplicity it is assumed that s and m are independent, so $f(m,s) = f(m)\ f(s)$.

The conditional mean of x is

$$\varepsilon(x|m,s) = \int_{-\infty}^{\infty} x\ f(x|m,s)\ dx$$

But $$f(x|m,s) = \frac{1}{(2\pi s)^{1/2}} e^{-1/2[(x - m)^2/s]}$$

so $$\varepsilon(x|m,s) = m \text{ independent of } s$$

The unconditional mean of x is

$$E(x) = \iint_{-\infty}^{\infty} \varepsilon(x|m,s)\ f(m,s)\ dm\ ds$$

$$= \int_{-\infty}^{\infty} m\ f(m)\ dm \triangleq \overline{m}$$

The conditional variance of x is

$$\varepsilon((x - m)^2|m,s) = \int_{-\infty}^{\infty} (x - m)^2\ f(x|m,s)\ dx = s$$

The unconditional variance of x is

$$E((x - \overline{m})^2) = \iint_{-\infty}^{\infty} \varepsilon((x - \overline{m})^2|m,s)\ f(m,s)\ dm\ ds$$

$$= \iint_{-\infty}^{\infty} (s + m^2 - \bar{m}^2)\ f(m)\ f(s)\ dm\ ds$$

$$= \bar{s} + \overline{m^2} - \bar{m}^2$$

where $$\bar{s} \triangleq \int_{-\infty}^{\infty} s\ f(s)\ ds$$

$$\overline{m^2} \triangleq \int_{-\infty}^{\infty} m^2\ f(m)\ dm$$

Note that $E((x - \bar{m})^2) \neq E((x - m)^2)$ **unless** $\overline{m^2} = \bar{m}^2$.

2.3 Maximum Likelihood State Estimation

In this section the theory of maximum likelihood estimation is discussed and applied to the estimation of the state of a linear dynamical system which is driven by white noise and observed by linear noisy measurements. Because of the relative simplicity of the equations for determining the state estimate, much can be said about the performance of the estimator. In more complicated situations, such as estimating the covariance of the measurement and driving noises, evaluation of the estimator behavior is considerably more difficult and requires a more thorough analysis. For this reason the discussion of these situations is deferred until Chapter 3.

Maximum likelihood estimation, as the name might imply, is concerned with finding the maximum of a likelihood function defined as a function of the parameters being estimated and the measurements on the system. Let Z denote the realized values of a set of measurements and $\alpha^T = (\alpha^1, \alpha^2,..,\alpha^m)$ be the vector of parameters belonging to a set of all possible parameter values Ω. Further, let $f(Z|\alpha)$ denote the conditional probability density function of the measurements Z given the value of the parameter α. The likelihood function is then defined by

$$l(\alpha, Z) = f(Z|\alpha) \tag{2.3.1}$$

The principle of maximum likelihood consists of accepting $\hat{\alpha}^T = (\hat{\alpha}^1, \hat{\alpha}^2, .., \hat{\alpha}^m)$ as the estimate of α^T, where

$$l(\hat{\alpha}, Z) = \max_{\alpha} l(\alpha, Z) \tag{2.3.2}$$

There may be a set of samples for which $\hat{\alpha}$ does not exist. Under suitable regularity conditions on $f(Z|\alpha)$, the frequency of such samples can be shown to be negligible.

In practice it is convenient to work with the natural logarithm of $l(\alpha,Z)$, in which case $\hat{\alpha}$ in (2.3.2) satisfies the equation

$$L(\hat{\alpha}, Z) = \ln l(\hat{\alpha}, Z) = \max_{\alpha} L(\alpha, Z) \tag{2.3.3}$$

When the maximum in (2.3.3) is attained at an interior point of Ω, and $L(\alpha,Z)$ is a differentiable function of α, then the partial derivatives vanish at that point, so that $\hat{\alpha}$ is a solution of the equation

$$\left(\frac{\partial L(\alpha, Z)}{\partial \alpha}\right)_{\hat{\alpha}} = 0 \tag{2.3.4}$$

Equation (2.3.4) is called the maximum likelihood equation and any solution of it a maximum likelihood estimate. The function $\hat{\alpha}$ defined by (2.3.3) over the sample space of observations Z is called a maximum likelihood estimator.

If a priori information about the parameters being estimated exists and if the a priori uncertainty in knowledge of these parameters can be formulated as an a priori probability density function for α, then a slightly different likelihood function can be defined so that this a priori information can be used in an optimal fashion. In such cases, the augmented likelihood function is defined by

$$l^A(\alpha,Z) = f(\alpha|Z) \qquad (2.3.5)$$

where $f(\alpha|Z)$ is the conditional probability density function of the parameters α given the measurements Z. By application of Bayes' rule it can be seen that

$$f(\alpha|Z) = \frac{f(Z|\alpha)\ f(\alpha)}{f(Z)}$$

where $f(\alpha)$ is the a priori probability density function of α and $f(Z)$ is the unconditional probability density function of Z, found by

$$f(Z) = \int_{\Omega} f(Z|\alpha)\ f(\alpha)\ d\alpha$$

In this case the logarithm of the augmented likelihood function (2.3.5) is

$$L^A(\alpha,Z) = \ln\, l^A(\alpha,Z) = \ln\, f(Z|\alpha) + \ln\, f(\alpha) - \ln\, f(Z) \qquad (2.3.6)$$

and $$\frac{\partial L^A(\alpha,Z)}{\partial\alpha} = \frac{\partial L(\alpha,Z)}{\partial\alpha} + \frac{\partial \ln\, f(\alpha)}{\partial\alpha} \qquad (2.3.7)$$

The inclusion of a priori information about α has a tendency to shift the zero points of (2.3.7) towards the peak of the a priori parameter density function. If a priori information about α exists, it is usually preferable to utilize the formulation $f(\alpha|Z)$ since this allows utilization of all information about the value of α, both from the a priori information and information derived from the measurements Z. However, it should be realized that if the assigned a priori probability density function of the parameters does not accurately represent possible variations in the parameters, the performance of the estimator may in fact be degraded by inclusion of a priori information. When studying the performance of an estimator, there is some justification for looking first at an estimator which does not utilize a priori information. This allows determination of how effectively a given estimator extracts information from the measurements without considering how this estimate might be incorporated with an a priori estimate to obtain a combined estimate.

In the derivation of the maximum likelihood state estimation equations, it is first assumed that a priori information about the state does exist so that the latter form of the likelihood function is employed. After the solution of this problem is obtained, the equations for estimating the state without a priori information will be given.

Both solutions of the state estimation equations should more correctly be called conditional maximum likelihood estimates because the optimality of such estimates is conditioned upon the assumption that the noise driving the state and corrupting the measurements of the state have a known distribution with precisely known parameters. If this assumption is not valid, then the state estimates are no longer the true maximum likelihood estimates and all guarantees of optimality are lost.

The purpose of this section is to establish certain results and notation which will be needed in later chapters. An excellent reference on the subject of maximum likelihood state estimation is by Rauch (Ref. 26).

Let the linear dynamical system being observed be defined by the recursive relationship

$$x_k = \Phi(k,k-1)\ x_{k-1} + \Gamma_k\ w_k \quad (\beta x1 \text{ vector}) \qquad (2.3.8)$$

and the linear noisy observations upon the system at time k be defined by

$$z_k = H_k\ x_k + v_k \quad (\gamma x1 \text{ vector}) \qquad (2.3.9)$$

where $\Phi(k,k-1)$ is the $\beta x \beta$ state transition matrix

H_k is the $\gamma x \beta$ observation matrix

Γ_k is the $\beta x \eta$ forcing function matrix

w_k is the $\eta x1$ driving noise vector

v_k is the $\Upsilon x1$ measurement noise vector

For this derivation it is assumed that v_k and w_k are independent zero mean normal random variables with known covariances R_k and Q_k respectively. Using the notation of Section 2.2,

$$\varepsilon(v_k) = \varepsilon(w_k) = 0 \qquad (2.3.10)$$

$$\varepsilon(v_k v_j^T) = R_k \ \delta_{jk}, \ \varepsilon(w_k w_j^T) = Q_k \ \delta_{jk}, \ \varepsilon(v_k w_j^T) = 0 \qquad (2.3.11)$$

where $\delta_{jk} = 1$ if $j = k$ and is zero otherwise.
The above conditional expectation operators are conditioned upon the assumed values of the means and covariances of the noises as well as their assumed independence.

Given the vector of n measurements $Z_n^T = (z_1^T,...,z_n^T)$ and an independent a priori estimate of the initial state, maximum likelihood estimation of the state x_n is based upon finding the particular value of the state which maximizes the conditional probability density function of the state, given all measurements of the state. Implicit in the definition of the likelihood function is that all values of R_k and Q_k, $k = 1,..,n$, be known precisely, as well as the covariance of the a priori state distribution, the elements of the state transition matrices, the observation matrices, and the forcing function matrices. To indicate this dependence of the likelihood function on these parameters, some of the parameters will appear as conditioning variables in the conditional likelihood function. This choice of parameters to thus indicate is motivated by the work of Chapter 3, when the values of certain parameters are to be estimated.

It is convenient to work with the natural logarithm of the likelihood function.

$$L_n^A(x_n, Z_n, R, Q) = \ln f(x_n | Z_n, R, Q) \tag{2.3.12}$$

where R and Q represent the known sequence of values $R_1, \ldots, R_n, Q_1, \ldots, Q_n$, the measurement and driving noise covariances.

The conditional probability density function of the state is found by use of Bayes' rule.

$$f(x_n | Z_n, R, Q) = \frac{f(x_n, Z_n, R, Q)}{f(Z_n, R, Q)}$$

$$= \frac{f(z_n | Z_{n-1}, x_n, R, Q)\ f(Z_{n-1}, x_n, R, Q)}{f(z_n | Z_{n-1}, R, Q)\ f(Z_{n-1}, R, Q)}$$

$$= f(x_n | Z_{n-1}, R, Q)\ \frac{f(z_n | Z_{n-1}, x_n, R, Q)}{f(z_n | Z_{n-1}, R, Q)} \tag{2.3.13}$$

On any one trial, the initial state x_o is not a random variable but assumes a certain value. However, this value is not precisely known. To model this uncertainty in the value of the initial state, x_o is assumed to be a random variable (over the ensemble of all possible initial conditions) having a normal probability density function $f(x_o)$ with mean $\overline{x}_o$ and covariance about the mean $P_{o|o}$. This distribution is presumed to be known a priori. The a priori state estimate is taken to be the mean of this distribution. Because of the symmetry of $f(x_o)$ about its mean, $\overline{x}_o$ is also the point of maximum probability of the distribution.

$$\varepsilon(x_o) = \overline{x}_o$$

$$\varepsilon[(x_o-\overline{x}_o)(x_o-\overline{x}_o)^T] = P_{o|o}$$

$$\hat{x}_{o|o} \overset{\Delta}{=} \overline{x}_o \qquad \text{the a priori state estimate}$$

The averaging here is performed over the ensemble of all possible initial conditions and is conditioned upon knowledge of $\overline{x}_o$ and $P_{o|o}$.

Let $\hat{x}_{n|n-1}$ be the maximum likelihood estimate of x_n immediately before the n^{th} measurement and let $P_{n|n-1}$ be the conditional covariance of x_n about its conditional mean $\hat{x}_{n|n-1}$.

$$P_{n|n-1} = \varepsilon[(x_n-\hat{x}_{n|n-1})(x_n-\hat{x}_{n|n-1})^T|Z_{n-1},R,Q]$$

The averaging here is over the ensemble of all possible measurement and driving noises and initial state conditions, all conditioned upon the values of R and Q. It can be shown that before the update at time n, the conditional probability density function of x_n is

$$f(x_n|Z_{n-1},R,Q) = \frac{e^{-1/2[(x_n-\hat{x}_{n|n-1})^T P_{n|n-1}^{-1}(x_n-\hat{x}_{n|n-1})]}}{(2\pi)^{\beta/2}|P_{n|n-1}|^{1/2}} \qquad (2.3.14)$$

From (2.3.9) $\qquad z_n = H_n x_n + v_n$

Since v_n is a normally distributed variable, independent of x_n, and x_n is also a normal variable, then z_n is a normally distributed variable with conditional mean

$$\varepsilon(z_n|Z_{n-1},x_n,R,Q) = \varepsilon(z_n|x_n) = H_n\ x_n$$

and conditional covariance

$$\varepsilon[(z_n-H_n x_n)(z_n-H_n x_n)^T|Z_{n-1},x_n,R,Q] = \varepsilon(v_n v_n^T|R) = R_n$$

Therefore the conditional probability density function of z_n is

$$f(z_n|Z_{n-1},x_n,R,Q) = \frac{e^{-1/2[(z_n-H_n x_n)^T R_n^{-1}(z_n-H_n x_n)]}}{(2\pi)^{\gamma/2}|R_n|^{1/2}} \qquad (2.3.15)$$

and from (2.3.12) and (2.3.13)

$$L_n^A(x_n,Z_n,R,Q) = \text{constant } -1/2[(x_n-\hat{x}_{n|n-1})^T P_{n|n-1}^{-1}(x_n-\hat{x}_{n|n-1})$$

$$+ (z_n-H_n x_n)^T R_n^{-1}(z_n-H_n x_n)] \qquad (2.3.16)$$

where "constant" includes all terms that are not functions of x_n.

The maximum likelihood estimate of x_n is that value of x_n which maximizes L_n^A, or makes

$$\left(\frac{\partial L_n^A}{\partial x_n}\right)_{x_n \to \hat{x}_{n|n}} = 0 \qquad (2.3.17)$$

It can be seen that

$$\frac{\partial L_n^A}{\partial x_n} = - (x_n-\hat{x}_{n|n-1})^T P_{n|n-1}^{-1} + (z_n-H_n x_n)^T R_n^{-1} H_n \qquad (2.3.18)$$

Then after some manipulation, the solution of (2.3.17) is

$$\hat{x}_{n|n} = (P_{n|n-1}^{-1} + H_n^T R_n^{-1} H_n)^{-1} (P_{n|n-1}^{-1} \hat{x}_{n|n-1} + H_n^T R_n^{-1} z_n) \qquad (2.3.19)$$

Upon using the matrix inversion lemma (see Appendix A)

$$\hat{x}_{n|n} = \hat{x}_{n|n-1} + A_n(z_n - H_n \hat{x}_{n|n-1}) \qquad (2.3.20)$$

where $$A_n = P_{n|n-1} H_n^T (R_n + H_n P_{n|n-1} H_n^T)^{-1} \qquad (2.3.21)$$

A_n is called the optimum gain to the measurement residual $(z_n - H_n \hat{x}_{n|n-1})$.

The conditional probability density function of x_n after the nth measurement can be shown to be

$$f(x_n | Z_n, R, Q) = \frac{e^{-1/2[(x_n-\hat{x}_{n|n})^T P_{n|n}^{-1} (x_n-\hat{x}_{n|n})]}}{(2\pi)^{\beta/2} |P_{n|n}|^{1/2}} \qquad (2.3.22)$$

where $P_{n|n}$ is the conditional covariance of x_n about $\hat{x}_{n|n}$ after the n^{th} measurement. It can be shown that

$$P_{n|n} = (P_{n|n-1}^{-1} + H_n^T R_n^{-1} H_n)^{-1}$$

$$= (I - A_n H_n) P_{n|n-1} (I - A_n H_n)^T + A_n R_n A_n^T \qquad (2.3.23)$$

The necessary quantities for computing $\hat{x}_{n|n}$ can be obtained recursively from the estimate at the previous time.

$$\hat{x}_{n|n-1} = \Phi(n,n-1)\, \hat{x}_{n-1|n-1} \qquad (2.3.24)$$

$$P_{n|n-1} = \Phi(n,n-1) P_{n-1|n-1} \Phi^T(n,n-1) + \Gamma_n Q_n \Gamma_n^T \qquad (2.3.25)$$

It should be noted that the above recursive state estimation equations are identical to those obtained by Kalman (Ref. 16) using the method of orthogonal projections and Lee (Ref. 20) using the method of weighted least squares. It is also easy to show that the state estimate is that estimate which minimizes the conditional covariance of the state estimation error at each stage of estimation.

If no a priori information about the state is used, the logarithm of the likelihood function is defined by

$$L_n(x_n, Z_n, R, Q) = \ln f(Z_n | x_n, R, Q) \qquad (2.3.26)$$

where $f(Z_n|x_n,R,Q)$ is the joint conditional probability density function of the measurements Z_n given x_n, R, and Q.

By application of Bayes' rule

$$f(Z_n|x_n,R,Q) = f(Z_{n-1}|x_n,R,Q)\ f(z_n|Z_{n-1},x_n,R,Q) \qquad (2.3.27)$$

By repeated application of Bayes' rule, it can be shown that

$$f(Z_n|x_n,R,Q) = \prod_{i=1}^{n} f(z_i|Z_{i-1},x_n,R,Q) \qquad (2.3.28)$$

It can be anticipated that until a sufficient number of measurements have been taken, the state estimate cannot be defined and there is no unique solution of the likelihood equations

$$\left(\frac{\partial L_n(x_n,Z_n,R,Q)}{\partial x_n}\right)_{x_n \to \hat{x}_{n|n}} = \left(\frac{\partial \ln f(Z_n|x_n,R,Q)}{\partial x_n}\right)_{x_n \to \hat{x}_{n|n}} = 0 \qquad (2.3.29)$$

The problem is conveniently broken into two parts, obtaining a minimal data set and then subsequent recursive estimation using the equations previously derived. A minimal data set is defined as the smallest set of measurements that is necessary to completely define the state. That is, for $n <$ some n_o, there is no unique solution of the likelihood equations for the state x_n.

The derivation of the estimation equations when no a priori information is used is considerably more complicated than the case previously studied when a priori information was used. Only the results of the derivation will be presented here. Fraser (Ref. 10) obtained the same equations given below using the criterion of minimum covariance.

Prior to obtaining a minimal data set no unique estimate of the state exists so an auxiliary variable must be introduced. Define

$$\hat{y}_{n|n} = F_{n|n} \hat{x}'_{n|n} \tag{2.3.30}$$

$$\hat{y}_{n|n-1} = F_{n|n-1} \hat{x}'_{n|n-1} \tag{2.3.31}$$

where $\hat{x}'_{n|n}$ and $\hat{x}'_{n|n-1}$ are the state estimates obtained without a priori information and $F_{n|n}$ and $F_{n|n-1}$ will be subsequently defined. It can be shown that a unique $\hat{y}_{n|n}$ and $\hat{y}_{n|n-1}$ exist at all times, but only if $F_{n|n}$ and $F_{n|n-1}$ are of full rank and possess inverses do unique $\hat{x}'_{n|n}$ and $\hat{x}'_{n|n-1}$ exist.

Recursive equations for $\hat{y}_{n|n}$, $F_{n|n}$, $\hat{y}_{n|n-1}$, and $F_{n|n-1}$ can be obtained with initial conditions

$$\hat{y}_{o|o} = 0$$

$$F_{o|o} = 0$$

Subsequently,

$$\hat{y}_{n|n-1} = (I - C_n\Gamma_n^T) \; \Phi^T(n-1,n) \; \hat{y}_{n-1|n-1} \tag{2.3.32}$$

$$\hat{y}_{n|n} = y_{n|n-1} + H_n^T \; R_n^{-1} \; z_n \tag{2.3.33}$$

$$F_{n|n-1} = S_n - S_n\Gamma_n D_n^{-1}\Gamma_n^T S_n \tag{2.3.34}$$

$$F_{n|n} = F_{n|n-1} + H_n^T R_n^{-1} H_n \qquad (2.3.35)$$

where $S_n = \Phi^T(n-1,n)\ F_{n-1|n-1}\ \Phi(n-1,n)$

$$D_n = Q_n^{-1} + \Gamma_n^T S_n \Gamma_n$$

$$C_n = S_n \Gamma_n D_n^{-1}$$

It can be shown that $F_{n|n-1}$ and $F_{n|n}$ are equal to the inverse of the state estimation error covariance matrix before and after the n^{th} measurement respectively. For $n < n_o$, $F_{n|n}$ is singular, implying that some or all elements of the error covariance matrix are infinite, this in turn implying that some or all of the elements of the state cannot be estimated on the basis of the measurements taken. However, once a minimal data set is obtained, the state estimate $\hat{x}'_{n|n}$ can be obtained from the equation below.

$$\hat{x}'_{n|n} = F_{n|n}^{-1}\ \hat{y}_{n|n} \qquad (2.3.36)$$

Subsequently, the usual state estimation equations (2.3.20) and (2.3.24) can be used with the solution of the minimal data set (2.3.36) used as the initial state estimate and $F_{n|n}^{-1}$ used as the covariance of the initial state estimation error.

The solution of the state estimation problem with no a priori information can be thought of as the limiting case of the solution with a priori information as $P_{o|o}^{-1} \rightarrow 0$. In other words, the covariance of the a priori estimation error distribution becomes arbitrarily large and in the limit becomes infinite. This is equivalent to having no a priori information about the state.

The state estimate obtained using a priori information can be shown to be completely equivalent to a linear combination of the state estimate obtained without use of a priori information and the propagated forward initial state estimate.

$$\hat{x}_{n|n} = P_{n|n}(P^{-1}_{n|o}\ \hat{x}_{n|o} + F_{n|n}\ \hat{x}'_{n|n}) \qquad (2.3.37)$$

where $\hat{x}_{n|n}$ is the combined state estimate

$\hat{x}'_{n|n}$ is the state estimate obtained without a priori information

$\hat{x}_{n|o}$ is the propagated forward initial state estimate

$$\hat{x}_{n|o} = \Phi(n,0)\ \hat{x}_{o|o}$$

$P_{n|o}$ is the covariance of the propagated forward initial state estimation error

$$P_{n|o} = \Phi(n,0)P_{o|o}\Phi^T(n,0) + \sum_{i=1}^{n}\Phi(n,i)\Gamma_i Q_i \Gamma_i^T \Phi^T(n,i)$$

$P_{o|o}$ is the covariance of the a priori state distribution

$P_{n|n}$ is the covariance of the combined state estimation error

$$P_{n|n} = (P_{n|o}^{-1} + F_{n|n})^{-1}$$

This result is also equivalent to setting the initial conditions on $\hat{y}_{o|o}$ and $F_{o|o}$ to $P_{o|o}^{-1}\hat{x}_{o|o}$ and $P_{o|o}^{-1}$ respectively.

It can be shown that in most situations (when the state is completely observable by the measurements and controllable by the driving noise) that as n → ∞,

$$P_{n|n}P_{n|o}^{-1} \rightarrow 0$$

in which case $\hat{x}_{n|n} \rightarrow \hat{x}'_{n|n}$

Thus as would be expected, for large n, the effect of any initial state estimate will become arbitrarily small.

If the true values of R and Q are not known precisely, then the measurement information cannot be processed optimally. Let R^* and Q^* represent the assumed value of the sequences R and Q, $\hat{x}^*_{n|n}$ represent the state estimate after n measurements using R^* and Q^* to compute the measurement residual gain matrices, and $P^*_{n|n}$ represent the "computed" state covariance matrix. Then

$$\hat{x}^*_{n|n} = \hat{x}^*_{n|n-1} + A^*_n\,(z_n - H_n\hat{x}^*_{n|n-1}) \qquad (2.3.38)$$

$$P^*_{n|n} = (I - A^*_nH_n)P^*_{n|n-1}(I - A^*_nH_n)^T + A^*_n\,R^*_n\,A^{*T}_n \qquad (2.3.39)$$

$$A^*_n = P^*_{n|n-1}H^T_n(R^*_n + H_nP^*_{n|n-1}H^T_n)^{-1} \qquad (2{\circ}3.40)$$

$$\hat{x}^*_{n|n-1} = \Phi(n,n-1)\ \hat{x}^*_{n-1|n-1} \tag{2.3.41}$$

$$P^*_{n|n-1} = \Phi(n,n-1)P^*_{n-1|n-1}\ \Phi^T(n,n-1) + \Gamma_n Q^*_n \Gamma^T_n \tag{2.3.42}$$

$P^*_{n|n}$ represents the conditional state covariance matrix after the n^{th} measurement, conditioned upon the assumption that $R = R^*$ and $Q = Q^*$. If this assumption is not valid, then $P^*_{n|n}$ does not accurately represent the state covariance matrix. It can easily be shown that the actual conditional covariance matrix can be computed recursively using the following equations.

$$P_{n|n} = (I - A^*_n H_n)P_{n|n-1}(I - A^*_n H_n)^T + A^*_n\ R_n\ A^{*T}_n \tag{2.3.43}$$

$$P_{n|n-1} = \Phi(n,n-1)P_{n-1|n-1}\Phi^T(n,n-1) + \Gamma_n\ Q_n\ \Gamma^T_n \tag{2.3.44}$$

$P_{n|n}$ represents the state covariance matrix under the assumptions that R^* and Q^* are used to compute the filter gains (2.3.40) while the true values of the noise covariances are R and Q. If the initial state covariance is presumed to be known, then $P_{o|o} = P^*_{o|o}$. Unless $R = R^*$ and $Q = Q^*$, $P_{n|n}$ will not be equal to $P^*_{n|n}$. Depending upon the values of R, R^*, Q, Q^*, this deviation can be very significant. Numerical results of a computer simulation of these equations for a particular system are given in Chapter 6.

Because of the linearity of the maximum likelihood equations in the state estimation problem, a strong statement can be made about the distribution of the estimation error. From the form of the state estimation equations it can be seen that if the initial state distribution is normal as well as the measurement and driving noises, then the state estimate is also a normal random variable. In order to completely specify the distribution of the estimation error, the mean and covariance of the distribution must be determined.

Conventionally, an estimator is said to be unbiased if over an ensemble of trials the expected value of the state estimate is equal to the expected value of the state. Implicit in this definition is averaging over the probability density functions of the measurement and driving noises as well as averaging over the ensemble of all initial conditions of the state. Even if incorrect values of R and Q are used to compute the measurement residual gain matrices, the state estimate remains unbiased in the above sense as long as the measurement gains are fixed numbers and are not random functions of the outcomes of the measurement process.

The conditional expected value of the state estimate (2.3.38) can be computed recursively.

$$\varepsilon(\hat{x}^*_{n|n}) = \varepsilon(\hat{x}^*_{n|n-1}) + \varepsilon[A^*_n(z_n - H_n\hat{x}^*_{n|n-1})] \qquad (2.3.45)$$

Under the assumption that A^*_n is not a random variable under the expectation operator,

$$\varepsilon[A_n^*(z_n - H_n\hat{x}^*_{n|n-1})] = A_n^*\ \varepsilon(v_n - H_n\tilde{x}^*_{n|n-1})$$

$$= -A_n^* H_n\ \varepsilon(\tilde{x}^*_{n|n-1})$$

where $\tilde{x}^*_{n|n-1} = \hat{x}^*_{n|n-1} - x_n$

But from (2.3.41)

$$\varepsilon(\hat{x}^*_{n|n-1}) = \Phi(n,n-1)\ \varepsilon(\hat{x}^*_{n-1|n-1}) \qquad (2.3.46)$$

Since $\varepsilon(x_n) = \Phi(n,n-1)\ \varepsilon(x_{n-1})$

and $\varepsilon(\hat{x}^*_{n|n-1}) = \varepsilon(x_n) + \varepsilon(\tilde{x}^*_{n|n-1})$

then $\varepsilon(\hat{x}^*_{n|n}) = \varepsilon(x_n) + (I - A_n^* H_n)\ \Phi(n,n-1)\ \varepsilon(\tilde{x}^*_{n-1|n-1})$

Repeating the above procedure, it can be shown that

$$\varepsilon(\hat{x}^*_{n|n}) = \varepsilon(x_n) + [\prod_{i=1}^{n} (I - A_i^* H_i)\ \Phi(i,i-1)]\ \varepsilon(\tilde{x}^*_{o|o}) \qquad (2.3.47)$$

With $\varepsilon(\tilde{x}^*_{o|o}) = \varepsilon(\hat{x}_{o|o} - x_o)$

and $\hat{x}_{o|o} = \varepsilon(x_o)$

then $\varepsilon(\tilde{x}^*_{o|o}) = 0$

and $\varepsilon(\hat{x}^*_{n|n}) = \varepsilon(x_n)$ for all n $\qquad$ (2.3.48)

This result is independent of the values of R, R^*, Q, Q^*. The maximum likelihood state estimate remains unbaised for any values of R^* and Q^*, but the covariance of such an estimate is a function of these quantities as expressed by (2.3.43) and (2.3.44). Thus it can be seen that over the ensemble of trials with all possible initial conditions, measurement noises, and driving noises, the state estimation error is zero mean normally distributed with covariance $P_{n|n}$, for any n.

Now the question is asked: Is the state estimate biased over the ensemble of trials with the same initial conditions? Or in other words, if the initial state were fixed and one averaged the estimate over all measurement and driving noises which might be experienced, would the state estimate be biased? The answer is yes if a priori information about the state is used and the initial state is different from the initial estimate. This can be shown in a fashion analogous to the previous work. Now all conditional expected values are additionally conditioned upon the value of x_o, the initial state. From (2.3.47),

$$\varepsilon(\hat{x}^*_{n|n}|x_o) = \varepsilon(x_n|x_o) + [\prod_{i=1}^{n}(I-A^*_iH_i)\Phi(i,i-1)]\varepsilon(\tilde{x}^*_{o|o}|x_o) \qquad (2.3.49)$$

Now $\varepsilon(\tilde{x}^*_{o|o}|x_o) = \varepsilon[(\hat{x}^*_{o|o} - x_o)|x_o]$

$$= \hat{x}^*_{o|o} - x_o$$

as averaging is not performed over x_o. Unless $x_o = \hat{x}^*_{o|o}$, then

$$\varepsilon(\hat{x}^*_{n|n} | x_o) \neq \varepsilon(x_n | x_o)$$

The bias of the estimator is due to the use of a priori information in the estimator. If no a priori information is used, it is easy to show that

$$\varepsilon(\hat{x}^*_{n|n} | x_o) = \varepsilon(x_n | x_o) = \Phi(n,0)\ x_o$$

However, even if initial information is used, as n becomes large the bias due to initial condition error becomes arbitrarily small. On the average, $x_o = \bar{x}_o = \hat{x}^*_{o|o}$ and the estimator is unbiased as shown before. But over the ensemble of all possible trials with the same initial conditions, the estimate is only asymptotically unbiased. However, the distribution of the estimate about this possibly biased value can be shown to be normal for any n.

A slightly different definition of unbiasedness is used in Chapter 3 in the discussion of maximum likelihood estimators of more general parameters. There, an estimator $\hat{\alpha}_n$ of the true value of the parameters α is said to be unbiased if

$$\varepsilon(\hat{\alpha}_n | \alpha_o) = \alpha_o$$

where α_o is the true value of α. This definition is really appropriate in situations when no a priori information about the parameters is used so that the parameter estimate is a function of the measurements alone. However, the asymptotic behavior of the estimator will be shown to be independent of the a priori estimate so that this definition is useful even if a priori information is used in obtaining the estimate.

Using this definition of unbiasedness, the maximum likelihood state estimate is unbiased if

$$\varepsilon(\hat{x}^*_{n|n}|x_n) = x_n$$

Using a procedure similar to that used to obtain (2.3.47) and (2.3.49), it can be shown that

$$\varepsilon(\hat{x}^*_{n|n}|x_n) = \varepsilon(x_n|x_n) + [\prod_{i=1}^{n}(I-A^*_i H_i)\Phi(i,i-1)]\varepsilon(\tilde{x}^*_{o|o}|x_n) \qquad (2.3.50)$$

But $$\varepsilon(\tilde{x}^*_{o|o}|x_n) = \varepsilon[(\hat{x}^*_{o|o} - x_o)|x_n]$$

$$= \hat{x}^*_{o|o} - \Phi(0,n)\ x_n$$

and $$\varepsilon(x_n|x_n) = x_n$$

Unless $\hat{x}^*_{o|o} = \Phi(0,n)x_n$, the maximum likelihood estimator is biased. But as before, if one looks at the asymptotic behavior of the estimator or studies an estimator which does not use a priori information about the state, then

$$\varepsilon(\hat{x}^*_{n|n}|x_n) = x_n$$

and the estimator is unbiased.

Now the question is asked: What is the effect of possible biases in the measurement and driving noises and what can be done to estimate these biases? In such a situation, the system state is given by the relationship

$$x_n = \Phi(n,n-1)x_{n-1} + \Gamma_n(w_n + B_w) \qquad (2.3.51)$$

where as before w_n is a zero mean random variable with covariance Q_n, with w_n independent of w_k for $k \neq n$. B_w is a constant bias independent of w_n with

$$\varepsilon(B_w) = 0$$

$$\varepsilon(B_w\ B_w^T) = \sigma^2_{B_w}$$

These conditional expected values are taken over the ensemble of all possible driving noise bias values. It is usually assumed that over the above mentioned ensemble, B_w is normally distributed with zero mean and covariance $\sigma^2_{B_w}$.

The measurement z_n is given by

$$z_n = H_n\ x_n + v_n + B_v \qquad (2.3.52)$$

where as before v_n is a zero mean random variable with covariance R_n, with v_n independent of v_k for $k \neq n$. B_v is a constant measurement bias independent of v_n and the driving noise bias B_w, with

$$\varepsilon(B_V) = 0$$

$$\varepsilon(B_V\ B_V^T) = \sigma^2_{B_V}$$

These conditional expected values are taken over the ensemble of all possible measurement noise bias values. Again it is usually assumed that B_V is normally distributed.

If the state x_n is estimated with the effects of these biases neglected, then the state estimate $\hat{x}^*_{n|n}$ is computed using (2.3.38) and (2.3.41), with the "computed" covariance matrix given by (2.3.39) and (2.3.42). It is assumed that the values of R and Q used to compute these matrices and the measurement residual gains are the correct values. Now however, the state estimate will not be an optimal estimate and $P^*_{n|n}$ will not correspond to the actual state estimation error covariance because of the neglected biases.

From (2.3.51) and (2.3.41) it can be seen that

$$\tilde{x}^*_{n|n-1} = \Phi(n,n-1)\ \tilde{x}_{n-1|n-1} - \Gamma_n(w_n + B_W) \qquad (2.3.53)$$

Then the actual state estimation error covariance matrix before the measurement at time n is given by

$$\begin{aligned} P_{n|n-1} &= \varepsilon(\tilde{x}^*_{n|n-1}\tilde{x}^{*T}_{n|n-1}) \\ &= \Phi(n,n-1)P_{n-1|n-1}\Phi^T(n,n-1) + \Gamma_n(Q_n + \sigma^2_{B_W})\Gamma^T_n \qquad (2.3.54) \\ &\quad - \Phi(n,n-1)\varepsilon(\tilde{x}^*_{n-1|n-1}B_W^T) - \varepsilon(B_W x^{*T}_{n-1|n-1})\Phi^T(n,n-1) \end{aligned}$$

So in order to compute $P_{n|n-1}$, the correlation between the driving noise bias B_w and the state estimation error $\tilde{x}^*_{n-1|n-1}$ must be determined. This will be done subsequently.

From (2.3.52) and (2.3.38) it can be seen that

$$\tilde{x}^*_{n|n} = \tilde{x}^*_{n|n-1} + A^*_n(v_n + B_v - H_n \tilde{x}^*_{n|n-1}) \tag{2.3.55}$$

Then the actual state estimation error covariance matrix after the measurement at time n is given by

$$\begin{aligned} P_{n|n} &= \varepsilon(\tilde{x}^*_{n|n} \tilde{x}^{*T}_{n|n}) \\ &= P_{n|n-1} + A^*_n(R_n + \sigma^2_{B_v} + H_n P_{n|n-1} H^T_n \qquad (2.3.56) \\ &\quad - H_n \varepsilon(\tilde{x}^*_{n|n-1} B^T_v) - \varepsilon(B_v \tilde{x}^{*T}_{n|n-1}) H^T_n) A^{*T}_n - P_{n|n-1} H^T_n A^{*T}_n \\ &\quad + A^*_n \varepsilon(B_v \tilde{x}^{*T}_{n|n-1}) + \varepsilon(\tilde{x}^*_{n|n-1} B^T_v) A^{*T}_n - A^*_n H_n P_{n|n-1} \end{aligned}$$

The correlation between B_v and $\tilde{x}^*_{n|n-1}$ must be determined in order to evaluate $P_{n|n}$.

Multiplying (2.3.53) by B_v and performing the conditional expected value,

$$\varepsilon(\tilde{x}^*_{n|n-1} B^T_v) = \Phi(n,n-1)\ \varepsilon(\tilde{x}^*_{n-1|n-1} B^T_v) \tag{2.3.57}$$

since it is assumed that B_v is independent of w_n and B_w.

$\varepsilon(\tilde{x}^*_{n|n} B^T_v)$ and $\varepsilon(\tilde{x}^*_{n|n} B^T_w)$ can be computed recursively. Multiplying (2.3.55) by B_v and performing the expected value,

$$\varepsilon(\tilde{x}^*_{n|n} B_V^T) = (I - A_n^* H_n)\ \varepsilon(\tilde{x}^*_{n|n-1} B_V^T) + A_n^*\ \sigma^2_{B_V} \qquad (2.3.58)$$

$$= (I - A_n^* H_n)\ \Phi(n,n-1)\ \varepsilon(\tilde{x}^*_{n-1|n-1} B_V^T) + A_n^*\ \sigma^2_{B_V}$$

Multiplying (2.3.55) by B_W and performing the expected value,

$$\varepsilon(\tilde{x}^*_{n|n} B_W^T) = (I - A_n^* H_n)\ \varepsilon(\tilde{x}^*_{n|n-1} B_W^T) \qquad (2.3.59)$$

since B_W is assumed to be independent of v_n and B_V. But from (2.3.53) it can be seen that

$$\varepsilon(\tilde{x}^*_{n|n-1} B_W^T) = \Phi(n,n-1)\ \varepsilon(\tilde{x}^*_{n-1|n-1} B_W^T) - \Gamma_n\ \sigma^2_{B_W} \qquad (2.3.60)$$

So (2.3.59) becomes

$$\varepsilon(\tilde{x}^*_{n|n} B_W^T) = (I - A_n^* H_n)\ \Phi(n,n-1)\ \varepsilon(\tilde{x}^*_{n-1|n-1} B_W^T)$$

$$- (I - A_n^* H_n)\ \Gamma_n\ \sigma^2_{B_W} \qquad (2.3.61)$$

It is assumed that the initial state estimation error is independent of B_W and B_V so the initial conditions on (2.3.57) and (2.3.61) are

$$\varepsilon(\tilde{x}^*_{o|o}\ B_W^T) = \varepsilon(\tilde{x}^*_{o|o}\ B_V^T) = 0$$

Using an analysis similar to that previously given, it can be shown that across the ensemble of all possible initial state conditions, measurement and driving noises, and measurement and driving noise biases, the state estimate $\hat{x}^*_{n|n}$ is unbiased. However, if the biases are present, the actual state estimation error covariance matrix is no longer accurately represented by $P^*_{n|n}$ but rather by $P_{n|n}$ as given above.

If there is a possibility that biases may be present in the measurement or driving noises, then it is usually preferable to estimate their values so that their effect upon the state estimator is diminished. This can easily be accomplished within the framework of maximum likelihood state estimation already established.

Define a new state variable

$$s^T_n = (x^T_n, B^T_w, B^T_v) \tag{2.3.62}$$

and a new state transition matrix

$$\Psi(n,n-1) = \begin{bmatrix} \Phi(n,n-1) & \Gamma_n & 0 \\ 0 & I & 0 \\ 0 & 0 & I \end{bmatrix} \tag{2.3.63}$$

and a new forcing function matrix

$$\lambda_n = \begin{bmatrix} \Gamma_n \\ 0 \\ 0 \end{bmatrix} \tag{2.3.64}$$

Then the augmented state s_n obeys the recursive relationship

$$s_n = \Psi(n,n-1)\ s_{n-1} + \lambda_n\ w_n \qquad (2.3.65)$$

Define a new observation matrix

$$G_n = (H_n,\ 0,\ I) \qquad (2.3.66)$$

Then the measurement z_n is given by

$$z_n = G_n\ s_n + v_n \qquad (2.3.67)$$

Now the problem is reduced to exactly the same form as the case when the noises were zero mean except that now the state vector is of increased dimension and includes all possible noise biases. The estimator for the augmented state s_n can be formulated in exactly the same way as before with initial conditions

$$\hat{s}^T_{o|o} = (\hat{x}^T_{o|o},\ 0,\ 0)$$

This says that the a priori estimates of the biases should always be zero since, if they were nonzero, they could be removed with the residual uncertainty in the bias values then zero mean.

The covariance of the initial augmented state estimation error is given by

$$E_{o|o} = \varepsilon(\tilde{s}_{o|o}\, \tilde{s}^T_{o|o}) = \begin{bmatrix} P_{o|o} & 0 & 0 \\ 0 & \sigma^2_{B_W} & 0 \\ 0 & 0 & \sigma^2_{B_V} \end{bmatrix}$$

where $P_{o|o}$ is the covariance of the unaugmented state estimate, $\sigma^2_{B_W}$ is the covariance of the driving noise bias, and $\sigma^2_{B_V}$ is the covariance of the measurement noise bias.

Thus the augmented state can be estimated using the same form of the equations as for the unaugmented state with the substitutions

$$\begin{aligned} H_n &\rightarrow G_n \\ \Gamma_n &\rightarrow \lambda_n \\ P_{n|n} &\rightarrow E_{n|n} \\ \hat{x}_{n|n} &\rightarrow \hat{s}_{n|n} \end{aligned}$$

If the true covariances of the random parts of the noises as well as the covariances of the bias parts of the noises are known precisely and used in the filter, then it can be shown that $E_{n|n}$ accurately represents the covariance of the augmented state estimation error, and the filter is optimal in a minimum covariance or maximum likelihood sense.

If instead of the measurement and driving noises having a bias, they have a component which is correlated with past noises, then a slightly different approach must be used. Only a limited type of correlation is easily treated so the

Front

following definitions are made.

It is assumed that the state obeys the relationship

$$x_n = \Phi(n,n-1)\ x_{n-1} + \Gamma_n(w_n + w_n^c) \qquad (2.3.68)$$

where w_n is uncorrelated zero mean noise such that

$$\varepsilon(w_n\ w_j^T) = Q_n\ \delta_{jn} \qquad (2.3.69)$$

and w_n^c is correlated zero mean noise such that

$$\varepsilon(w_n^c\ w_j^{cT}) = Q_c\ e^{-\,|n-j|/\tau_w} \qquad (2.3.70)$$

τ_w is the "correlation time" of the driving noise. It is also assumed that w_n and w_n^c are mutually uncorrelated so that

$$\varepsilon(w_n\ w_j^c) = 0 \qquad (2.3.71)$$

The correlated noise w_n^c can be generated by considering w_n^c to be composed of two parts.

$$w_n^c = w_n^* + (e^{-1/\tau_w})\ w_{n-1}^c \qquad (2.3.72)$$

where w_n^* is a zero mean random noise that is independent of all past noises with

$$\varepsilon(w_n^*\ w_n^{*T}) = Q_c\ (1 - e^{-2/\tau_w}) \qquad (2.3.73)$$

It is easy to show that the correlated noise defined by (2.3.72) has the proper correlation between the noises at different times as given by (2.3.70).

It is also assumed that the measurement z_n is given by

$$z_n = H_n x_n + v_n + v_n^c \tag{2.3.74}$$

where v_n is uncorrelated zero mean noise such that

$$\varepsilon(v_n v_j^T) = R_n \delta_{jn} \tag{2.3.75}$$

and v_n^c is correlated zero mean noise such that

$$\varepsilon(v_n^c v_j^{cT}) = R_c e^{-|n-j|/\tau_v} \tag{2.3.76}$$

τ_v is the "correlation time" of the measurement noise. It is again assumed that v_n and v_n^c are mutually uncorrelated with the further assumption that all measurement noises are uncorrelated with all driving noises.

Again it is convenient to define the correlated measurement noise by

$$v_n^c = v_n^* + (e^{-1/\tau_v}) v_{n-1}^c \tag{2.3.77}$$

where v_n^* is a zero mean random noise that is independent of all past noises with

$$\varepsilon(v_n^* v_n^{*T}) = R_c (1 - e^{-2/\tau_v}) \tag{2.3.78}$$

It is easy to show that the correlated measurement noise defined by (2.3.77) has the proper correlation between the noises at different times as given by (2.3.75).

It should be noted that when the correlation time of the noises becomes very large, the correlated noises approach constant biases, whereas as the correlation times become small, the noises become uncorrelated.

If it is assumed that the state x_n is estimated neglecting this correlation, the state estimate $\hat{x}^*_{n|n}$ is computed using (2.3.38) and (2.3.41), with the "computed" covariance matrix given by (2.3.39) and (2.3.42). Again $\hat{x}^*_{n|n}$ will not be an optimal estimate and $P^*_{n|n}$ will not correspond to the actual state estimation error covariance matrix because of the neglected correlation in the noises.

From (2.3.68) and (2.3.41) it can be seen that

$$\tilde{x}^*_{n|n-1} = \Phi(n,n-1)\ \tilde{x}^*_{n-1|n-1} - \Gamma_n(w_n + w^c_n) \tag{2.3.79}$$

Then the actual state estimation error covariance matrix before the measurement at time n is given by

$$\begin{aligned} P_{n|n-1} = {} & \Phi(n,n-1)P_{n-1|n-1}\Phi^T(n,n-1) + \Gamma_n(Q_n + Q_c)\Gamma^T_n \\ & - \Gamma_n\varepsilon(w^c_n\ \tilde{x}^{*T}_{n-1|n-1})\Phi^T(n,n-1) - \Phi(n,n-1)\varepsilon(\tilde{x}^*_{n-1|n-1}w^{cT}_n)\Gamma^T_n \end{aligned} \tag{2.3.80}$$

In order to compute $P_{n|n-1}$, the correlation between the driving noise w^c_n and the state estimation error $\tilde{x}^*_{n-1|n-1}$ must be computed. This will be done subsequently.

From (2.3.74) and (2.3.38) it can be seen that

$$\tilde{x}^*_{n|n} = \tilde{x}^*_{n|n-1} + A^*_n \left(v_n + v^c_n - H_n \tilde{x}^*_{n|n-1}\right) \qquad (2.3.81)$$

Then the actual state estimation error covariance matrix after the measurement at time n is given by

$$P_{n|n} = (I-A^*_n H_n) P_{n|n-1} (I-A^*_n H_n)^T + A^*_n (R_n + R_c) A^{*T}_n \qquad (2.3.82)$$

$$+ A^*_n \varepsilon(v^c_n \tilde{x}^{*T}_{n|n-1}) (I-A^*_n H_n)^T + (I-A^*_n H_n) \varepsilon(\tilde{x}^*_{n|n-1} v^{cT}_n) A^{*T}_n$$

The correlation between v^c_n and $\tilde{x}^*_{n|n-1}$ must be computed in order to evaluate $P_{n|n}$.

Multiplying (2.3.79) by v^c_n and performing the conditional expected value

$$\varepsilon(\tilde{x}_{n|n-1} v^{cT}_n) = \Phi(n,n-1)\ \varepsilon(\tilde{x}^*_{n-1|n-1} v^{cT}_n) \qquad (2.3.83)$$

since it is assumed that v^c_n is independent of w^c_n. But using (2.3.77) plus the independence of v^*_n,

$$\varepsilon(\tilde{x}^*_{n-1|n-1} v^{cT}_n) = (e^{-1/\tau_v})\ \varepsilon(\tilde{x}^*_{n-1|n-1} v^{cT}_{n-1}) \qquad (2.3.84)$$

Similarly it can be seen that

$$\varepsilon(\tilde{x}^*_{n-1|n-1} w^{cT}_n) = (e^{-1/\tau_w})\ \varepsilon(\tilde{x}^*_{n-1|n-1} w^{cT}_{n-1}) \qquad (2.3.85)$$

$\varepsilon(\tilde{x}^*_{n-1|n-1} v^{cT}_{n-1})$ and $\varepsilon(\tilde{x}^*_{n-1|n-1} w^{cT}_{n-1})$ can be computed recursively. Multiplying (2.3.81) by v^c_n and performing the expected value,

$$\varepsilon(\tilde{x}^*_{n|n} v^{cT}_n) = (I - A^*_n H_n)\ \varepsilon(\tilde{x}^*_{n|n-1} v^{cT}_n) + A^*_n\ R_c \tag{2.3.86}$$

$$= (I - A^*_n H_n)\ \Phi(n,n-1)(e^{-1/\tau_v})\ \varepsilon(\tilde{x}^*_{n-1|n-1} v^{cT}_{n-1})$$

$$+ A^*_n\ R_c$$

Multiplying (2.3.81) by w^c_n and performing the expected value,

$$\varepsilon(\tilde{x}^*_{n|n}\ w^{cT}_n) = (I - A^*_n H_n)\ \varepsilon(\tilde{x}^*_{n|n-1}\ w^{cT}_n) \tag{2.3.87}$$

$$= (I - A^*_n H_n)\ \Phi(n,n-1)\ \varepsilon(\tilde{x}^*_{n-1|n-1}\ w^{cT}_n) - \Gamma_n\ Q_c$$

$$= (I - A^*_n H_n)\ \Phi(n,n-1)(e^{-1/\tau_w})\ \varepsilon(\tilde{x}^*_{n-1|n-1}\ w^{cT}_{n-1}) - \Gamma_n Q_c$$

It is assumed that the initial state estimation error is uncorrelated with the measurement and driving noises, so the initial conditions on the recursive equations (2.3.86) and (2.3.87) are

$$\varepsilon(\tilde{x}^*_{o|o}\ v^{cT}_o) = \varepsilon(\tilde{x}^*_{o|o}\ w^{cT}_o) = 0$$

By analogy with the estimation of possible noise biases, it is possible to estimate the correlated part of the measurement and driving noises.

Define a new state variable

$$s_n^T = (x_n^T, w_n^{cT}, v_n^{cT}) \tag{2.3.88}$$

and a new state transition matrix

$$\Psi(n,n-1) = \begin{bmatrix} \Phi(n,n-1) & (e^{-1/\tau_w})\Gamma_n & 0 \\ 0 & (e^{-1/\tau_w})I & 0 \\ 0 & 0 & (e^{-1/\tau_v})I \end{bmatrix} \tag{2.3.89}$$

and a new forcing function matrix

$$\lambda_n = \begin{bmatrix} \Gamma_n & \Gamma_n & 0 \\ I & 0 & 0 \\ 0 & 0 & I \end{bmatrix} \tag{2.3.90}$$

and a new "driving noise" vector

$$u_n^T = (w_n^{*T}, w_n^T, v_n^{*T}) \tag{2.3.91}$$

It can be seen that the new state s_n satisfies the relationship

$$s_n = \Psi(n,n-1)\, s_{n-1} + \lambda_n\, u_n \tag{2.3.92}$$

and the measurement z_n is given by

$$z_n = G_n\, s_n + v_n \tag{2.3.93}$$

where G_n is defined by

$$G_n = (H_n, 0, I) \qquad (2.3.94)$$

Now the problem is reduced to exactly the same form as the cases when the noises are uncorrelated except that now the state vector is of increased dimension and includes all possible correlated noises. The estimator for the augmented state s_n can be formulated in exactly the same way as before with initial conditions

$$\hat{s}^T_{o|o} = (\hat{x}^T_{o|o}, 0, 0)$$

The covariance of the initial augmented state estimation error is given by

$$E_{o|o} = \varepsilon(\tilde{s}_{o|o} \, \tilde{s}^T_{o|o}) = \begin{bmatrix} P_{o|o} & 0 & 0 \\ 0 & Q_c & 0 \\ 0 & 0 & R_c \end{bmatrix}$$

Thus the augmented state can be estimated using the same form of the equations as for the unaugmented state without correlated noises.

If the true covariance of the correlated and uncorrelated parts of the noises as well as the proper correlation times are known precisely and used in the filter, then it can be shown that $E_{n|n}$ as computed by the filter accurately represents the covariance of the augmented state estimation error, and the filter is optimal in a minimum covariance or maximum likelihood sense.

Chapter 3

MAXIMUM LIKELIHOOD ESTIMATION OF NOISE COVARIANCE PARAMETERS AND THE SYSTEM STATE

3.1 Introduction

In Chapter 2 the theory of maximum likelihood estimation was briefly discussed and then applied to the problem of state estimation. The resulting equations were derived under the assumption that the probability density functions of the measurement and driving noises as well as the initial state probability density function are known a priori. It was shown that if the second order statistics of the noises are not known precisely, the state estimation becomes suboptimal. The purpose of this chapter is to utilize the concepts of maximum likelihood to remove the restriction that R and Q be known precisely a priori in order to obtain an optimal state estimate.

In Section 3.2 important definitions are given and a summary of some classical results of maximum likelihood estimation discussed. These results concern the asymptotic properties of maximum likelihood estimators, but they cannot be directly applied to the problem of state and noise covariance estimation.

In Section 3.3 the likelihood functions appropriate for the solution of a set of closely related problems are derived, all of which concern the estimation of the noise covariance parameters. Section 3.4 is devoted to demonstrating the asymptotic properties of these estimators.

The remainder of this chapter concerns the application of the theoretical results to the problem of state and noise covariance estimation.

3.2 Summary of Previous Results in Maximum Likelihood Estimation

Maximum likelihood estimation has been studied by many authors and many useful results have been obtained concerning the properties of maximum likelihood estimators. These results apply directly only to a limited set of problems, when the measurements are independent and identically distributed. However, they provide a base upon which the analysis of more general problems can rest. The purpose of this section is to summarize the important results and definitions which will be needed to extend the analysis to more general problems.

First several important definitions must be made. These definitions apply equally well to any situation when the values of certain parameters are to be estimated on the basis of observations of a random variable which is a function of these parameters. They are not limited to situations when the criterion of maximum likelihood is used to define the estimate.

The estimator of the true value of the parameter α is an observable random variable, say $\hat{\alpha}_n(z_1,..,z_n)$ which is a function of the sample elements $(z_1,..,z_n)$ and whose distribution is, in some sense, concentrated about the true value of α. As in linear estimation, it will be found that the covariance of the estimate is often a reasonable criterion for measuring the concentration. If the realized (observed) value of $\hat{\alpha}_n$ corresponding to a realized (observed) value of $(z_1,..,z_n)$ is used for α_o, the true value of α, then the random variable $\hat{\alpha}_n$ is called a point estimate or estimator for α_o. This use of $\hat{\alpha}_n$ normally would be made, of course, only when the value of α_o is unknown.

If when $\alpha = \alpha_o$, $\varepsilon(\hat{\alpha}_n|\alpha_o) = \alpha_o$, then $\hat{\alpha}_n$ is called an unbiased estimator for α_o. This is the last definition of unbiasedness that was used in Chapter 2 in the discussion of maximum likelihood state estimation.

If an estimator $\hat{\alpha}_n$ converges to α_o as $n \to \infty$, it is called a consistent estimator for α_o. A necessary condition for $\hat{\alpha}_n$ to be a consistent estimator is that it be unbiased and have a covariance which goes to zero as $n \to \infty$.

If $\hat{\alpha}_n$ is an unbiased estimator for α_o having finite covariance and has the further property that no other unbiased estimator has a smaller covariance than $\hat{\alpha}_n$, it is called an efficient estimator.

The following results of maximum likelihood estimation have been obtained by Rao (Ref. 25), Wilks (Ref. 37), and Deutsch (Ref. 6) after certain assumptions have been made about the nature of the likelihood function.

Let $Z_n^T = (z_1^T,..,z_n^T)$ be a vector of n independent identically distributed observations and α be the m x 1 vectors of parameters being estimated. Then the joint conditional probability density function of Z_n can be found by application of Bayes' rule.

$$f(Z_n|\alpha) = f(Z_{n-1}|\alpha)\ f(z_n|Z_{n-1},\alpha) \tag{3.2.1}$$

where $f(z_n|Z_{n-1},\alpha)$ is the conditional probability density function of z_n given Z_{n-1} and α. Because of the assumed independence of the z_i,

$$f(z_n|Z_{n-1},\alpha) = f(z_n|\alpha) \tag{3.2.2}$$

By repeated application of Bayes' rule, it can be seen that

$$f(Z_n|\alpha) = \prod_{i=1}^{n} f(z_i|\alpha) \tag{3.2.3}$$

It is assumed that the likelihood function is chosen to be the probability density function (3.2.3), in which case the natural logarithm of the likelihood function has the form

$$\begin{aligned} L_n(Z_n,\alpha) &= \ln f(Z_n|\alpha) \\ &= \sum_{i=1}^{n} \ln f(z_i|\alpha) \end{aligned} \tag{3.2.4}$$

Then

$$\frac{\partial L_n(Z_n,\alpha)}{\partial\alpha} = \sum_{i=1}^{n} \frac{\partial \ln f(z_i|\alpha)}{\partial\alpha} \tag{3.2.5}$$

As stated in Chapter 2, maximum likelihood estimation is concerned with finding the value of the parameters α such that

$$\left(\frac{\partial L_n(Z_n,\alpha)}{\partial \alpha}\right)_{\hat{\alpha}_n} = 0$$

For notational convenience, define

$$f_n = f(Z_n|\alpha)$$

$$f = f(z|\alpha)$$

The following assumptions are made about the likelihood function.

1) The derivatives $\frac{\partial L_n}{\partial \alpha}$, $\frac{\partial^2 L_n}{\partial \alpha^2}$, $\frac{\partial^3 L_n}{\partial \alpha^3}$ exist for almost all Z_n in an interval Ω of α.

2) $\varepsilon\left[\frac{1}{f_n}\frac{\partial f_n}{\partial \alpha}|\alpha_o\right] = 0$, $\varepsilon\left[\frac{1}{f_n}\frac{\partial^2 f_n}{\partial \alpha \partial \alpha}|\alpha_o\right] = 0$

3) $\varepsilon\left[\frac{1}{f_n^2}\left(\frac{\partial f_n}{\partial \alpha}\right)^T \frac{\partial f_n}{\partial \alpha}|\alpha_o\right]$ is positive definite

4) For every α in Ω

$$\frac{1}{n}\left|\frac{\partial^3 \ln f_n}{\partial \alpha^i \partial \alpha^j \partial \alpha^k}\right| < M(Z_n)$$

with $\varepsilon[M(Z_n)|\alpha_o] < K$ for some K which is independent of α and n.

Define

$$S(z,\alpha) = \frac{\partial \ln f}{\partial \alpha} \quad \text{the m x 1 single measurement score}$$

$$S_n(Z_n,\alpha) = \frac{\partial \ln f_n}{\partial \alpha} \quad \text{the m x 1 total measurement score}$$

$$J(\alpha_o,\alpha) = \varepsilon[S^T(z,\alpha)S(z,\alpha)|\alpha_o]$$

$$= \int S^T(z,\alpha)S(z,\alpha)\ f(z|\alpha_o)\ dz$$

$$J_n(\alpha_o,\alpha) = \varepsilon[S_n^T(Z_n,\alpha)\ S_n(Z_n,\alpha)|\alpha_o]$$

$$= \int\cdot\cdot\int S_n^T(Z_n,\alpha)\ S_n(Z_n,\alpha)\ f(Z_n|\alpha_o)\ dZ_n$$

$J(\alpha_o) = J(\alpha_o,\alpha_o)$ the m x m single measurement conditional information matrix

$J_n(\alpha_o) = J_n(\alpha_o,\alpha_o)$ the m x m total measurement conditional information matrix

The following theorems are from Wilks. The proofs will not be repeated here but will be discussed subsequently.

Asymptotic Distribution of the Score

Suppose $(z_1,..,z_n)$ is a sample from the probability density function $f(z|\alpha_o)$. Let $f(z|\alpha)$ possess finite first derivatives with respect to α in the range Ω. Then if $J_n(\alpha,\alpha)$ is positive definite for α in Ω, the total measurement score $S_n(Z_n,\alpha_o)$ is asymptotically distributed for large n as a zero mean normal random variable with covariance $J_n(\alpha_o)$.

Convergence of the Maximum Likelihood Estimator

Suppose $(z_1,..,z_n)$ is a sample from the probability density function $f(z|\alpha_o)$ where $f(z|\alpha)$ possesses finite first derivatives with respect to α in Ω. Let $S^j(z,\alpha)$, the j^{th} component of the vector $S(z,\alpha)$, be a continuous function of α in Ω for all values of z except possibly for a set of zero probability. Then there exists a sequence of solutions of

$$S_n^j(Z_n,\alpha) = 0 \qquad (3.2.6)$$

which converges almost certainly to α_o. If the solution is a unique vector $\hat{\alpha}_n$ for $n \geq$ some n_o, the sequence of vectors converges almost certainly to α_o as $n \to \infty$.

Asymptotic Distribution of the Maximum Likelihood Estimator

If $(z_1,..,z_n)$ is a sample from the probability density function $f(z|\alpha_o)$ where $f(z|\alpha)$ possesses finite first and second derivatives with respect to α in the range Ω, and if the maximum likelihood estimator satisfying (3.2.6) is unique for some $n \geq$ some n_o, then it is asymptotically normally distributed for large n with mean α_o and covariance $[J_n(\alpha_o)]^{-1}$.

Thus under the assumptions previously given, the maximum likelihood estimator of the parameters α_o is asymptotically unbiased and normally distributed for any value of α_o in the range Ω, with

$$\varepsilon(\hat{\alpha}_n|\alpha_o) = \alpha_o$$

$$\varepsilon[(\hat{\alpha}_n-\alpha_o)(\hat{\alpha}_n-\alpha_o)^T|\alpha_o] = [J_n(\alpha_o)]^{-1}$$

Now the distribution of the estimation error over the ensemble of all possible true values of α_o is sought. An analytic expression for the unconditional probability density function of $\hat{\alpha}_n$ cannot be found in most situations. Formally

$$f(\hat{\alpha}_n) = \int_\Omega f(\hat{\alpha}_n, \alpha_o)\, d\alpha_o$$

$$= \int_\Omega f(\hat{\alpha}_n|\alpha_o)\, f(\alpha_o)\, d\alpha_o$$

Even if $f(\hat{\alpha}_n|\alpha_o)$ is a normal density function, the above integral is usually nonanalytic for any nontrivial $f(\alpha_o)$. However, even if the unconditional distribution of $\hat{\alpha}_n$ is not known, two useful moments of the distribution, the mean and covariance, can be evaluated.

The unconditional mean of the estimate is defined by

$$E(\hat{\alpha}_n) = \int_\Omega \varepsilon(\hat{\alpha}_n|\alpha_o)\, f(\alpha_o)\, d\alpha_o$$

$$= \int_\Omega \alpha_o\, f(\alpha_o)\, d\alpha_o \triangleq \overline{\alpha_o}$$

where $\overline{\alpha_o}$ is the mean of the distribution $f(\alpha_o)$.

The unconditional covariance of the estimate is defined by

$$\mathrm{cov}(\hat{\alpha}_n) = E[(\hat{\alpha}_n - E(\hat{\alpha}_n))(\hat{\alpha}_n - E(\hat{\alpha}_n))^T]$$

$$= E[(\hat{\alpha}_n - \overline{\alpha_o})(\hat{\alpha}_n - \overline{\alpha_o})^T]$$

But

$$E[(\hat{\alpha}_n - \overline{\alpha_o})(\hat{\alpha}_n - \overline{\alpha_o})^T] = E[(\hat{\alpha}_n - \alpha_o)(\hat{\alpha}_n - \alpha_o)^T] + E[(\overline{\alpha_o} - \alpha_o)(\overline{\alpha_o} - \alpha_o)^T]$$

$$+ E[(\hat{\alpha}_n - \alpha_o)(\alpha_o - \overline{\alpha_o})^T] + E[(\alpha_o - \overline{\alpha_o})(\hat{\alpha}_n - \alpha_o)^T]$$

and $E[(\hat{\alpha}_n - \alpha_o)(\alpha_o - \overline{\alpha_o})^T] = E[\varepsilon(\hat{\alpha}_n - \alpha_o)(\alpha_o - \overline{\alpha_o})^T]$

$$= 0$$

So $\mathrm{cov}(\hat{\alpha}_n) = E[(\hat{\alpha}_n - \alpha_o)(\hat{\alpha}_n - \alpha_o)^T] + E[(\alpha_o - \overline{\alpha_o})(\alpha_o - \overline{\alpha_o})^T]$

But $E[(\hat{\alpha}_n - \alpha_o)(\hat{\alpha}_n - \alpha_o)^T] = E[\varepsilon((\hat{\alpha}_n - \alpha_o)(\hat{\alpha}_n - \alpha_o)^T)]$

$$= E[(J_n(\alpha_o))^{-1}]$$

$$\triangleq \overline{J_n^{-1}}$$

and $E[(\alpha_o - \overline{\alpha_o})(\alpha_o - \overline{\alpha_o})^T] = \mathrm{cov}(\alpha_o)$ the covariance of the α_o distribution

Then $\mathrm{cov}(\hat{\alpha}_n) = \overline{J_n^{-1}} + \mathrm{cov}(\alpha_o)$

$\overline{J_n^{-1}}$ represents the mean square estimation error matrix, which for any nontrivial $f(\alpha_o)$ is nonanalytic. Formally

$$\overline{J_n^{-1}} = \int_{\Omega} [J_n(\alpha_o)]^{-1} f(\alpha_o)\, d\alpha_o$$

There are several approximate techniques for evaluating this integral which are discussed in Section 3.7.

3.3 Derivation of the Likelihood Function

In this section several closely related problems are studied and the likelihood function appropriate for the solution of each derived. It will be shown that the asymptotic behavior of the solutions of each problem is the same so that if the asymptotic behavior of any one is found, the results can be applied to the others. The notation and definitions of Section 2.3 are used with the additional assumption that the measurement and driving noise covariance matrices are diagonal and time invariant. The technique of maximum likelihood estimation is not restricted to cases when this assumption is valid, but the estimation problem becomes much more complicated if this assumption is not made. A discussion of the problem when this restriction is not employed is given in Chapter 7.

Estimation of Noise Covariance Parameters with No A Priori Noise Covariance Information

The first problem considered is estimating the diagonal elements of the measurement and driving noise covariance matrices without the use of a priori information about these quantities. The maximum likelihood estimate of the noise covariance parameters is defined by

$$l(\hat{R}_n,\hat{Q}_n,Z_n) = \max_{R,Q} l(R,Q,Z_n) \qquad (3.3.1)$$

where $l(R,Q,Z_n)$ is the likelihood function which is chosen to be the conditional probability density function

$$l(R,Q,Z_n) = f(Z_n|R,Q) \qquad (3.3.2)$$

By application of Bayes' rule

$$f(Z_n|R,Q) = f(Z_{n-1}|R,Q)\ f(z_n|Z_{n-1},R,Q)$$

Repeating the above procedure to find $f(Z_{n-1}|R,Q)$, it can be shown that

$$f(Z_n|R,Q) = \prod_{i=1}^{n} f(z_i|Z_{i-1},R,Q) \qquad (3.3.3)$$

where $f(z_i|Z_{i-1},R,Q)$ is the conditional probability density function of z_i given Z_{i-1}, R, and Q.

Using the results of Section 2.3, it can be shown that z_i is a normally distributed random variable with conditional mean

$$\varepsilon(z_i|Z_{i-1},R,Q) = H_i\ \hat{x}_{i|i-1}$$

and conditional covariance

$$\varepsilon(\Delta z_i \; \Delta z_i^T | Z_{i-1}, R, Q) = R + H_i P_{i|i-1} H_i^T$$

where $\Delta z_i \triangleq z_i - H_i \; \hat{x}_{i|i-1}$

$\hat{x}_{i|i-1}$ is the maximum likelihood estimate of x_i after i-1 measurements using the true values of R and Q to compute the proper filter gains, and $P_{i|i-1}$ is the conditional covariance of x_i about $\hat{x}_{i|i-1}$.

$$P_{i|i-1} = \varepsilon[(x_i - \hat{x}_{i|i-1})(x_i - \hat{x}_{i|i-1})^T | Z_{i-1}, R, Q]$$

It is assumed that a priori information about the state is used in forming the above state estimates so that a unique $\hat{x}_{i|i-1}$ exists for all i.

Define

$$B_i = R + H_i P_{i|i-1} H_i^T$$

Then the conditional probability density function of z_i is given by

$$f(z_i | Z_{i-1}, R, Q) = \frac{1}{(2\pi)^{\gamma/2} |B_i|^{1/2}} \, e^{-1/2(\Delta z_i^T B_i^{-1} \Delta z_i)} \qquad (3.3.4)$$

As in Chapter 2, it is convenient to work with the natural logarithm of the likelihood function (3.3.2).

$$L_n(R, Q, Z_n) = \ln l(R, Q, Z_n)$$

After algebraic manipulation,

$$L_n(R,Q,Z_n) = \text{constant} - 1/2\left[\sum_{i=1}^{n} \ln|B_i| + \Delta z_i^T B_i^{-1} \Delta z_i\right] \quad (3.3.5)$$

where "constant" includes all terms that are not functions of R or Q.

It is convenient to introduce an auxiliary variable.

$$\xi^T = (R^{11},..,R^{\gamma\gamma},Q^{11},..,Q^{\eta\eta})$$

ξ is the $(\gamma + \eta) \times 1$ vector of the diagonal elements of R and Q.

The likelihood equations are obtained by equating the derivatives of $L_n(R,Q,Z_n)$ with respect to ξ to zero. Using the identities of Appendix A, after algebraic manipulation,

$$\frac{\partial L_n}{\partial \xi^j} = - 1/2 \sum_{i=1}^{n} \mathrm{Tr}[(B_i^{-1} - B_i^{-1}\Delta z_i \Delta z_i^T B_i^{-1})\frac{\partial B_i}{\partial \xi^j} - 2B_i^{-1}\Delta z_i \frac{\partial \hat{x}^T_{i|i-1}}{\partial \xi^j} H_i^T] \quad (3.3.6)$$

$\hat{\xi}_n$ is found as the solution of

$$\left(\frac{\partial L_n}{\partial \xi}\right)_{\hat{\xi}_n} = 0 \quad (3.3.7)$$

In general there is no closed form solution of (3.3.7) for $\hat{\xi}_n$ so an iterative solution like those described in Section 3.6 must be employed.

Estimation of Noise Covariance Parameters with A Priori Noise Covariance Information

In this problem the measurement and driving noise covariance matrices are not known precisely a priori but rather knowledge of them is described by a joint probability density function f(R,Q), where it is assumed that f(R,Q) is known a priori. The maximum likelihood estimate of the noise covariance parameters in this case is defined by

$$l^A(\hat{R}_n,\hat{Q}_n,Z_n) = \max_{R,Q} l^A(R,Q,Z_n) \tag{3.3.8}$$

where $l^A(R,Q,Z_n)$ is the augmented likelihood function which is chosen to be the conditional probability density function

$$l^A(R,Q,Z_n) = f(R,Q|Z_n) \tag{3.3.9}$$

By application of Bayes' rule

$$f(R,Q|Z_n) = \frac{f(Z_n|R,Q)\ f(R,Q)}{f(Z_n)} \tag{3.3.10}$$

$f(Z_n)$ need not be evaluated as it is not a function of R or Q. Formally

$$f(Z_n) = \iint_{\Omega} f(Z_n|R,Q)\ f(R,Q)\ dR\ dQ$$

All R and Q dependence is integrated out.

Define $$L_n^A(R,Q,Z_n) = \ln l^A(R,Q,Z_n) \qquad (3.3.11)$$

Then it can be seen that

$$L_n^A\ (R,Q,Z_n) = L_n(R,Q,Z_n) + \ln f(R,Q) - \ln f(Z_n) \qquad (3.3.12)$$

It is assumed that R and Q are independent random variables, in which case

$$f(R,Q) = f(r)\ f(Q)$$

It is further assumed that the diagonal elements of R and Q are mutually independent, so

$$f(R) = \prod_{i=1}^{\gamma} f(R^{ii})$$

$$f(Q) = \prod_{i=1}^{\eta} f(Q^{ii})$$

Then $$L_n^A(R,Q,Z_n) = \text{constant} - 1/2 \left[\sum_{i=1}^{n} \ln|B_i| + \Delta z_i^T B_i^{-1} \Delta z_i \right]$$

$$+ \sum_{i=1}^{\gamma} \ln\ f(R^{ii}) + \sum_{i=1}^{\eta} \ln\ f(Q^{ii}) \qquad (3.3.14)$$

where "constant" includes all terms that are not functions of R and Q.

$$\frac{L_n^A(R,Q,Z_n)}{\partial\xi^j} = -\,1/2\left[\sum_{i=1}^{n} Tr[(B_i^{-1}-B_i^{-1}\Delta z_i\Delta z_i^T B_i^{-1})\,\frac{\partial B_i}{\partial\xi^j} - 2B_i^{-1}\Delta z_i\,\frac{\partial \hat{x}^T_{i\;i-1}}{\partial\xi^j}\,H_i^T]\right] + \frac{\partial \ln\, f(\xi^j)}{\partial\xi^j} \qquad (3.3.14)$$

$$= \frac{\partial L_n(R,Q,Z_n)}{\partial\xi^j} + \frac{\partial \ln\, f(\xi^j)}{\partial\xi^j}$$

(3.3.14) is then set to zero and solved for $\hat{\xi}_n$. Again there is no general closed form solution so some iterative procedure must be employed. However, it can be seen that the inclusion of a priori information has a tendency to shift the solution point towards the peak of the a priori distribution of ξ.

Estimation of Noise Covariance Parameters and the System State with No A Priori Noise Covariance Information

In this problem the noise covariance parameters and the state are to be estimated simultaneously. No a priori information about the noise covariance parameters is to be used, but as before it is assumed that a priori state information is used. The maximum likelihood estimate of these quantities is defined by

$$l(\hat{R}_n,\hat{Q}_n,\hat{x}_{n|n},Z_n) = \max_{R,Q,x_n}\; l(R,Q,x_n,Z_n) \qquad (3.3.15)$$

where $l(R,Q,x_n,Z_n)$ is the likelihood function which is chosen to be the conditional probability density function

$$l(R,Q,x_n,Z_n) = f(x_n,Z_n|R,Q) \qquad (3.3.16)$$

where $f(x_n, Z_n | R,Q)$ is the joint conditional probability density function of the state x_n and the measurements Z_n given R and Q. By application of Bayes' rule

$$f(x_n, Z_n | R,Q) = f(x_n | Z_n, R,Q) \; f(Z_n | R,Q) \tag{3.3.17}$$

Define $$L_n(R,Q,x_n,Z_n) = \ln l(R,Q,x_n,Z_n) \tag{3.3.18}$$

The set of parameters to be estimated is now

$$\alpha^T = (x_n^T, \; \xi^T)$$

Using (2.3.22) and (3.3.5) it can be seen that

$$L_n(R,Q,x_n,Z_n) = \text{constant} - 1/2 \Big[\ln|P_{n|n}| + \Delta x_n^T P_{n|n}^{-1} \Delta x_n \tag{3.3.19}$$
$$+ \sum_{i=1}^{n} \ln|B_i| + \Delta z_i^T B_i^{-1} \Delta z_i \Big]$$

where $$\Delta x_n \triangleq x_n - \hat{x}_{n|n}$$

and "constant" includes all terms that are not functions of x_n, R, or Q.

The likelihood equations are obtained by equating the derivatives of L_n with respect to α to zero. Dealing first with finding the state estimate,

$$\frac{\partial L_n}{\partial x_n} = - (x_n - \hat{x}_{n|n})^T P_{n|n}^{-1} \tag{3.3.20}$$

Then the solution of

$$\left(\frac{\partial L_n}{\partial x_n}\right)_{\substack{x_n \to \hat{x}_{n|n} \\ \xi \to \hat{\xi}_n}} = 0 \qquad (3.3.21)$$

is clearly $\quad x_n \to \hat{x}_{n|n}(Z_n, \hat{R}_n, \hat{Q}_n) \qquad (3.3.22)$

This says that the maximum likelihood estimate of the state x_n after n measurements is just the maximum likelihood state estimate which uses the estimates of R and Q to compute the filter gains.

The simultaneous estimates for ξ (R and Q) are found as the solutions of

$$\left(\frac{\partial L_n}{\partial \xi}\right)_{\substack{x_n \to \hat{x}_{n|n} \\ \xi \to \hat{\xi}_n}} = 0 \qquad (3.3.23)$$

Using the identities of Appendix A, after algebraic manipulation,

$$\frac{\partial L_n}{\partial \xi^j} = -\frac{1}{2}\Bigg[\mathrm{Tr}\,[(P^{-1}_{n|n} - P^{-1}_{n|n}\Delta x_n \Delta x_n^T P^{-1}_{n|n})\frac{\partial P_{n|n}}{\partial \xi^j} - 2\,P^{-1}_{n|n}\Delta x_n \frac{\partial \hat{x}^T_{n|n}}{\partial \xi^j}]$$

$$+ \sum_{i=1}^{n} \mathrm{Tr}\,[(B_i^{-1} - B_i^{-1}\Delta z_i \Delta z_i^T B_i^{-1})\frac{\partial B_i}{\partial \xi^j} - 2\,B_i^{-1}\Delta z_i \frac{\partial \hat{x}^T_{i|i-1}}{\partial \xi^j} H_i^T]\Bigg] \qquad (3.3.24)$$

Substituting the solution of (3.3.21) into (3.3.24),

$$\left(\frac{\partial L_n}{\partial \xi^j}\right)_{\substack{x_n\to\hat{x}_{n|n}\\ \xi\to\hat{\xi}_n}} = -\frac{1}{2}\left[\mathrm{Tr}\left(P_{n|n}^{-1}\frac{\partial P_{n|n}}{\partial \xi^j}\right) + \sum_{i=1}^{n}\mathrm{Tr}[(B_i^{-1}-B_i^{-1}\Delta z_i\Delta z_i^T B_i^{-1})\frac{\partial B_i}{\partial \xi^j}\right.$$

$$\left. - 2\, B_i^{-1}\Delta z_i \frac{\partial \hat{x}^T_{i|i-1}}{\partial \xi^j} H_i^T]\right]_{\xi\to\hat{\xi}_n} = 0 \qquad (3.3.25)$$

As before there is no general closed form solution of (3.3.25) for $\hat{\xi}_n$ so some iterative procedure must be employed. However, when there is no driving noise ($Q = 0$) a considerable simplification occurs.

By use of Bayes' rule, the likelihood function (3.3.16) can be rewritten in the following form.

$$f(x_n,Z_n|R,Q) = f(x_n,Z_{n-1}|R,Q)\; f(z_n|Z_{n-1},x_n,R,Q) \qquad (3.3.26)$$

By repeated application of Bayes' rule, it can be shown that

$$f(x_n,Z_n|R,Q) = f(x_n|R,Q) \prod_{i=1}^{n} f(z_i|Z_{i-1},x_n,R,Q) \qquad (3.3.27)$$

When $Q = 0$, it is easy to show that

$$f(x_n|R,Q) = \frac{1}{(2\pi)^{\beta/2}|P_{n|o}|^{1/2}}\, e^{-\frac{1}{2}[(x_n-\hat{x}_{n|o})^T P_{n|o}^{-1}(x_n-\hat{x}_{n|o})]}$$

where

$$\hat{x}_{n|o} = \Phi(n,o)\hat{x}_{o|o}$$

$$P_{n|o} = \Phi(n,0)P_{o|o}\Phi^T(n,0)$$

and $$f(z_i|Z_{i-1},x_n,R,Q) = \frac{1}{(2\pi)^{\beta/2}|R|^{1/2}} e^{-\frac{1}{2}(\Delta z^T_{i|n}R^{-1}\Delta z_{i|n})}$$

where $$\Delta z_{i|n} \triangleq z_i - H_i\ \Phi(i,n)x_n$$

Then (3.3.19) becomes

$$L_n(R,Q,x_n,Q_n) = \text{constant} - \frac{1}{2}\Big[\ln|P_{n|o}| + (x_n-\hat{x}_{n|o})^T P^{-1}_{n|o}(x_n-\hat{x}_{n|o})$$
$$+ \sum_{i=1}^{n} \ln|R| + (z_i-H_i\Phi(i,n)x_n)^T R^{-1}(z_i-H_i\Phi(i,n)x_n)\Big] \quad (3.3.28)$$

Then

$$\frac{\partial L_n}{\partial x_n} = -(x_n-\hat{x}_{n|o})^T P^{-1}_{n|o} + \sum_{i=1}^{n}(z_i-H_i\Phi(i,n)x_n)^T R^{-1}H_i\Phi(i,n) \quad (3.3.29)$$

Define $$F_{n|n}(\hat{R}_n) = \sum_{i=1}^{n}\Phi^T(i,n)H_i^T\hat{R}_n^{-1}H_i\Phi(i,n)$$

Then after algebraic manipulation, the solution of (3.3.21) for $\hat{x}_{n|n}$ is

$$\hat{x}_{n|n} = (P^{-1}_{n|o} + F_{n|n})^{-1}(P^{-1}_{n|o}\hat{x}_{n|o} + \sum_{i=1}^{n}\Phi^T(i,n)H_i^T\hat{R}_n^{-1}z_i) \quad (3.3.30)$$

Using the identities of Appendix A, it can be shown that

$$\frac{\partial L_n}{\partial \xi^j} = -\frac{1}{2}\sum_{i=1}^{n}\mathrm{Tr}[(R^{-1} - R^{-1}\Delta z_{i|n}\Delta z^T_{i|n}\ R^{-1})\frac{\partial R}{\partial \xi^j}] \quad (3.3.31)$$

The solution of (3.3.25) for $\hat{R}_n^{jj}$ then becomes

$$\hat{R}_n^{jj} = \frac{1}{n} \sum_{i=1}^{n} [(z_i - H_i\Phi(i,n)\hat{x}_{n|n})(z_i - H_i\Phi(i,n)\hat{x}_{n|n})^T]^{jj} \qquad (3.3.32)$$

A closed form solution of (3.3.30) and (3.3.32) for $\hat{x}_{n|n}$ and $\hat{R}_n$ is not possible except in the trivial case of a scalar measurement and when no a priori information about the state is used. In this case, $P_{n|o}^{-1} = 0$ and (3.3.30) becomes

$$\hat{x}_{n|n} = \left[\sum_{i=1}^{n} \Phi^T(i,n)H_i^TH_i\Phi(i,n)\right]^{-1} \sum_{i=1}^{n} \Phi^T(i,n)H_i^Tz_i \qquad (3.3.33)$$

From (3.3.33) it can be seen that $\hat{x}_{n|n}$ is not a function of $\hat{R}_n$ so that $\hat{x}_{n|n}$ can be computed independently of what value of $\hat{R}_n$ is obtained from (3.3.32).

In any other case a numerical solution of (3.3.30) and (3.3.32) must be performed. However, even if a closed form solution is not obtained, the estimation equations in this no driving noise case have a particularly simple form.

Estimation of the Noise Covariance Parameters and the System State with A Priori Noise Covariance Information

In this problem the state and noise covariance parameters are to be simultaneously estimated when a priori information about R and Q is used. The maximum likelihood estimate of these quantities in this case is defined by

$$1^A(\hat{R}_n,\hat{Q}_n,\hat{x}_{n|n},Z_n) = \max_{R,Q,x_n} 1^A(R,Q,x_n,Z_n) \qquad (3.3.34)$$

where $l^A(R,Q,x_n,Z_n)$ is the augmented likelihood function which is chosen to be the conditional probability density function

$$l^A(R,Q,x_n,Z_n) = f(R,Q,x_n|Z_n) \qquad (3.3.35)$$

By use of Bayes' rule

$$f(R,Q,x_n|Z_n) = f(x_n|R,Q,Z_n)\ f(R,Q|Z_n) \qquad (3.3.36)$$

From (3.3.10)

$$f(R,Q|Z_n) = \frac{f(Z_n|R,Q)\ f(R,Q)}{f(Z_n)}$$

Assuming that all the diagonal elements of R and Q are mutually independent, it can be shown that

$$L_n^A(R,Q,x_n,Z_n) = \ln l^A(R,Q,x_n,Z_n) = L_n(R,Q,x_n,Z_n) + \sum_{i=1}^{\gamma} \ln f(R^{ii}) + \sum_{i=1}^{\eta} \ln f(Q^{ii}) \qquad (3.3.37)$$

So

$$\frac{\partial L_n^A(R,Q,x_n,Z_n)}{\partial\alpha} = \frac{\partial L_n(R,Q,x_n,Z_n)}{\partial\alpha} + \frac{\partial \ln f(\xi)}{\partial\alpha} \qquad (3.3.38)$$

where $\frac{\partial L_n(R,Q,x_n,Z_n)}{\partial\alpha}$ is given by (3.3.20) and (3.3.24).

It can be seen that the likelihood equation for the state is unchanged by the inclusion of a priori information about ξ

since f(ξ) is not a function of x_n. The likelihood equations for the noise covariance parameters are modified by the addition of the term related to the a priori probability density function of the parameters ξ.

Several comments should be made about the four problems just discussed. In each problem it was assumed that a priori information about the state was used in forming the state estimates. This assumption greatly simplifies the formulation and solution of the problem while not being unreasonably restrictive. If the initial state estimate is believed to be of poor quality, then setting its covariance to a large positive definite matrix will effectively result in not using the a priori information about the state. The assumption that the initial state uncertainty has a normal distribution is a realistic assumption in most applications.

However, it was felt that a distinction should be made between noise covariance estimators which do or do not use a priori information about these parameters. The derivation of the estimation equations with no a priori noise covariance information is important because an arbitrary selection of an a priori distribution of these quantities does not have to be made. The proper choice of a distribution for the covariance parameters is much less clear than was the case in choosing a distribution of the initial state estimation error. The case of no a priori information could be handled within the framework of the estimator that uses a priori information by setting the covariance of the a priori noise covariance parameter distribution to a large quantity but with relatively little additional effort the two cases can be treated separately.

The most physically motivated problem is the last of the four given above, that of maximizing the joint conditional probability density function of the state and noise covariance parameters. The solution of this problem gives the most probable values of the state and noise covariances based upon the measurements and the a priori information. However, as will be seen, the asymptotic behavior of the solution of this problem is most easily obtained in terms of the asymptotic behavior of the simpler problem of estimating the noise covariance parameters alone. This is the primary motivation for separately treating these two problems.

3.4 Asymptotic Properties of Noise Covariance and System State Maximum Likelihood Estimators

In Section 3.2 the asymptotic properties of a restricted set of maximum likelihood estimators were given, namely that class of estimators for which the measurements were independent and identically distributed. Now the asymptotic properties of four maximum likelihood estimators that do not fit in the above category are sought.

1) noise covariance estimation with no a priori information

2) noise covariance estimation with a priori information

3) noise covariance and system state estimation with no a priori information

4) noise covariance and system state estimation with a priori information

As will be shown, if the asymptotic properties of the first of the above estimators are found, the properties of the other three follow immediately. Therefore, the asymptotic properties of the noise covariance estimator with no a priori information will be found first.

The maximum likelihood estimate of R and Q was defined as the solution of (3.3.7). Define the single measurement score

$$S^j(Z_i,\xi) = -\frac{1}{2}\,\mathrm{Tr}[(B_i^{-1}-B_i^{-1}\Delta z_i \Delta z_i^T B_i^{-1})\frac{\partial B_i}{\partial \xi^j} - 2\, B_i^{-1}\Delta z_i \frac{\partial \hat{x}_{i|i-1}^T}{\partial \xi^j} H_i^T] \tag{3.4.1}$$

(3.4.1) differs from the single measurement score of Section 3.2 because it is a function of all measurements up to and including the i^{th} measurement. Define the total measurement score

$$S_n^j(Z_n,\xi) = \sum_{i=1}^{n} S^j(Z_i,\xi) \tag{3.4.2}$$

and

$$[J(\xi_o,\xi)]^{jk} = \varepsilon[S^j(Z_i,\xi)S^k(Z_i,\xi)\,|\,\xi_o] \tag{3.4.3}$$

$$= \int\!\cdots\!\int S^j(Z_i,\xi)S^k(Z_i,\xi)\ f(Z_i|\xi_o)dZ_i$$

$$[J_n(\xi_o,\xi)]^{jk} = \varepsilon[S_n^j(Z_n,\xi)S_n^k(Z_n,\xi)\,|\,\xi_o] \tag{3.4.4}$$

$$= \int\!\cdots\!\int S_n^j(Z_n,\xi)S_n^k(Z_n,\xi)\ f(Z_n|\xi_o)dZ_n$$

$J(\xi_o) = J(\xi_o,\xi_o)$ the single measurement conditional information matrix

$J_n(\xi_o) = J_n(\xi_o,\xi_o)$ the total measurement conditional information matrix

Then the likelihood equations (3.3.6) become

$$\frac{\partial L_n}{\partial \xi^j} = S_n^j(Z_n,\xi) = \sum_{i=1}^{n} S^j(Z_i,\xi) \qquad (3.4.5)$$

It can be shown that when $\xi = \xi_o$, the true value of the parameters, the measurement residuals Δz_i are zero mean normal variables with covariance B_i, with the further property that the residuals at different times are independent. Or

$$\varepsilon(\Delta z_i | \xi_o) = 0$$

$$\varepsilon(\Delta z_i \Delta z_l^T | \xi_o) = B_i(\xi_o)\ \delta_{il}$$

It can also be shown that

$$\varepsilon[\mathrm{Tr}(B_i^{-1}\Delta z_i \frac{\partial \hat{x}^T_{i|i-1}}{\partial \xi^j} H_i^T) | \xi_0] = 0$$

$$\varepsilon\left[\mathrm{Tr}(B_i^{-1}\Delta z_i \frac{\partial \hat{x}^T_{i|i-1}}{\partial \xi^j} H_i^T)\ \mathrm{Tr}(B_l^{-1}\Delta z_l \frac{\partial \hat{x}^T_{l|l-1}}{\partial \xi^k} H_l^T) | \xi_o\right] = \mathrm{Tr}(B_i^{-1} H_i G^{jk}_{i|i-1} H_i^T)\,\delta_{il}$$

where $$G^{jk}_{i\ i-1} = \varepsilon\left[\frac{\partial \hat{x}_{i|i-1}}{\partial \xi^j}\ \frac{\partial \hat{x}^T_{i|i-1}}{\partial \xi^k}\ | \xi_o\right]$$

and $\varepsilon \left[\mathrm{Tr}[(B_i^{-1} - B_i^{-1} \Delta z_i \Delta z_i^T B_i^{-1}) \frac{\partial B_i}{\partial \xi^j}] \; \mathrm{Tr}(B_l^{-1} \Delta z_l \; \frac{\partial \hat{x}^T_{l|l-1}}{\partial \xi^k} \; H_l^T) \,|\, \xi_o \right] = 0$

Therefore, after algebraic manipulation it can be shown that

$$\varepsilon [S^j(Z_i, \xi_o) \,|\, \xi_o] = 0 \qquad (3.4.6)$$

$$\varepsilon [S^j(Z_i, \xi_o) S^k(Z_l, \xi_o) \,|\, \xi_o] = \frac{1}{2} \mathrm{Tr}(B_i^{-1} \frac{\partial B_i}{\partial \xi^j} B_i^{-1} \frac{\partial B_i}{\partial \xi^k}) \delta_{il} \qquad (3.4.7)$$

$$+ \; \mathrm{Tr}(B_i^{-1} H_i G^{jk}_{i|i-1} H_i^T) \; \delta_{il}$$

From (3.4.7) it can be seen that $S(Z_i, \xi_o)$ is independent of $S(Z_l, \xi_o)$ for $i \neq l$. Then it follows immediately that

$$\varepsilon [S_n^j(Z_n, \xi_o) \,|\, \xi_o] = 0 \qquad (3.4.8)$$

$$\varepsilon [S_n^j(Z_n, \xi_o) S_n^k(Z_n, \xi_o) \,|\, \xi_o] = \frac{1}{2} \sum_{i=1}^{n} \left[\mathrm{Tr}(B_i^{-1} \frac{\partial B_i}{\partial \xi^j} \; B_i^{-1} \frac{\partial B_i}{\partial \xi^k}) \right. \qquad (3.4.9)$$

$$\left. + \; 2 \; \mathrm{Tr}(B_i^{-1} H_i G^{jk}_{i|i-1} H_i^T) \right]$$

(3.4.7) and (3.4.9) represent respectively the single and total measurement conditional information matrices.

Because of the independence of the measurement residuals when $\xi = \xi_o$ and the other relationships shown above, the asymptotic properties of the maximum likelihood noise covariance estimator can be found relatively easily. These properties are quite similar to those mentioned in Section 3.2 even though the measurements are not now identically distributed.

Asymptotic Distribution of the Score

Suppose $(z_1,..,z_n)$ is a sample from the probability density function $f(z_i|Z_{i-1},\xi_o)$. Let $f(z_i|Z_{i-1},\xi)$ possess finite first derivatives with respect to ξ in the range Ω. Then if $J_n(\xi,\xi)$ is positive definite for ξ in Ω, $S_n(Z_n,\xi_o)$ is asymptotically distributed for large n as a zero mean normal random variable with covariance $J_n(\xi_o)$.

Proof: It has already been shown that $S_n(Z_n,\xi_o)$ is a zero mean random variable with covariance $J_n(\xi_o)$. Now all that remains to show is that S_n is asymptotically normally distributed. From the definition of $S_n(Z_n,\xi_o)$,

$$S_n(Z_n,\xi_o) = \sum_{i=1}^{n} S(Z_i,\xi_o)$$

It was shown that $S(Z_i,\xi_o)$ was independent of $S(Z_1,\xi_o)$ for $i \neq 1$. If it is assumed that no term dominates the above sum by having a large value with appreciable probability, then by use of the central limit theorem concerning the sum of independent random variables, the score $S_n(Z_n,\xi_o)$ can be shown to be asymptotically normally distributed for large n.

Convergence of the Maximum Likelihood Estimator

Suppose $(z_1,..,z_n)$ is a sample from the probability density function $f(z_i|Z_{i-1},\xi_o)$. Let $f(z_i|Z_{i-1},\xi)$ possess finite first derivatives with respect to ξ in Ω. Let $S^j(Z_i,\xi)$ be a continuous function of ξ in Ω for all values of Z_i

except possibly for a set of zero probability. If as $n \to \infty$,

$$[J_n(\xi_o)]^{-1} \to 0$$

then there exists a sequence of solutions of

$$S_n^j(Z_n,\xi) = 0 \qquad (3.4.10)$$

which converges in probability to ξ_o. If for $n \geq$ some n_o the solution is a unique vector $\hat{\xi}_n$, the sequence of vectors converges in probability to ξ_o as $n \to \infty$.

Proof: Define

$$A_i^j(\xi_o,\xi) = \varepsilon[S^j(Z_i,\xi)|\xi_o]$$

$$= \int\!\!..\!\!\int S^j(Z_i,\xi)\ f(Z_i|\xi_o)\ dZ_i$$

$$\overline{A^j(\xi_o,\xi)} = \frac{1}{n} \sum_{i=1}^{n} A_i^j(\xi_o,\xi)$$

Then $\frac{1}{n} S_n^j(Z_n,\xi)$ is the mean of a sample of size n from a population having mean $\overline{A^j(\xi_o,\xi)}$ if ξ_o is the true value of ξ. From the weak law of large numbers, $\frac{1}{n} S_n^j$ converges in probability to $\overline{A^j(\xi_o,\xi)}$. Without loss of generality, define Ω' to be $(\xi_o-\delta,\ \xi_o+\delta)$ with $\delta > 0$. It can be shown that $\overline{A^j(\xi_o,\xi)}$ is monotonically decreasing over this interval, and since $A^j(\xi_o,\xi_o) = 0$,

$$\overline{A^j(\xi_o, \xi_o-\delta)} > 0$$

$$\overline{A^j(\xi_o, \xi_o+\delta)} < 0$$

Therefore there exists an $n(\delta,\varepsilon)$ so that the probability exceeds $1 - \varepsilon$ that both of the following inequalities hold for any $n > n(\delta,\varepsilon)$ if ξ_o is the true value of ξ.

$$\frac{1}{n} S_n^j(Z_n,\xi) > 0 \qquad \text{if } \xi = \xi_o-\delta$$

$$\frac{1}{n} S_n^j(Z_n,\xi) < 0 \qquad \text{if } \xi = \xi_o+\delta$$

Since $S^j(Z_i,\xi)$ is continuous in ξ over Ω for all Z_i except for a set of probability zero, a similar statement holds for $\frac{1}{n} S_n^j(Z_n,\xi)$. Therefore, for any fixed $n > n(\delta,\varepsilon)$ for some in Ω',

$$P[\frac{1}{n} S_n^j(Z_n,\xi) = 0|\xi_o] > 1 - \varepsilon$$

This is equivalent to the statement that a sequence of roots of (3.4.10) exists which converges in probability to ξ_o. In particular if (3.4.10) has a unique solution $\hat{\xi}_n$ for $n = n_o$, $n_o + 1,..,$ for some integer n_o, then the sequence $\hat{\xi}_n$, $n > n_o$, converges in probability to ξ_o.

Asymptotic Distribution of the Maximum Likelihood Estimator

If $(z_1,..,z_n)$ is a sample from the probability density function $f(z_i|Z_{i-1},\xi_o)$ where $f(z_i|Z_{i-1},\xi)$ possesses finite first and second derivatives with respect to ξ in the range Ω, and if the maximum likelihood estimator $\hat{\xi}_n$ satisfying (3.4.10) is unique for $n \geq$ some n_o, then it is asymptotically normally distributed for large n with mean ξ_o and covaraince $[J_n(\xi_o)]^{-1}$.

Proof: First it will be shown that

$$J_n(\xi_o)(\hat{\xi}_n-\xi_o) - S_n^T(Z_n,\xi_o) \rightarrow 0 \qquad (3.4.11)$$

with large probability. This will then be used to show that $\hat{\xi}_n$ is an efficient estimator and the asymptotic distribution of $(\hat{\xi}_n-\xi_o)$ is normal with zero mean and covariance $[J_n(\xi_o)]^{-1}$.

Since $\hat{\xi}_n$ satisfies the likelihood equation

$$S_n^j(Z_n,\hat{\xi}_n) = 0$$

then by a Taylor series expansion of S_n^j at ξ_o,

$$S_n^j(Z_n,\hat{\xi}_n) = 0 \simeq S_n^j(Z_n,\xi_o) + \left(\frac{\partial S_n^j}{\partial \xi^k}\right)_{\xi_o} \Delta\xi_o^k + \frac{1}{2}\left(\frac{\partial^2 S_n^j}{\partial \xi^k \partial \xi^l}\right)_{\xi_o} \Delta\xi_o^k \Delta\xi_o^l \qquad (3.4.12)$$

where $\Delta\xi_o = \hat{\xi}_n - \xi_o$

Here as elsewhere index summation notation is used. If an index appears more than once on the right side of an equation with no comparable index on the left side, a summation over that index is implied.

Define

$$C_n^{jk}(Z_n,\xi_o) = \left(\frac{\partial S_n^j}{\partial \xi^k}\right)_{\xi_o} + \frac{1}{2}\left(\frac{\partial^2 S_n^j}{\partial \xi^k \partial \xi^l}\right)_{\xi_o} \Delta\xi_o^l$$

Then (3.4.12) becomes

$$0 = S_n^T(Z_n,\xi_o) + C_n(Z_n,\xi_o)\Delta\xi_o$$

Assuming that C_n is of full rank,

$$\Delta\xi_o = - [C_n(Z_n,\xi_o)]^{-1} S_n^T(Z_n,\xi_o) \qquad (3.4.13)$$

Define $\qquad b_n = J_n(\xi_o)[C_n(Z_n,\xi_o)]^{-1}$

Multiplying (3.4.13) by $J_n(\xi_o)$ and rearranging terms,

$$J_n(\xi_o)\Delta\xi_o - S_n^T(Z_n,\xi_o) = - (b_n + I) S_n^T(Z_n,\xi_o) \qquad (3.4.14)$$

It will now be shown that $b_n \to - I$ with large probability, in which case the right hand side of (3.4.14) → 0, establishing the desired result.

As before, define

$$L_n(Z_n,\xi) = \ln f(Z_n|\xi)$$

and $\qquad f_n = f(Z_n|\xi)$

Now define

$$D_n(\xi_o,\xi) = \varepsilon\left[\frac{1}{f_n}\frac{\partial^2 f_n}{\partial\xi\partial\xi}\,|\xi_o\right]$$

$$= \int\!\!\cdot\!\!\cdot\!\!\int \frac{1}{f_n(Z_n,\xi)}\frac{\partial^2 f_n(Z_n,\xi)}{\partial\xi\partial\xi} f_n(Z_n,\xi_o)dZ_n$$

Then $$D_n(\xi_o,\xi_o) = \int\!\!\cdot\!\!\int \frac{\partial^2 f_n(Z_n,\xi_o)}{\partial\xi_o\partial\xi_o}\, dZ_n$$

Assuming that differentiation with respect to ξ_o can be taken outside the integral

$$D_n(\xi_o,\xi_o) = \frac{\partial^2}{\partial\xi_o\partial\xi_o} \int\!\!\cdot\!\!\int f_n(Z_n,\xi_o)\, dZ_n$$

$$= \frac{\partial^2}{\partial\xi_o\partial\xi_o} (1) = 0$$

But $$\frac{1}{f_n}\frac{\partial^2 f_n}{\partial\xi\partial\xi} = \frac{\partial^2 L_n}{\partial\xi\partial\xi} + \frac{1}{f_n^2}\left(\frac{\partial f_n}{\partial\xi}\right)^T \frac{\partial f_n}{\partial\xi}$$

$$= \frac{\partial S_n(Z_n,\xi)}{\partial\xi} + S_n^T(Z_n,\xi)\; S_n(Z_n,\xi)$$

So $$D_n(\xi_o,\xi_o) = 0 = \varepsilon\left[\frac{\partial S_n(Z_n,\xi_o)}{\partial\xi_o} \;|\xi_o\right] + J_n(\xi_o)$$

Or $$\varepsilon\left[\frac{\partial S_n(Z_n,\xi_o)}{\partial\xi_o} \;|\xi_o\right] = - J_n(\xi_o)$$

But $$\frac{\partial S_n(Z_n,\xi_o)}{\partial\xi_o} = \sum_{i=1}^{n} \frac{\partial S(Z_i,\xi_o)}{\partial\xi_o}$$

As n becomes large, by application of the strong law of large numbers, it can be shown that

$$\frac{\partial S_n(Z_n,\xi_o)}{\partial\xi_o} \rightarrow \varepsilon\left[\frac{\partial S_n(Z_n,\xi_o)}{\partial\xi_o} \;|\xi_o\right] = - J_n(\xi_o)$$

Analogous to the assumption made in Section 3.2, it is assumed that

$$\frac{1}{n}\left|\frac{\partial^3 L_n}{\partial\xi^j\partial\xi^k\partial\xi^l}\right| < K$$

with large probability as n → ∞, where K is independent of ξ and n. Since $\Delta\xi \to 0$, the product

$$\frac{1}{2n}\left(\frac{\partial^2 S_n^j}{\partial\xi^k\partial\xi^l}\right)_{\xi_o} \Delta\xi_o^l \to 0$$

with large probability. Assuming that for large n,

$$J_n(\xi_o) \geq n\ K_1$$

where K_1 is a positive definite matrix independent of n, then

$$C_n(Z_n,\xi_o) \to -\ J_n(\xi_o)$$

and $\qquad b_n \to -\ I \qquad$ with large probability

Thus it has been shown that

$$J_n(\xi_o)(\hat{\xi}_n-\xi_o) - S_n^T(Z_n,\xi_o) \to 0 \qquad (3.4.15)$$

It has already been shown that $S_n(Z_n,\xi_o)$ is asymptotically distributed as a normal random variable with zero mean and covariance $J_n(\xi_o)$. From this and (3.4.15) it can be concluded that $(\hat{\xi}_n-\xi_o)$ is normally distributed with zero mean and covariance $[J_n(\xi_o)]^{-1}$ as $n \to \infty$.

Wilks has shown that (3.4.15) is a necessary and sufficient condition for stating that $\hat{\xi}_n$ is an asymptotically efficient estimator for ξ_o.

Thus it has been shown that the maximum likelihood estimator for the noise covariance parameters using no a priori information about these parameters is: 1) consistent, 2) asymptotically unbiased, 3) asymptotically normally distributed, and 4) asymptotically efficient. Now the asymptotic properties of the three closely related estimators previously mentioned are sought.

If a priori information about ξ is used, the maximum likelihood estimator was defined to be the solution of (3.3.14).

$$\left(\frac{\partial L_n^A(\xi,Z_n)}{\partial\xi}\right)_{\hat{\xi}_1} = \left(\frac{\partial L_n(\xi,Z_n)}{\partial\xi}\right)_{\hat{\xi}_1} + \left(\frac{\partial \ln f(\xi)}{\partial\xi}\right)_{\hat{\xi}_1} = 0 \qquad (3.4.16)$$

The estimator in the absence of a priori information is the solution of (3.3.6).

$$\left(\frac{\partial L_n(\xi,Z_n)}{\partial\xi}\right)_{\hat{\xi}_2} = 0 \qquad (3.4.17)$$

where $\hat{\xi}_1$ is the estimator using a priori information and $\hat{\xi}_2$ is the estimator without using a priori information. Expanding (3.4.16) in a Taylor series about $\hat{\xi}_2$,

$$\left(\frac{\partial L_n^A(\xi,Z_n)}{\partial\xi}\right)_{\hat{\xi}_1} = \left(\frac{\partial L_n^A(\xi,Z_n)}{\partial\xi}\right)_{\hat{\xi}_2} + \left(\frac{\partial^2 L_n^A(\xi,Z_n)}{\partial\xi\partial\xi}\right)_{\hat{\xi}_2}(\hat{\xi}_1-\hat{\xi}_2) \qquad (3.4.18)$$

$$+ \;\ldots\ldots \;= 0$$

But $$\left(\frac{\partial L_n^A(\xi,Z_n)}{\partial\xi}\right)_{\hat{\xi}_2} = \left(\frac{\partial \ln f(\xi)}{\partial\xi}\right)_{\hat{\xi}_2}$$

and $$\left(\frac{\partial^2 L_n^A(\xi,Z_n)}{\partial\xi\partial\xi}\right)_{\hat{\xi}_2} = \left(\frac{\partial^2 L_n(\xi,Z_n)}{\partial\xi\partial\xi}\right)_{\hat{\xi}_2} + \left(\frac{\partial^2 \ln f(\xi)}{\partial\xi\partial\xi}\right)_{\hat{\xi}_2}$$

It has been shown that for large n,

$$\left(\frac{\partial^2 L_n(\xi,Z_n)}{\partial\xi\partial\xi}\right)_{\xi_o} \simeq \varepsilon\left[\frac{\partial^2 L_n(\xi_o,Z_n)}{\partial\xi_o\partial\xi_o} \mid \xi_o\right]$$

But $$\varepsilon \frac{\partial^2 L_n(\xi_o,Z_n)}{\partial\xi_o\partial\xi_o} \mid \xi_o = - J_n(\xi_o)$$

so $$\left(\frac{\partial^2 L_n(\xi,Z_n)}{\partial\xi\partial\xi}\right)_{\xi_o} \simeq - J_n(\xi_o)$$

It has already been assumed that as n → ∞, $[J_n(\xi_o)]^{-1} \rightarrow 0$. Now the assumption is made that $\hat{\xi}_2$ is sufficiently close to ξ_o so that the following approximation is valid.

$$\left(\frac{\partial^2 L_n(\xi,Z_n)}{\partial\xi\partial\xi}\right)_{\hat{\xi}_2} \simeq \left(\frac{\partial^2 L_n(\xi,Z_n)}{\partial\xi\partial\xi}\right)_{\xi_o} = - J_n(\xi_o)$$

Then $$\left(\frac{\partial^2 L_n^A(\xi,Z_n)}{\partial\xi\partial\xi}\right)_{\hat{\xi}_2} \simeq - J_n(\xi_o) + \left(\frac{\partial^2 \ln f(\xi)}{\partial\xi\partial\xi}\right)_{\hat{\xi}_2} \qquad (3.4.19)$$

It is also assumed that as n → ∞, $- J_n(\xi_o)$ dominates in (3.4.19) so that

$$\left(\frac{\partial^2 L_n^A(\xi,Z_n)}{\partial\xi\partial\xi}\right)_{\hat{\xi}_2} \simeq - J_n(\xi_o)$$

and then (3.4.18) becomes

$$\left(\frac{\partial \ln f(\xi)}{\partial \xi}\right)_{\hat{\xi}_2} - J_n(\xi_o)\ (\hat{\xi}_1 - \hat{\xi}_2) = 0$$

The first linear correction to the solution $\hat{\xi}_2$ due to inclusion of a priori information is then

$$\hat{\xi}_1 = \hat{\xi}_2 + [J_n(\xi_o)]^{-1} \left(\frac{\partial \ln f(\xi)}{\partial \xi}\right)_{\hat{\xi}_2}$$

But as $n \to \infty$, $[J_n(\xi_o)]^{-1} \to 0$, so assuming that all elements of $\left(\frac{\partial \ln f(\xi)}{\partial \xi}\right)_{\hat{\xi}_2}$ are finite,

$$\hat{\xi}_1 \to \hat{\xi}_2 \text{ as } n \to \infty$$

Therefore, under a wide set of conditions, the estimator which utilizes a priori information behaves asymptotically as the estimator which does not utilize this a priori information.

If the state and noise covariance parameters are estimated without a priori information about ξ, the maximum likelihood estimator was defined to be the solution of (3.3.21) and (3.3.23). Or

$$\left(\frac{\partial \ln f(x_n, Z_n | R, Q)}{\partial x_n}\right)_{\substack{x_n \to \hat{x}_{n|n} \\ \xi \to \hat{\xi}_1}} = 0 \qquad (3.4.20)$$

$$\left(\frac{\partial \ln f(x_n, Z_n | R, Q)}{\partial \xi}\right)_{\substack{x_n \to \hat{x}_{n|n} \\ \xi \to \hat{\xi}_1}} = 0 \qquad (3.4.21)$$

The estimator for ξ alone with no a priori information about ξ was defined to be the solution of (3.3.6). Or

$$\left(\frac{\partial \ln f(Z_n|R,Q)}{\partial \xi}\right)_{\xi \to \hat{\xi}_2} = 0 \qquad (3.4.22)$$

where $\hat{\xi}_1$ is the estimate of ξ found simultaneously with $\hat{x}_{n|n}$ and $\hat{\xi}_2$ is the estimate of ξ found independently.

It can be seen that

$$\frac{\partial \ln f(x_n,Z_n|R,Q)}{\partial \xi} = \frac{\partial \ln f(Z_n|R,Q)}{\partial \xi} + \frac{\partial \ln f(x_n|Z_n,R,Q)}{\partial \xi} \qquad (3.4.23)$$

Expanding (3.4.21) in a Taylor series about $\hat{\xi}_2$,

$$\left(\frac{\partial \ln f(x_n,Z_n|R,Q)}{\partial \xi}\right)_{\substack{x_n \to \hat{x}_{n|n} \\ \xi \to \hat{\xi}_1}} = \left(\frac{\partial \ln f(x_n,Z_n|R,Q)}{\partial \xi}\right)_{\substack{x_n \to \hat{x}_{n|n} \\ \xi \to \hat{\xi}_2}} \qquad (3.4.24)$$

$$+\left(\frac{\partial^2 \ln f(x_n,Z_n|R,Q)}{\partial \xi \partial \xi}\right)_{\substack{x_n \to \hat{x}_{n|n} \\ \xi \to \hat{\xi}_2}} (\hat{\xi}_1-\hat{\xi}_2)+\ldots$$

But $\left(\frac{\partial \ln f(x_n,Z_n|R,Q)}{\partial \xi}\right)_{\substack{x_n \to \hat{x}_{n|n} \\ \xi \to \hat{\xi}_2}} = \left(\frac{\partial \ln f(x_n|Z_n,R,Q)}{\partial \xi}\right)_{\substack{x_n \to \hat{x}_{n|n} \\ \xi \to \hat{\xi}_2}}$

and $\left(\frac{\partial^2 \ln f(x_n,Z_n|R,Q)}{\partial \xi \partial \xi}\right)_{\substack{x_n \to \hat{x}_{n|n} \\ \xi \to \hat{\xi}_2}} = \left(\frac{\partial^2 \ln f(Z_n|R,Q)}{\partial \xi \partial \xi}\right)_{\xi \to \hat{\xi}_2}$

$$+ \left(\frac{\partial^2 \ln f(x_n|Z_n,R,Q)}{\partial \xi \partial \xi}\right)_{\substack{x_n \to \hat{x}_{n|n} \\ \xi \to \hat{\xi}_2}}$$

It has been shown that for large n,

$$\left(\frac{\partial^2 \ln f(Z_n|\xi)}{\partial\xi\partial\xi}\right)_{\xi_o} \simeq \varepsilon\left[\frac{\partial^2 \ln f(Z_n|\xi_o)}{\partial\xi\;\partial\xi}\;|\xi_o\right]$$

$$= - J_n(\xi_o)$$

where $J_n(\xi_o)$ is the conditional information matrix. Analogous to an assumption previously made, it is assumed that $\hat{\xi}_2$ is sufficiently close to ξ_o so that

$$\left(\frac{\partial^2 \ln f(Z_n|R,Q)}{\partial\xi\partial\xi}\right)_{\xi\to\hat{\xi}_2} \simeq - J_n(\xi_o)$$

Then $$\left(\frac{\partial^2 \ln f(x_n,Z_n|R,Q)}{\partial\xi\partial\xi}\right)_{\substack{x_n\to\hat{x}_{n|n}\\ \xi\to\hat{\xi}_2}} \simeq - J_n(\xi_o) + \left(\frac{\partial^2 \ln f(x_n|Z_n,R,Q)}{\partial\xi\partial\xi}\right)_{\substack{x_n\to\hat{x}_{n|n}\\ \xi\to\hat{\xi}_2}}$$

(3.4.25)

But $$\left(\frac{\partial^2 \ln f(x_n|Z_n,R,Q)}{\partial\xi\partial\xi}\right)_{\substack{x_n\to\hat{x}_{n|n}\\ \xi\to\hat{\xi}_2}} = - \frac{1}{2}\left[\,\mathrm{Tr}(P^{-1}_{n|n}\frac{\partial^2 P_{n|n}}{\partial\xi\partial\xi}\right.$$

$$\left. - \mathrm{Tr}(P^{-1}_{n|n}\frac{\partial P_{n|n}}{\partial\xi}\;P^{-1}_{n|n}\frac{\partial P_{n|n}}{\partial\xi}) + 2\;\mathrm{Tr}(P^{-1}_{n|n}\;\frac{\partial \hat{x}_{n|n}}{\partial\xi}\;\frac{\hat{x}^T_{n|n}}{\partial\xi})\right]_{\hat{\xi}_2}$$

Assuming that as n → ∞, $- J_n(\xi_o)$ dominates in (3.4.25),

$$\left(\frac{\partial^2 \ln f(x_n,Z_n|R,Q)}{\partial\xi\partial\xi}\right)_{\substack{\xi\to\hat{\xi}_2\\ x_n\to\hat{x}_{n|n}}} \simeq - J_n(\xi_o)$$

and then (3.4.24) becomes

$$\left(\frac{\partial \ln f(x_n | Z_n, R, Q)}{\partial \xi}\right)_{\substack{x_n \to \hat{x}_{n|n} \\ \xi \to \hat{\xi}_2}} - J_n(\xi_o)(\hat{\xi}_1 - \hat{\xi}_2) = 0$$

But $$\left(\frac{\partial \ln f(x_n | Z_n, R, Q)}{\partial \xi}\right)_{\substack{x_n \to \hat{x}_{n|n} \\ \xi \to \hat{\xi}_2}} = -\frac{1}{2}\left[\mathrm{Tr}\left(P_{n|n}^{-1} \frac{\partial P_{n|n}}{\partial \xi}\right)\right]_{\hat{\xi}_2}$$

The first linear correction to the solution $\hat{\xi}_1$ due to simultaneously estimating the state is then

$$\hat{\xi}_1 = \hat{\xi}_2 - \frac{1}{2}[J_n(\xi_o)]^{-1}\left[\mathrm{Tr}\left(P_{n|n}^{-1} \frac{\partial P_{n|n}}{\partial \xi}\right)\right]_{\hat{\xi}_2}$$

But as n → ∞, $[J_n(\xi_o)]^{-1} \to 0$, so assuming that $\left[\mathrm{Tr}\left(P_{n|n}^{-1} \frac{\partial P_{n|n}}{\partial \xi}\right)\right]_{\hat{\xi}_2}$ remains finite,

$$\hat{\xi}_1 \to \hat{\xi}_2 \text{ as } n \to \infty$$

Therefore, the estimator of ξ when the state is also estimated behaves asymptotically as the estimator which does not simultaneously estimate the state. As was shown, the estimator of ξ alone converges to the true value of ξ_o so that the state estimator which then uses this estimated value of ξ converges to the true maximum likelihood state estimator discussed in Chapter 2.

Using similar arguments, the inclusion of a priori information about ξ in the simultaneous state and noise covariance parameter estimator does not affect its asymptotic properties.

3.5 Selection of the A Priori Noise Covariance Distribution

The choice of f(R) and f(Q) is somewhat arbitrary as these functions are introduced so that uncertainty in knowledge of R and Q can be properly treated. However, once selected, they can strongly influence the solutions of the likelihood equations. They must be selected to realistically represent possible variations in the values of R and Q while not being mathematically intractable. Caution should be observed in their selection because the simplest and seemingly realistic distributions may be unsuited for use in a maximum likelihood estimator.

Suppose that f(R) or f(Q) is defined to be nonzero only over some finite range of R or Q and is zero outside this range. Then all solutions of the likelihood equations for R and Q must also lie within this range. This can be seen by considering the following example.

Let $f(z|\xi)$ be the conditional probability density function of a random variable z, assumed to be normally distributed with zero mean and variance ξ. Let $f(\xi)$ be the a priori probability density function of ξ, defined over some finite range

$$\begin{aligned} f(\xi) &= f(\xi^*) \qquad && \xi_o < \xi < \xi_1 \\ &= 0 && \text{otherwise} \end{aligned}$$

By application of Bayes' rule

$$f(\xi|z) = \frac{f(z|\xi)\ f(\xi)}{f(z)}$$

where
$$f(z) = \int_{\xi_o}^{\xi_1} f(z|\xi)\ f(\xi)\ d\xi$$

For any finite value of ξ, $f(z|\xi)$ is zero only at $z = \pm\infty$, and it is assumed that $f(\xi)$ is selected so that $f(z)$ is also zero only at $z = \pm\infty$. Then from the above it can be seen that $f(\xi|z)$ is zero outside the range (ξ_o, ξ_1). This says that regardless of the shape of $f(\xi|z)$ within the range (ξ_o, ξ_1), there can be no legitimate solutions of

$$\frac{f(\xi|z)}{\partial\xi} = 0$$

outside this range. If the range is too small and happens to exclude the true value of ξ, the maximum likelihood equations cannot have a valid solution for the true value of ξ. So if $f(R)$ and $f(Q)$ are defined only over some finite or semi-infinite range of R or Q, this range must be large enough to include all possible true values of R and Q.

Since the diagonal elements of R and Q represent variances, it is clear that the a priori probability density functions for these quantities must be zero for all negative values of the diagonal elements. From the preceding discussion it can be seen that all solutions of the likelihood equations for $\hat{R}_n^{jj}$ and $\hat{Q}_n^{jj}$ must be positive.

Perhaps the simplest possible distribution for R and Q is a rectangular distribution for any diagonal element, denoted by ξ.

$$f(\xi) = \frac{1}{\xi_1-\xi_o} \qquad \xi_o \leq \xi \leq \xi_1, \ \xi_o > 0 \qquad (3.5.1)$$

$$= 0 \qquad \text{otherwise}$$

It can be seen that this distribution does not possess finite derivatives with respect to ξ for any value of ξ. The derivatives are either zero or infinite. Therefore

$$\frac{\partial f(\xi|z)}{\partial \xi} = \frac{1}{f(z)} \left[\frac{\partial f(z|\xi)}{\partial \xi} f(\xi) + f(z|\xi) \frac{\partial f(\xi)}{\partial \xi}\right]$$

$$= \frac{f(\xi)}{f(z)} \frac{\partial f(z|\xi)}{\partial \xi} \qquad \xi \neq \xi_o \text{ or } \xi_1$$

This says that if $\xi_o < \xi < \xi_1$, then the maximum of $f(\xi|z)$ occurs at the same point as the maximum of $f(z|\xi)$ and that no valid maximum can exist outside the range (ξ_o,ξ_1). The solution for $\hat{\xi}$ in this case would be identical to the solution obtained by considering that no a priori information about the value of ξ exists, as long as such a solution is within the range (ξ_o,ξ_1).

This is the distribution that would be used, at least in theory, if the only a priori information about ξ is that ξ must be positive. In such a case

$$f(\xi) = \lim_{\substack{\xi_o \to 0 \\ \xi_1 \to \infty}} \frac{1}{\xi_1 - \xi_o} \qquad \xi_o < \xi < \xi_1$$

$$= 0 \qquad \text{otherwise}$$

It should be noted that if a rectangular distribution of ξ is used, then in the absence of any measurements, no unique maximum likelihood estimate of ξ exists. This is a consequence of the fact that all values of ξ within the range (ξ_o, ξ_1) are equally likely to occur, so that there is no preferred value from the viewpoint of maximum likelihood. If another estimation criterion is used, there may be a preferred value. In the case of a minimum variance estimation criterion, the mean of the distribution of ξ would be the minimum variance estimate.

In many situations more may be known about ξ than merely that its value lies in some range with equal probability of occurence in that range. In such situations a more complex $f(\xi)$ should be assigned. Two possible distributions are given below, a truncated normal distribution and a Gamma distribution.

Truncated Normal Distribution

If ξ has a truncated normal distribution, then its probability density function is given by

$$f(\xi) = K\, e^{-1/2\left[(\xi-\mu)^2/\sigma^2\right]} \qquad \xi_o < \xi < \xi_1 \qquad (3.5.2)$$

$$= 0 \qquad \text{otherwise}$$

where $$K = \frac{2}{(2\pi)^{1/2}\,\sigma[\mathrm{erf}(s_1) - \mathrm{erf}(s_o)]}$$

$$s_o = \frac{\xi_o-\mu}{\sigma}, \qquad s_1 = \frac{\xi_1-\mu}{\sigma}$$

erf( ) is the error function

μ is the mean of the untruncated distribution

σ^2 is the variance of the untruncated distribution

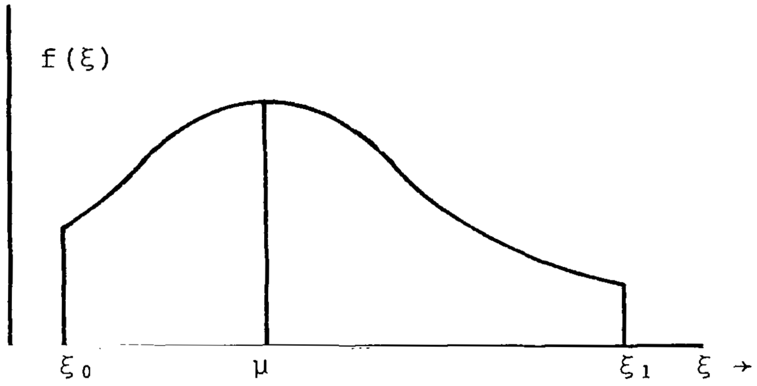

Fig. 3.1 Truncated Normal Distribution

The mean of the truncated distribution is

$$\overline{\xi} = \int_{-\infty}^{\infty} \xi\ f(\xi)\ d\xi = \mu + \Delta\mu \qquad (3.5.3)$$

where $$\Delta\mu = \sigma^2 \, K(e^{-s_1^2} - e^{-s_2^2})$$

and the variance of the truncated distribution is

$$\sigma_\xi^2 = \int_{-\infty}^{\infty} (\xi - \overline{\xi})^2 \, f(\xi) \, d\xi \qquad (3.5.4)$$

$$= \sigma^2 + \Delta\sigma^2 - \Delta\mu^2$$

where $$\Delta\sigma^2 = \sqrt{2} \, K\sigma^3 (s_o \, e^{-\frac{1}{2}s_o^2} - s_1 \, e^{-\frac{1}{2}s_1^2})$$

Gamma Distribution

If ξ has a Gamma distribution, then its probability density function is given by

$$f(\xi) = \frac{a}{\mu\Gamma(a)} \left[\frac{a\xi}{\mu}\right]^{a-1} e^{-a\xi/\mu} \qquad \xi \geq 0 \qquad (3.5.5)$$

$$= 0 \qquad \xi < 0$$

where a and μ are parameters of the distribution, and $a > 0$. $\Gamma(a)$ is the Gamma function.

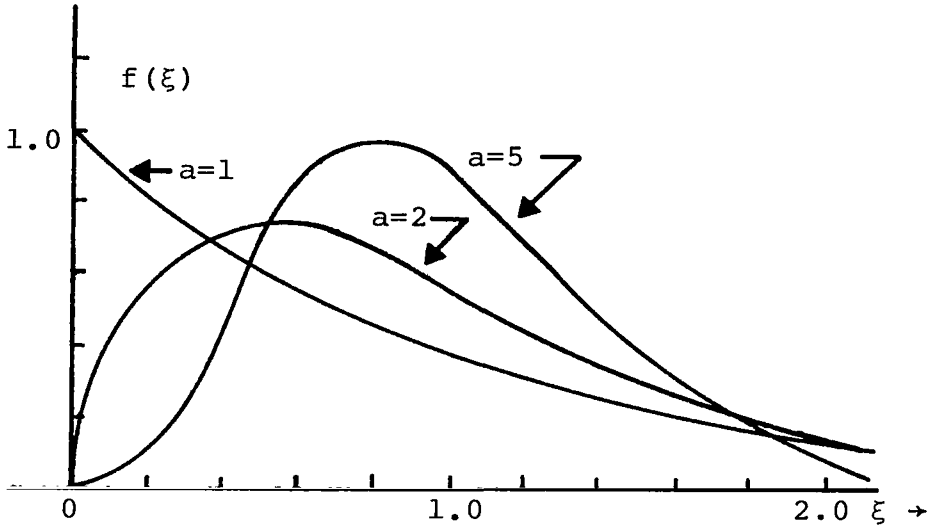

Fig. 3.2 Gamma Distribution with μ = 1

The mean of the distribution is

$$\bar{\xi} = \int_0^\infty \xi \ f(\xi) \ d\xi = \mu \tag{3.5.6}$$

and the variance of the distribution is

$$\sigma_\xi^2 = \int_0^\infty (\xi-\bar{\xi})^2 \ f(\xi) \ d\xi = \frac{\mu^2}{a} \tag{3.5.7}$$

In Chapter 2, the a priori state estimate was defined as the mean of the normal a priori state probability density function. Because of the symmetry of the normal density function, the mean is located at the point of maximum probability or likelihood. Now the a priori values of R and Q must be defined in terms of parameters of their respective distributions. The Gamma distribution is not symmetric about its mean so that the point of maximum probability occurs at a different point than the mean of the distribution. The same is true for the truncated normal distribution if the points of truncation are not chosen to be equidistant from the mean. Because the criterion of maximum likelihood is used to define the optimal estimates of the state and noise covariance parameters, it would be consistent to define the a priori estimates of these quantities as the points of maximum likelihood of their respective a priori probability density functions. If $\hat{\xi}_o^k$ denotes the a priori estimate of the kth component of ξ, then

$$\hat{\xi}_o^k = \mu^k \quad \text{for the truncated normal distribution}$$

$$\hat{\xi}_o^k = \frac{(a^k-1)}{a^k}\,\mu^k \quad \text{for the Gamma distribution}$$

Actually, if the parameters of the respective distributions are defined, there is no need to separately define the a priori estimates of ξ when solving the likelihood equations. The solution is a function of the parameters of the distribution, not $\hat{\xi}_o$. However, in subsequent sections when approximate solutions are discussed, it becomes convenient to introduce the a priori estimates as separate entities, although they will be related to the parameters of their distributions as shown above.

If a rectangular distribution of ξ is selected, then no point of maximum likelihood of this distribution exists. In this case, the a priori estimate of ξ is defined as the mean of the rectangular distribution. In fact, any point within the nonzero range of the distribution could be selected as the a priori estimate without affecting the solution, but for the sake of uniqueness, the above definition is made.

3.6 Computation of the Estimate

The likelihood equations for estimating the state and noise covariance parameters with and without the use of a priori information have been derived but in general the equations are so complicated that solutions cannot be obtained in closed form. In this section techniques for a numerical solution of the equations are discussed. For simplicity, only one of the several possible cases are treated, that of simultaneously estimating the state and noise covariance parameters when a priori information is used. The solution of this problem includes all of the features that are necessary for the solution of the others, so that only slight modification of the discussion below is necessary in the other cases.

The solutions of the augmented likelihood equations

$$\left(\frac{\partial L_n^A(\alpha, Z_n)}{\partial \alpha}\right)_{\hat{\alpha}_n} = \left(\frac{\partial \ln f(\alpha | Z_n)}{\partial \alpha}\right)_{\hat{\alpha}_n} = 0$$

are sought. A general method of solution would be to assume a trial solution and derive linear equations for small additive corrections. This process can be repeated until the corrections become negligible. If $\hat{\alpha}_o$ is the trial value of the estimate, then expanding $\frac{\partial L_n^A}{\partial \alpha}$ in a Taylor series and retaining only the first power of $\Delta\alpha_o = \hat{\alpha}_n - \hat{\alpha}_o$, leads to

$$\left(\frac{\partial L_n^A}{\partial \alpha}\right)_{\hat{\alpha}_n}^T = \left(\frac{\partial L_n^A}{\partial \alpha}\right)_{\hat{\alpha}_o}^T + \left(\frac{\partial^2 L_n^A}{\partial \alpha^2}\right)_{\hat{\alpha}_o} \Delta\alpha_o = 0 \qquad (3.6.1)$$

Assuming that $\left(\frac{\partial^2 L_n^A}{\partial \alpha^2}\right)_{\hat{\alpha}_o}$ is of full rank, the first linear correction to $\hat{\alpha}_o$ is

$$\Delta\alpha_o = - \left(\frac{\partial^2 L_n^A}{\partial \alpha^2}\right)_{\hat{\alpha}_o}^{-1} \left(\frac{\partial L_n^A}{\partial \alpha}\right)_{\hat{\alpha}_o}^T \qquad (3.6.2)$$

The next trial value is then $\hat{\alpha}_o + \Delta\hat{\alpha}_o$.

Clearly this method has several drawbacks. Computation of $\frac{\partial^2 L_n^A}{\partial \alpha^2}$ and its inverse is very complicated, and once a stable solution is found, another computation, the conditional information matrix, must be performed before any evaluation of the performance of the estimator can be undertaken. A mechanization introduced by Rao eliminates these drawbacks. It is quite similar to the above method but employs one approximation which greatly reduces the number of computations. For this iterative solution, the approximation is made

$$\left(\frac{\partial^2 L_n^A}{\partial \alpha^2}\right)_{\hat{\alpha}_o} = - J_n^A(\hat{\alpha}_o) \qquad (3.6.3)$$

where $J_n^A(\hat{\alpha}_o)$ is the augmented conditional information matrix defined by

$$J_n^A(\hat{\alpha}_o) = \varepsilon \left[\left(\frac{\partial L_n^A(\hat{\alpha}_o, Z_n)}{\partial \alpha}\right)^T \frac{\partial L_n^A(\hat{\alpha}_o, Z_n)}{\partial \alpha} \, | \hat{\alpha}_o \right]$$

Thus the additive correction $\Delta\alpha_o$ becomes

$$\Delta\alpha_o = [J_n^A(\hat{\alpha}_o)]^{-1} \left(\frac{\partial L_n^A}{\partial \alpha}\right)^T_{\hat{\alpha}_o} \qquad (3.6.4)$$

In large samples with a given $\alpha = \hat{\alpha}_o$, the difference between $\left(\frac{\partial^2 L_n^A}{\partial \xi^2}\right)_{\hat{\alpha}_o}$ and $- J_n^A(\hat{\alpha}_o)$ will be of order 1/n, so that the above approximation holds to first order of small quantities.

When a stable solution of $\hat{\alpha}_n$ is obtained, the asymptotic estimation error is zero mean normally distributed with conditional covariance $[J_n^A(\alpha)]^{-1}$ which is closely approximated by the computed $[J_n^A(\hat{\alpha}_n)]^{-1}$.

In this method the main difficulty is the computation and inversion of the information matrix at each stage of the iteration. In practice this is found to be unnecessary. The information matrix can be kept fixed after some stage and only the score recalculated. At the final stage when stable values are reached, the information matrix can be recomputed at the estimate value to obtain the covariance of the estimation error.

Whenever an iterative solution to a set of nonlinear equations is proposed, there is always a question of convergence. This question is reasonably well resolved in the case of the likelihood equations. Deutsch discusses this problem and references several other works on the subject. The results of his discussion are given below.

If $\hat{\alpha}_o$ is selected as the initial estimate of the solution of the likelihood equations, if $\hat{\alpha}_j$ is the j^{th} iteration value of the estimate, and if $\hat{\alpha}$ is the "true" maximum likelihood estimate, then the iteration process converges if $|\hat{\alpha}_j - \hat{\alpha}|$ decreases as j increases and tends to zero as $j \rightarrow \infty$. The iteration process is defined as follows: Let $g(\alpha)$ be a differentiable function which has no zero in the neighborhood of the root $\hat{\alpha}$ for the likelihood equation. The existence of $\hat{\alpha}$ is postulated. Define

$$h(\alpha) = \alpha - g(\alpha) \frac{\partial}{\partial \alpha} \ln L$$

where L is a likelihood function. The general iteration process is then

$$\hat{\alpha}_{j+1} = [h(\alpha)]_{\alpha=\hat{\alpha}_j}$$

$$= \hat{\alpha}_j - g(\hat{\alpha}_j)[\frac{\partial}{\partial \alpha} \ln L]_{\alpha=\hat{\alpha}_j}$$

If $$\varepsilon_j = |\hat{\alpha}_j - \hat{\alpha}|$$

is the estimation error at the j^{th} iteration, then $g(\alpha)$ must be chosen such that $\varepsilon_{j+1} < \varepsilon_j$ and $\varepsilon_j \rightarrow 0$ as $j \rightarrow \infty$. This condition assures the convergence of the iteration process to the value $\hat{\alpha}$. By using the asymptotic properties of the maximum likelihood estimator for large sample sizes, the two previously given iterative techniques for the computation of the estimate can be shown to be convergent.

3.7 Computation of the Information Matrix

By calculation of the information matrix, the asymptotic covariance of the maximum likelihood estimate can be obtained. Care must be taken to distinguish between $[J_n^A(\alpha_o)]^{-1}$ and $\overline{[J_n^A(\alpha_o)]^{-1}}$, the former being the conditional covariance of the estimate for a given value of α_o, the latter being the average conditonal covariance of the estimate, averaged over the ensemble of all possible true values of α_o.

$$J_n^A(\alpha_o) = \varepsilon\left[\left(\frac{\partial L_n^A}{\partial \alpha_o}\right)^T \frac{\partial L_n^A}{\partial \alpha_o} \,|\alpha_o\right] \qquad (3.7.1)$$

$$\overline{J_n^A(\alpha_o)^{-1}} = \int_0^\infty [J_n^A(\alpha_o)]^{-1} f(\alpha_o)\, d\alpha_o \qquad (3.7.2)$$

where $$L_n^A = L_n^A(\alpha_o, Z_n)$$

$[J_n^A(\alpha_o)]^{-1}$ is a highly nonlinear function of α_o, so the average conditional covariance cannot be explicitly calculated. Fortunately, $\overline{J_n^A(\alpha_o)^{-1}}$ is not needed in finding $\hat{\alpha}_n$, but is only used in evaluation of the estimator performance over the ensemble of all possible α_o. To find $\overline{J_n^A(\alpha_o)^{-1}}$ some numerical evaluation of (3.7.2) is necessary.

From (3.3.37) and (3.3.20),

$$\frac{\partial L_n^A}{\partial x_n} = - (x_n - \hat{x}_{n|n})^T P_{n|n}^{-1} \qquad (3.7.3)$$

From (3.3.37) and (3.3.24)

$$\frac{\partial L_n^A}{\partial \xi^j} = - \frac{1}{2}\Bigg(\mathrm{Tr}\left[(P_{n|n}^{-1} - P_{n|n}^{-1}\Delta x_n \Delta x_n^T P_{n|n}^{-1})\frac{\partial P_{n|n}}{\partial \xi^j} - 2\ P_{n|n}^{-1}\ \Delta x_n\ \frac{\partial \hat{x}_{n|n}^T}{\partial \xi}\right]$$

$$+ \sum_{i=1}^{n} \mathrm{Tr}\left[(B_i^{-1} - B_i^{-1}\Delta z_i \Delta z_i^T B_i^{-1})\frac{\partial B_i}{\partial \xi^j} - 2\ B_i^{-1}\Delta z_i \frac{\partial \hat{x}_{i|i-1}^T}{\partial \xi^j}\ H_i^T\right]\Bigg)$$

$$+ \frac{\partial \ln\ f(\xi^j)}{\partial \xi^j} \qquad (3.7.4)$$

where
$$\Delta x_n = x_n - \hat{x}_{n|n}$$

$$\Delta z_i = z_i - H_i \hat{x}_{i|i-1}$$

Then it follows that

$$\varepsilon\left[\left(\frac{\partial L_n^A}{\partial x_n}\right)^T \frac{\partial L_n^A}{\partial x_n}\ |\alpha_o\right] = [P_{n|n}(\xi_o)]^{-1} \qquad (3.7.5)$$

Using the same procedures as in obtaining (3.4.9), after algebraic manipulation, it can be shown that

$$\varepsilon\left[\frac{\partial L_n^A}{\partial \xi^j}\ \frac{\partial L_n^A}{\partial \xi^k}\ |\alpha_o\right] = \frac{1}{2}\Big[\mathrm{Tr}(P_{n|n}^{-1}\frac{\partial P_{n|n}}{\partial \xi^j}\ P_{n|n}^{-1}\frac{\partial P_{n|n}}{\partial \xi^k} + 2\ P_{n|n}^{-1} G_{n|n}^{jk})$$

$$+ \sum_{i=1}^{n} \mathrm{Tr}(B_i^{-1}\frac{\partial B_i}{\partial \xi^j}\ B_i^{-1}\frac{\partial B_i}{\partial \xi^k} + 2\ B_i^{-1} H_i G_{i|i-1}^{jk} H_i^T)\Big]$$

$$+ \frac{\partial \ln\ f(\xi^j)}{\partial \xi^j}\ \frac{\partial \ln\ f(\xi^k)}{\partial \xi^k} \qquad (3.7.6)$$

where $$G^{jk}_{i|i-1} = \varepsilon\left[\frac{\partial \hat{x}_{i|i-1}}{\partial \xi^j} \frac{\partial \hat{x}^T_{i|i-1}}{\partial \xi^k} \,|\xi_o\right]$$

$$G^{jk}_{n|n} = \varepsilon\left[\frac{\partial \hat{x}_{n|n}}{\partial \xi^j} \frac{\partial \hat{x}^T_{n|n}}{\partial \xi^k} \,|\xi_o\right]$$

It can also be shown that

$$\varepsilon\left[\left(\frac{\partial L^A_n}{\partial x_n}\right)^T \frac{\partial L^A_n}{\partial \xi^j} \,|\alpha_o\right] = 0 \qquad (3.7.7)$$

If the diagonal elements of R and Q (ξ) are mutually independent and are distributed with a truncated normal distribution, then

$$\frac{\partial \ln f(\xi^k)}{\partial \xi^k} = -\frac{(\xi^k - \mu^k)}{\sigma^2_k} \qquad (3.7.8)$$

where ξ^k represents the appropriate element of R or Q and μ^k and σ^2_k are the mean and variance of the corresponding untruncated normal distribution.

If the diagonal elements of R and Q are distributed with a Gamma distribution, then

$$\frac{\partial \ln f(\xi^k)}{\partial \xi^k} = \frac{a^k - 1}{\xi^k} - \frac{a^k}{\mu^k} \qquad (3.7.9)$$

where a^k and μ^k are parameters of the corresponding Gamma distribution.

All of the necessary quantities appearing in (3.7.6) can be computed using recursive relationships.

$$P_{n|n} = (I - A_n H_n) P_{n|n-1} (I - A_n H_n)^T + A_n\ R\ A_n^T \qquad \textbf{(3.7.10)}$$

$$P_{n|n-1} = \Phi(n,n-1) P_{n-1|n-1}\ \Phi^T(n,n-1) + \Gamma_n\ Q\ \Gamma_n^T \qquad (3.7.11)$$

$$\frac{\partial P^{st}_{n|n}}{\partial R^{jj}} = [(I - A_n H_n) \frac{\partial P_{n|n-1}}{\partial R^{jj}} (I - A_n H_n)^T]^{st} + A_n^{sj}\ A_n^{tj} \qquad (3.7.12)$$

$$\frac{\partial P_{n|n-1}}{\partial R^{jj}} = \Phi(n,n-1)\ \frac{\partial P_{n-1|n-1}}{\partial R^{jj}}\ \Phi^T(n,n-1) \qquad (3.7.13)$$

$$\frac{\partial P_{n|n}}{\partial Q^{jj}} = (I - A_n H_n)\ \frac{\partial P_{n|n-1}}{\partial Q^{jj}}\ (I - A_n H_n)^T \qquad (3.7.14)$$

$$\frac{\partial P^{st}_{n|n-1}}{\partial Q^{jj}} = [\Phi(n,n-1)\ \frac{\partial P_{n-1|n-1}}{\partial Q^{jj}}\ \Phi^T(n,n-1)]^{st} + \Gamma_n^{sj}\ \Gamma_n^{tj} \qquad (3.7.15)$$

$$G^{jk}_{n|n} = (I - A_n H_n)\ G^{jk}_{n|n-1}\ (I - A_n H_n)^T + \frac{\partial A_n}{\partial \xi^j}\ B_n\ \frac{\partial A_n^T}{\partial \xi^k} \qquad (3.7.16)$$

$$G^{jk}_{n|n-1} = \Phi(n,n-1)\ G^{jk}_{n-1|n-1}\ \Phi^T(n,n-1) \qquad (3.7.17)$$

where $$\frac{\partial A_n}{\partial \xi^j} = \left(\frac{\partial P_{n|n}}{\partial \xi^j} H_n^T - A_n \frac{\partial R}{\partial \xi^j} \right) R^{-1} \qquad (3.7.18)$$

The proper initial conditions for these recursive relationships are:

$$\frac{\partial P_{o|o}}{\partial R^{jj}} = \frac{\partial P_{o|o}}{\partial Q^{jj}} = G^{jk}_{o|o} = 0$$

$J_n^A(\alpha_o)$ can be partitioned into submatrices, corresponding to x_n and ξ.

$$J_n^A(\alpha_o) = \begin{bmatrix} P_{n|n}^{-1} & 0 \\ 0 & W_n^{-1} \end{bmatrix}$$

where $$W_n^{-1} = \varepsilon\left[\left(\frac{\partial L_n^A}{\partial \xi_o}\right)^T \frac{\partial L_n^A}{\partial \xi_o} \,|\, \xi_o\right]$$

Then $$[J_n^A(\alpha_o)]^{-1} = \begin{bmatrix} P_{n|n} & 0 \\ 0 & W_n \end{bmatrix}$$

and $$\overline{J_n^A(\alpha_o)^{-1}} = \begin{bmatrix} \overline{P_{n|n}} & 0 \\ 0 & \overline{W_n} \end{bmatrix}$$

where $$\overline{P_{n|n}} = \int P_{n|n}\ (\xi_o)\ f(\xi_o)\ d\xi_o$$

and $$\overline{W_n} = \int W_n\ (\xi_o)\ f(\xi_o)\ d\xi_o$$

Neither $\overline{P_{n|n}}$ nor $\overline{W_n}$ can be computed analytically. A first order approximation to $\overline{P_{n|n}}$ and $\overline{W_n}$ could be computed by expanding $P_{n|n}$ and W_n about $\overline{\xi}$.

$$P_{n|n}^{ij}(\xi_o) \simeq P_{n|n}^{ij}(\overline{\xi}) + \left(\frac{\partial P_{n\ n}^{ij}}{\partial \xi}\right)_{\overline{\xi}} \Delta\xi_o + \frac{1}{2}\,\Delta\xi_o^T \left(\frac{\partial^2 P_{n|n}^{ij}}{\partial \xi^2}\right)_{\overline{\xi}} \Delta\xi_o \qquad (3.7.19)$$

$$W_n^{ij}(\xi_o) \simeq W_n^{ij}(\overline{\xi}) + \left(\frac{\partial W_n^{ij}}{\partial \xi}\right)_{\overline{\xi}} \Delta\xi_o + \frac{1}{2}\,\Delta\xi_o^T \left(\frac{\partial^2 W_n^{ij}}{\partial \xi^2}\right)_{\overline{\xi}} \Delta\xi_o \qquad (3.7.20)$$

where $$\Delta\xi_o = \xi_o - \bar{\xi}$$

But $$E(\xi_o) = \bar{\xi}$$

and $$E(\Delta\xi_o \; \Delta\xi_o^T) = \mathrm{cov}(\xi_o)$$

where $\bar{\xi}$ is the mean of the a priori distribution of true parameter values ξ_o and $\mathrm{cov}(\xi_o)$ is the covariance of this distribution. Then

$$\overline{P^{ij}_{n|n}} \simeq P^{ij}_{n|n}\;(\bar{\xi}) + \frac{1}{2}\,\mathrm{Tr}\left[\left(\frac{\partial^2 P^{ij}_{n|n}}{\partial\xi^2}\right)_{\bar{\xi}} \mathrm{cov}(\xi_o)\right]$$

$$\overline{W^{ij}_n} \simeq W^{ij}_n\;(\bar{\xi}) + \frac{1}{2}\,\mathrm{Tr}\left[\left(\frac{\partial^2 W^{ij}_n}{\partial\xi^2}\right)_{\bar{\xi}} \mathrm{cov}(\xi_o)\right]$$

It is obvious that extensive computation is necessary to compute these quantities so that this technique is not particularly attractive.

An alternate method of evaluating $\overline{P_{n|n}}$ and $\overline{W_n}$ would be to select a sample of ξ chosen from the distribution $f(\xi)$ and then employ the approximations

$$\overline{P_{n|n}} \simeq \frac{1}{K} \sum_{j=1}^{K} P_{n|n}\;(\xi_j)$$

$$\overline{W_n} \simeq \frac{1}{K} \sum_{j=1}^{K} W_n\;(\xi_j)$$

Of course, the sample size K must be sufficiently large to ensure that this approximation is reasonably good.

The simplest approximation to make would be

$$\overline{P_{n|n}} \approx P_{n|n}(\overline{\xi})$$

$$\overline{W_n} \approx W_n(\overline{\xi})$$

This approximation may be adequate in applications where the range of ξ is limited, but caution should be employed in its use.

Chapter 4

SUBOPTIMAL SOLUTIONS OF THE ESTIMATION PROBLEM

4.1 Introduction

An exact or iterative solution of the likelihood equations of Chapter 3 requires extensive computation as the solution is generally found only after several passes over the measurement data. In many applications such computation is not feasible or a "real time" solution is needed. In such situations, approximate solutions are necessary, either to reduce the required computation and/or to obtain a real time solution of the parameter estimation problem. As would be expected, the quality of the estimator is degraded in such cases, but often the degradation is not serious. However, there are certain special cases when some of the approximate solutions are not unique or are so highly biased that their use is questionable.

This chapter deals with the derivation and evaluation of several suboptimal approximate solutions. Also included is a summary of possible parameter estimators suggested by other authors. The list of approximate solutions is not exhaustive but is meant to illustrate several techniques that are available to obtain an adequate solution of the problem.

4.2 Linearized Maximum Likelihood Solution

The iterative solution of the maximum likelihood equations of Chapter 3 was based upon successive relinearization of the maximum likelihood equations about trial values of the parameters obtained from the previous iteration, continuing the process until convergence. If the initial trial value of the parameter is "sufficiently close" to the true value, a single correction to the initial estimate based upon a linear approximation to the equations is often adequate for the solution. This single linearization is the basis of the linearized maximum likelihood solution.

As in Chapter 3, the solution of

$$\left(\frac{\partial L_n^A}{\partial \alpha}\right)_{\hat{\alpha}} = 0$$

is sought. If $\hat{\alpha}_o$ is the trial (a priori) value of the estimate, then from (3.6.4) the linearized maximum likelihood solution $\hat{\alpha}_\ell$ is found from the equation

$$\hat{\alpha}_\ell = \hat{\alpha}_o + [J_n^A(\hat{\alpha}_o)]^{-1} \left(\frac{\partial L_n^A}{\partial \alpha}\right)_{\hat{\alpha}_o}^T \tag{4.2.1}$$

The linearized solution $\hat{\alpha}_\ell$ can be found as long as $J_n^A(\hat{\alpha}_o)$ is of full rank. Both $J_n^A(\hat{\alpha}_o)$ and $\left(\frac{\partial L_n^A}{\partial \alpha}\right)_{\hat{\alpha}_o}$ can be evaluated in real time since they represent the conditional information matrix and the score evaluated at the <u>a priori</u> estimate of the parameter α. The conditional information matrix $J_n^A(\hat{\alpha}_o)$ is expressible as a linear combination of the conditional information matrix at the previous time, $J_{n-1}^A(\hat{\alpha}_o)$, and a term which represents the additional information about the parameters contained in the measurement at time n. Similarly, the score $\left(\frac{\partial L_n^A}{\partial \alpha}\right)_{\hat{\alpha}_o}$ is expressible as a linear combination of the score at the previous time, $\left(\frac{\partial L_{n-1}^A}{\partial \alpha}\right)_{\hat{\alpha}_o}$, and a term which is a function of the measurement at time n. Thus as the measurements are taken, the conditonal information matrix and the score can be computed as running sums, and the linearized solution (4.2.1) can be found in real time.

Because $\frac{\partial L_n^A}{\partial \alpha}$ is a highly nonlinear function of α, there is no simple way to determine when the above linearizing approximation is valid, or more importantly, when the linearized solution is "closer" to the true value than the a priori estimate. Several measures can be used to determine if the linearized solution is closer to the true solution. If the linearized solution is valid, the following inequality should be satisfied.

$$\left[\frac{\partial L_n^A}{\partial \alpha}\,[J_n^A(\alpha)]^{-1}\left(\frac{\partial L_n^A}{\partial \alpha}\right)^T\right]_{\hat{\alpha}_\ell} < \left[\frac{\partial L_n^A}{\partial \alpha}\,[J_n^A(\alpha)]^{-1}\left(\frac{\partial L_n^A}{\partial \alpha}\right)^T\right]_{\hat{\alpha}_o}$$

If this is not satisfied, another trial value of $\hat{\alpha}_o$ must be found and the procedure repeated. Evaluation of this measure requires a recomputation of the score and the conditional information matrix at the value $\alpha = \hat{\alpha}_\ell$, so in this sense the linearized solution is not real time. However, numerical results indicate that this linearized solution converges over a wide range of $\hat{\alpha}_o$ so that in many applications this check is not necessary.

The asymptotic conditional covariance of the linearized solution is approximately $[J_n^A(\hat{\alpha}_o)]^{-1}$. A better approximation can be obtained if computational capacity allows evaluation of $[J_n^A(\hat{\alpha}_\ell)]^{-1}$.

If it is known that there may be a significant error in the a priori estimate of α, then use of the linearized technique may be questionable. However, in this situation a combination of an iterative solution plus a linearized solution could be used. Sufficient measurements are taken to obtain a relatively good estimate of α by use of the iterative procedures of Chapter 3. Subsequently, the linearized solution is employed, using the results of the iterative procedure as the point about which to linearize.

A third procedure, sequential relinearization, could also be used. It is quite similar to the linearized solution except at regular intervals of time, which may encompass several measurement times, the best linearized estimate of α is used to compute subsequent values of the information matrix and the score. At each relinearization, the score must be corrected to account for having used a different value of α in its computation than the newly obtained value. Let $\hat{\alpha}_1$ be the estimate of α that was obtained at the previous relinearization and used from then until the present in the computation of the score, and let $\hat{\alpha}_2$ be the current linearized estimate. Expanding the score in a Taylor series about $\hat{\alpha}_1$,

$$\left(\frac{\partial L_n^A}{\partial \alpha}\right)_{\hat{\alpha}_2} \simeq \left(\frac{\partial L_n^A}{\partial \alpha}\right)_{\hat{\alpha}_1} + \left(\frac{\partial^2 L_n^A}{\partial \alpha \partial \alpha}\right)_{\hat{\alpha}_1} \qquad (\hat{\alpha}_2-\hat{\alpha}_1)+\ldots.$$

Using the approximation

$$\left(\frac{\partial^2 L_n^A}{\partial\alpha\partial\alpha}\right)_{\hat{\alpha}_1} \simeq \varepsilon\left[\frac{\partial^2 L_n^A}{\partial\alpha\partial\alpha}\,\middle|\,\hat{\alpha}_1\right] \simeq -\ J_n^A(\hat{\alpha}_1)$$

the corrected score is given by

$$\left(\frac{\partial L_n^A}{\partial\alpha}\right)_{\hat{\alpha}_2} = \left(\frac{\partial L_n^A}{\partial\alpha}\right)_{\hat{\alpha}_1} - J_n^A(\hat{\alpha}_1)\,(\hat{\alpha}_2-\hat{\alpha}_1)$$

As with the linearized solution, this procedure should be used only after a sufficiently accurate estimate of α is obtained, either from the a priori estimate or through use of the iterative procedure.

4.3 Near Maximum Likelihood Solution

By a suitable approximation to (3.3.38) a "near maximum likelihood" solution can be found which reduces the necessary computations considerably. In this solution, the state estimate is defined to be the maximum likelihood estimate which uses the near maximum likelihood estimates of R and Q (ξ) to compute the filter gains, and estimates of ξ are found from the solution of the "pseudo" likelihood equations:

$$\frac{\partial\Lambda_n}{\partial\xi^j} = -\ \frac{1}{2}\sum_{i=1}^{n} \mathrm{Tr}\left[(B_i^{-1}-B_i^{-1}\Delta z_i\Delta z_i^T B_i^{-1})\frac{\partial B_i}{\partial\xi^j}\right] + \frac{\partial \ln\ f(\xi^j)}{\partial\xi^j} = 0 \qquad (4.3.1)$$

where Λ_n is the "pseudo" likelihood function defined by (4.3.1). This equation is obtained from (3.3.38) by retaining only the most significant terms. The savings in computation arise from not having to compute $\frac{\partial\hat{x}_{i|i-1}}{\partial\xi}$ appearing in the likelihood function and $G^{kj}_{i|i-1}$ appearing in the expression for the conditional information matrix (3.7.6). $\frac{\partial \hat{x}_{i|i-1}}{\partial \xi}$ is an array with $\beta x(\gamma+\eta)$ elements and $G^{kj}_{i|i-1}$ is an array with $\beta^2 x(\gamma+\eta)^2$ elements. If all of the symmetry properties of $G^{kj}_{i|i-1}$ are utilized, the number of independent elements is $\frac{\beta(\beta+1)}{2} \times \frac{(\gamma+\eta)(\gamma+\eta+1)}{2}$. If the state, driving noise, and measurement are of moderate dimension, the number of computations involved in calculating these quantities can be considerable, so that not having to perform these calculations can result in a significant saving in computer time.

If convergence of (4.3.1) to a unique solution is obtained, the asymptotic distribution of $\hat{x}_{n|n}$ and $\hat{\xi}_n$ are approximately normal with conditional covariances

$$\varepsilon[(x_n-\hat{x}_{n|n})(x_n-\hat{x}_{n|n})^T] = P_{n|n}(\xi)$$

$$\varepsilon[(\xi-\hat{\xi}_n)(\xi-\hat{\xi}_n)^T] = \left[\varepsilon[\left(\frac{\partial \Lambda_n}{\partial \xi}\right)^T \frac{\partial \Lambda_n}{\partial \xi} \mid \xi]\right]^{-1} \triangleq J_n^{-1}(\xi)$$

The conditional information matrix $J_n(\xi)$ is not the same as the information matrix of Chapter 3 because of the omitted terms in the likelihood function. Here

$$[J_n(\xi)]^{kj} = \sum_{i=1}^{n} \mathrm{Tr}\left[(B_i^{-1}\frac{\partial B_i}{\partial \xi^k}B_i^{-1}\frac{\partial B_i}{\partial \xi^j})\right] + \frac{\partial \ln f(\xi^k)}{\partial \xi^k}\frac{\partial \ln f(\xi^j)}{\partial \xi^j} \qquad (4.3.2)$$

A comparison of (4.3.2) with (3.7.6) will show that the above information matrix is smaller* than the information matrix of Chapter 3. Thus, as would be expected, the covariance of the parameter estimates will be larger when the pseudo likelihood equations are used than when the full likelihood equations are solved.

Numerical results indicate that the iterative solution of the pseudo likelihood equations when the information matrix (4.3.2) is used as an approximation to the negative gradient of the likelihood equations may present difficulties. This is because in some circumstances J_n given above may be nearly singular and using its inverse in the solution may result in an unstable iterative procedure. However, these same numerical results show that the pseudo likelihood equations do have a unique solution, but they must be found using other techniques in the iteration algorithm, say a fixed step size sweep looking for zeros of the pseudo likelihood equations.

4.4 Explicit Suboptimal Solutions

In this section, explicit "real time" solutions for the estimates of R and Q are sought. As will be shown, such estimates are approximations to the maximum likelihood solutions and on any given trial may be highly biased. However, if the

* a positive definite matrix A is said to be smaller than another positive definite matrix B if the matrix (A-B) is negative semi-definite.

a priori estimates of R and Q are sufficiently close to the true values, such estimators will provide reasonable estimates with considerably less computation than the estimators previously discussed. Even if the estimates are biased, they provide useful information. If the estimates differ consistently and significantly from the assumed a priori values, then there is good reason to doubt the accuracy of the a priori values, even though the biased estimates do not necessarily represent better estimates of R and Q. In other words, the explicit estimates will indicate if there is a significant error in the a priori values of R and Q even if they do not tell how to correct this error. In this sense their use is related to testing a hypothesis on the values of R and Q as discussed in Chapter 5.

These approximate estimators are obtained as approximate solutions of the pseudo likelihood equations (4.3.1).

$$\frac{\partial \Lambda_n}{\partial \xi^j} = -\frac{1}{2} \sum_{i=1}^{n} \mathrm{Tr} \left[(B_i^{-1} - B_i^{-1} \Delta z_i \Delta z_i^T B_i^{-1}) \frac{\partial B_i}{\partial \xi^j} \right] + \frac{\partial \ln f(\xi^j)}{\partial \xi^j} = 0$$

The last term allows introduction of a distribution function of R and Q so that a priori estimates can be weighted with the estimates derived from the measurements alone. For this approximate solution it is convenient to form estimates of R and Q which are independent of this distribution function, and after such estimates are obtained, then the a priori estimates and their associated covariances are considered in obtaining a combined estimate for R and Q. Thus, initially, the solutions of the following equation are sought.

$$\sum_{i=1}^{n} \operatorname{Tr}(\Delta B_i^{-1} \frac{\partial B_i}{\partial \xi^j}) = 0 \qquad (4.4.1)$$

where $$\Delta B_i^{-1} = B_i^{-1} - B_i^{-1} \Delta z_i \Delta z_i^T B_i^{-1}$$

Using the results of Appendix A, (4.4.1) becomes

$$\sum_{i=1}^{n} [(\Delta B_i^{-1})^{jj} + \operatorname{Tr}(\Delta B_i^{-1} H_i \frac{\partial P_{i|i-1}}{\partial R^{jj}} H_i^T)] = 0 \qquad (4.4.2)$$

$$\sum_{i=1}^{n} \operatorname{Tr}(\Delta B_i^{-1} H_i \frac{\partial P_{i|i-1}}{\partial Q^{jj}} H_i^T) = 0 \qquad (4.4.3)$$

As the equations stand, no explicit solution for estimates of R and Q is possible, so further approximations must be made. When these approximations are made, there is a real question of existence of independent solutions of the resulting equations for the unknown elements of R and Q. Even if there are sufficient independent equations, there is no general way to obtain a closed form solution of the nonlinear relationships. If R or Q is to be estimated separately, there is no difficulty in obtaining a reasonable solution to the problem. Unfortunately the question of simultaneous estimation of these quantities from the above equations is not well resolved. The solutions given below represent separate estimation of R with Q known and estimation of Q with R known. The two solutions can be used, with caution, to simultaneously estimate R and Q, realizing beforehand that the resulting estimates are not independent. This dependency can result in biased estimates which fail to distinguish between errors in R and Q. However, as mentioned previously, some useful information can be derived from such biased estimates.

It can be shown that for many applications

$$H_i \frac{\partial P_{i|i-1}}{\partial R^{jj}} H_i^T << I$$

so that (4.4.2) becomes

$$\sum_{i=1}^{n} (B_i^{-1} - B_i^{-1} \Delta z_i \Delta z_i^T B_i^{-1})^{jj} = 0 \qquad (4.4.4)$$

But $$B_i^{-1} = R^{-1} - R^{-1} H_i P_{i|i} H_i^T R^{-1}$$

Since $$\hat{x}_{i|i} = \hat{x}_{i|i-1} + P_{i|i} H_i^T R^{-1} (z_i - H_i \hat{x}_{i|i-1})$$

it can be seen that

$$z_i - H_i \hat{x}_{i|i} = z_i - H_i \hat{x}_{i|i-1} - H_i P_{i|i} H_i^T R^{-1} (z_i - H_i \hat{x}_{i|i-1})$$

$$= (I - H_i P_{i|i} H_i^T R^{-1})(z_i - H_i \hat{x}_{i|i-1})$$

$$= R(R^{-1} - R^{-1} H_i P_{i|i} H_i^T R^{-1}) \Delta z_i$$

$$= R B_i^{-1} \Delta z_i$$

Defining $$\Delta z_i' = z_i - H_i \hat{x}_{i|i}$$

then $$B_i^{-1} \Delta z_i = R^{-1} \Delta z_i'$$

and (4.4.4) becomes

$$\sum_{i=1}^{n} [R^{-1}(R - H_i P_{i|i} H_i^T - \Delta z_i' \Delta z_i'^T) R^{-1}]^{jj} = 0$$

Or $$R^{jj} - \frac{1}{n} \sum_{i=1}^{n} (\Delta z_i' \Delta z_i'^T + H_i P_{i|i} H_i^T)^{jj} = 0 \qquad (4.4.5)$$

It is still not possible to solve (4.4.5) for R as $P_{i|i}$ and $\Delta z_i'$ are highly nonlinear functions of both R and Q. However, if either the a priori values of R and Q or some estimates of these quantities are used to compute $\Delta z_i'$ and $P_{i|i}$, then the estimate of R can be defined as

$$\hat{R}_n^{jj} = \frac{1}{n} \sum_{i=1}^{n} (\Delta z_i'^{*} \Delta z_i'^{*T} + H_i P_{i|i}^{*} H_i^T)^{jj} \qquad (4.4.6)$$

where $\Delta z_i'^{*}$ and $P_{i|i}^{*}$ are computed as functions of either the a priori estimates of R and Q or some previously obtained estimates.

A recursive relationship for $\hat{R}_n^{jj}$ can be obtained if $\Delta z_n'^{*}$ and $P_{n|n}^{*}$ are not functions of $\hat{R}_n$ or $\hat{Q}_n$.

$$\hat{R}_n^{jj} = \frac{n-1}{n} \hat{R}_{n-1}^{jj} + \frac{1}{n} (\Delta z_n'^{*} \Delta z_n'^{*T} + H_n P_{n|n}^{*} H_n^T)^{jj} \qquad (4.4.7)$$

Equation (4.4.7) is not the only approximate solution that could be reasonably obtained from (4.4.4). Rewriting (4.4.4)

$$\sum_{i=1}^{n} [B_i^{-1}(B_i - \Delta z_i \Delta z_i^T) B_i^{-1}]^{jj} = 0$$

Or $$\sum_{i=1}^{n} [B_i^{-1}(R + H_i P_{i|i} H_i^T - \Delta z_i \ \Delta z_i^T) B_i^{-1}]^{jj} = 0 \qquad (4.4.8)$$

If the estimation process has reached a steady state, that is, $B_i \approx$ constant for all i, then an estimate of R can be defined by

$$\hat{R}_n^{jj} = \frac{1}{n} \sum_{i=1}^{n} (\Delta z_i^* \ \Delta z_i^{*T} - H_i P_{i|i-1}^* H_i^T)^{jj} \qquad (4.4.9)$$

where Δz_i^* and $P_{i|i-1}^*$ are equal to Δz_i and $P_{i|i-1}$ computed as a function of a priori values of R and Q or some past estimates of these quantities. The form of (4.4.9) is not as desirable as (4.4.6) because $\hat{R}_n$ is not necessarily positive definite. If some of the squared residuals are small compared with $H_i P_{i|i-1}^* H_i^T$, then some of the terms in the above sum can be negative. If this occurs often, then the resulting estimate of R may have negative diagonal elements. However, the estimator has the advantage of not being a function of the value of R that is used to update Δz_n^* and $P_{n|n-1}^*$ at time n. This can reduce possible bias problems in the feedback estimator discussed later. The estimator of the form (4.4.6) is the one studied further.

Obtaining an explicit estimate of Q is not as straightforward as obtaining the estimate of R. There are many approximate solutions to (4.4.3) for $\hat{Q}_n$ depending upon the nature of the approximations made. The solution given below is but one of several possible solutions, but it is felt that it has the advantage of simplicity and wide applicability.

By manipulation of (4.4.3) it can be shown that

$$\sum_{i=1}^{n} \mathrm{Tr}(\Delta B_i^{-1} H_i \frac{\partial P_{i|i-1}}{\partial Q^{jj}} H_i^T)$$

$$= \sum_{i=1}^{n} \mathrm{Tr}\left[P_{i|i-1}^{-1}(P_{i|i-1} - P_{i|i} - \Delta x_i \Delta x_i^T) P_{i|i-1}^{-1} \frac{\partial P_{i|i-1}}{\partial Q^{jj}}\right]$$

where $\Delta x_i \triangleq \hat{x}_{i|i} - \hat{x}_{i|i-1} = P_{i|i-1} H_i^T B_i^{-1} \Delta z_i$

Define $U_i = \Phi(i,i-1) P_{i-1|i-1} \Phi^T(i,i-1)$

Then (4.4.3) becomes

$$\sum_{i=1}^{n} \mathrm{Tr}\left[P_{i|i-1}^{-1}(\Gamma_i Q \Gamma_i^T - \Delta x_i \Delta x_i^T - P_{i|i} + U_i) P_{i|i-1}^{-1} \frac{\partial P_{i|i-1}}{\partial Q^{jj}}\right] = 0$$

But $P_{i|i-1} = U_i + \Gamma_i Q \Gamma_i^T$

So $\frac{\partial P_{i|i-1}}{\partial Q^{jj}} = \frac{\partial U_i}{\partial Q^{jj}} + \Gamma_i \frac{\partial Q}{\partial Q^{jj}} \Gamma_i^T$

In many applications, $\Gamma_i \frac{\partial Q}{\partial Q^{jj}} \Gamma_i^T >> \frac{\partial U_i}{\partial Q^{jj}}$, so (4.4.3) becomes

$$\sum_{i=1}^{n} \left[\Gamma_i^T P_{i|i-1}^{-1}(\Gamma_i Q \Gamma_i^T - \Delta x_i \Delta x_i^T - P_{i|i} + U_i) P_{i|i-1}^{-1} \Gamma_i\right]^{jj} = 0 \qquad (4.4.10)$$

Equation (4.4.10) cannot be solved explicitly for Q, so additional approximations are necessary. If it is assumed that Γ_i and $P_{i|i-1}$ are approximately constant for all i, then (4.4.10) becomes

$$\left[\Gamma_n^T P_{n|n-1}^{-1} \sum_{i=1}^{n} (\Gamma_i Q \Gamma_i^T - \Delta x_i \Delta x_i^T - P_{i|i} + U_i)\ P_{n|n-1}^{-1} \Gamma_n\right]^{jj} = 0$$

The equation above is satisfied if

$$\sum_{i=1}^{n} (\Gamma_i Q \Gamma_i^T - \Delta x_i \Delta x_i^T - P_{i|i} + U_i) = 0 \qquad (4.4.11)$$

Or $$Q^{jj} - \frac{1}{n} \sum_{i=1}^{n} \left[\Gamma_i^{-1} (\Delta x_i \Delta x_i^T + P_{i|i} - U_i) \Gamma_i^{T-1}\right]^{jj} = 0$$

If Γ_i^{-1} does not exist, the generalized inverse of Γ_i is to be used. (See Appendix A for discussion of generalized inverse.) In general the dimension of the driving noise vector is less than or equal to the dimension of the state, in which case $(\Gamma_i^T \Gamma_i)^{-1}$ exists and the generalized inverse of Γ_i is

$$\Gamma_i^{\#} = (\Gamma_i^T \Gamma_i)^{-1} \Gamma_i^T$$

The estimate of Q is defined as

$$\hat{Q}_n^{jj} \triangleq \frac{1}{n} \sum_{i=1}^{n} \left[\Gamma_i^{-1} (\Delta x_i^* \Delta x_i^{*T} + P_{i|i}^* - U_i^*) \Gamma_i^{T-1}\right]^{jj} \qquad (4.4.12)$$

where Δx_i^*, $P_{i|i}^*$, and U_i^* are computed as functions of the a priori estimates of R and Q or some past estimates. If Δx_n^*, $P_{n|n}^*$, and U_n^* are not functions of $\hat{R}_n$ or $\hat{Q}_n$, a recursive relationship can be obtained.

$$\hat{Q}_n^{jj} = \frac{n-1}{n} \hat{Q}_{n-1}^{jj} + \frac{1}{n}\left[\Gamma_n^{-1} (\Delta x_n^* \Delta x_n^* + P_{n|n}^* - U_n^*) \Gamma_n^{T-1}\right]^{jj} \qquad (4.4.13)$$

Two classes of estimators of the form (4.4.7) and (4.4.13) exist depending upon what use is made of past estimates of R and Q.

1) no feedback estimators
2) feedback estimators

In the no feedback case, a priori values of R and Q are used to compute the quantities denoted by a * in the estimator equations. In the feedback case these quantities are computed as functions of past estimates of R and Q. At each stage the best available estimates of R and Q are used to update the starred quantities. If feedback is employed and the variance estimation process converges to the true values of R and Q, then the state estimate $\hat{x}^*_{n|n}$ will converge in most applications to the optimal state estimate that would be obtained if the true values of R and Q were known a priori. However, using this estimation scheme, convergence is not guaranteed. In fact, numerical results indicate that if the a priori values of R and Q are significantly in error, the process will converge but to biased and incorrect estimates of the variance parameters. Techniques for evaluating the performance of the feedback and no feedback estimators are given next.

The two measures which seem appropriate for evaluating the performance of the explicit suboptimal estimators are the mean and mean square error of the estimates of R and Q. In the preceding section, estimators for the diagonal elements of R and Q were developed, resulting in $(\gamma+\eta)$ estimator equations. The mean square error matrix of such estimates is a $(\gamma+\eta) \times (\gamma+\eta)$ matrix, which includes the mean of all quadratic functions of the errors in each component of the diagonal elements of R and Q. Such a matrix is most difficult to compute, so for the purposes of this development, only the diagonal elements of such a matrix will be considered.

As mentioned in Chapter 2, a distinction must be made between conditional and unconditional expectation operators. The same notation as in that chapter will be used to make this distinction.

First, the performance of the no feedback estimator will be discussed. From (4.4.7)

$$\hat{R}_n^{jj} = \frac{n-1}{n} \hat{R}_{n-1}^{jj} + \frac{1}{n}(\Delta z_n'^{*} \; \Delta z_n'^{*T} + H_n P_{n|n}^{*} H_n^T)^{jj}$$

The conditional expected value of $\hat{R}_n^{jj}$ is

$$\varepsilon(\hat{R}_n^{jj}) = \frac{n-1}{n} \varepsilon(\hat{R}_{n-1}^{jj}) + \frac{1}{n}\left[\varepsilon(\Delta z_n'^{*} \; \Delta z_n'^{*T}) + H_n \varepsilon(P_{n|n}^{*}) H_n^T\right]^{jj}$$

This conditional expected value is conditioned upon the fact that the a priori estimates $\hat{R}_o$ and $\hat{Q}_o$ are used to compute the filter gains while the true values of these covariances are R and Q. Averaging is performed over the ensemble of all driving and measurement noises as well as all possible initial state conditions.

$$\Delta z_n'^{*} = z_n - H_n\hat{x}^*_{n|n} = v_n - H_n\tilde{x}^*_{n|n}$$

So $$\varepsilon(\Delta z_n'^{*}\ \Delta z_n'^{*T}) = \varepsilon(v_n v_n^T) + H_n\ \varepsilon(\tilde{x}^*_{n|n}\ \tilde{x}^{*T}_{n|n})H_n^T$$

$$- H_n\ \varepsilon(\tilde{x}^*_{n|n}v_n^T) - \varepsilon(v_n\tilde{x}^{*T}_{n|n})H_n^T$$

$$= R + H_nP_{n|n}H_n^T - H_nA_n^*R - RA_n^{*T}H_n^T$$

where $$A_n^* = P^*_{n|n-1}H_n^T(\hat{R}_o + H_nP^*_{n|n-1}H_n^T)^{-1}$$

$$P_{n|n} = \varepsilon(\tilde{x}^*_{n|n}\tilde{x}^{*T}_{n|n})$$ (not $P^*_{n|n}$ unless $\hat{R}_o = R$ and $\hat{Q}_o = Q$)

In the no feedback case, $P^*_{n|n}$ is not a random variable under the expectation operator, so $\varepsilon(P^*_{n|n}) = P^*_{n|n}$ and

$$\varepsilon(\hat{R}_n^{jj}) = \frac{n-1}{n}\ \varepsilon(\hat{R}_{n-1}^{jj}) + \frac{1}{n}(R^{jj} + \Delta F_n^{jj}) \qquad (4.4.14)$$

where $$\Delta F_n = H_n(P_{n|n} + P^*_{n|n})H_n^T - H_nA_n^*R - R\ A_n^{*T}H_n^T$$

This can be expressed as

$$\varepsilon(\hat{R}_n^{jj}) = R^{jj} + F_n^{jj}$$

where $$F_n \triangleq \frac{1}{n}\sum_{i=1}^{n} \Delta F_i$$

If $\hat{R}_o = R$ and $\hat{Q}_o = Q$, then $P^*_{n|n} = P_{n|n}$, $A_n^*R = P_{n|n}H_n^T$, and from the definition of ΔF_n it can be seen that $\Delta F_i = 0$, for all i. Then

$$\varepsilon(\hat{R}_n^{jj}) = R^{jj}$$

If $\hat{R}_o \neq R$ or $\hat{Q}_o \neq Q$, then $\Delta F_i \neq 0$ and $\hat{R}_n$ is biased, the bias equalling F_n.

The unconditional expected value of $\hat{R}_n$ follows from the above.

$$E(\hat{R}_n^{jj}) = E(R^{jj}) + E(F_n^{jj})$$

Here averaging is done over the ensembles mentioned above and also over the ensemble of all possible R and Q.

By definition $E(R) = \overline{R}$, where $\overline{R}$ is the mean of the distribution of all possible R values, and

$$E(F_n) = \frac{1}{n} \sum_{i=1}^{n} E(\Delta F_i)$$

But $E(\Delta F_i) = H_i E(P_{i|i}) H_i^T + H_i P^*_{i|i} H_i^T - H_i A_i^* E(R) - E(R) A_i^{*T} H_i^T$

$E(P_{i|i})$ can be computed recursively using (2.3.43) and (2.3.44)

$$E(P_{i|i}) = (I - A_i^* H_i)\ E(P_{i|i-1})(I - A_i^* H_i)^T + A_i^* \overline{R} A_i^{*T}$$

$$E(P_{i|i-1}) = \Phi(i,i-1)\ E(P_{i-1|i-1}) \cdot \Phi^T(i,i-1) + \Gamma_i \overline{Q} \Gamma_i^T$$

where $\overline{Q}$ is the mean of the distribution of all possible Q values.

Define $$\bar{P}_{i|i} = E(P_{i|i})$$

Then $$E(\Delta F_i) = H_i\bar{P}_{i|i}H_i^T + H_iP^*_{i|i}H_i^T - H_iA_i^*\bar{R} - \bar{R}A_i^{*T}H_i^T$$

and $$E(\hat{R}_n^{jj}) = \bar{R}^{jj} + E(F_n^{jj})$$

If the a priori values of R and Q are assumed to be equal to the means of their respective distributions, then

$$\hat{R}_o = \bar{R} \quad \text{and} \quad \hat{Q}_o = \bar{Q}$$

and it can be shown that

$$\bar{P}_{i|i} = P^*_{i|i}$$

$$E(\Delta F_i) = 0$$

Then $$E(\hat{R}_n^{jj}) = \bar{R}^{jj}$$

Thus $\hat{R}_n$ is an unbiased estimator of R across the ensemble of all possible R and Q. However, if $\hat{R}_o \neq \bar{R}$ or $\hat{Q}_o \neq \bar{Q}$, then $E(\Delta F_i) \neq 0$ and $\hat{R}_n$ is biased, the bias equalling $E(F_n)$.

The measures of error of the estimator are chosen to be the expected squared deviation of the R estimate from the true value, or $\varepsilon[(\hat{R}_n^{jj} - R^{jj})^2]$ and $E[(\hat{R}_n^{jj} - R^{jj})^2]$.

$$\varepsilon[(\hat{R}_n^{jj} - R^{jj})^2] = \varepsilon[(\hat{R}_n^{jj} - \varepsilon(\hat{R}_n^{jj}))^2] + [\varepsilon(\hat{R}_n^{jj} - R^{jj})]^2$$

$\varepsilon[(\hat{R}_n^{jj} - \varepsilon(\hat{R}_n^{jj}))^2]$ can be computed recursively by noting that

$$\varepsilon[(\hat{R}_n^{jj} - \varepsilon(\hat{R}_n^{jj}))^2] = \varepsilon[(\hat{R}_n^{jj})^2] - [\varepsilon(\hat{R}_n^{jj})]^2$$

The diagonal elements of (4.4.7) are squared and the conditional expected value then evaluated. Use is made of the fact that the residuals $\Delta z_i^{*}{}'$ are zero mean normal variables in the no feedback case, and the approximation

$$\varepsilon[\hat{R}_{n-1}^{jj}(\Delta z_n^{*}{}',^j)^2] \simeq \varepsilon(\hat{R}_{n-1}^{jj})\ \varepsilon[(\Delta z_n^{*}{}',^j)^2]$$

is used. It can be shown that as the filter approaches optimality $(\hat{R}_o \to R,\ \hat{Q}_o \to Q)$, the above approximation is identically satisfied. Using the above approximation and after extensive algebraic manipulation,

$$\varepsilon[(\hat{R}_n^{jj} - \varepsilon(\hat{R}_n^{jj}))^2] = G_n^{jj} \tag{4.4.15}$$

where $$G_n^{jj} = \left(\frac{n-1}{n}\right)^2 G_{n-1}^{jj} + \frac{2}{n^2}\,(R^{jj} + \Delta F_n^{jj} - (H_n P_{n|n}^{*} H_n^T)^{jj})^2$$

So $$\varepsilon[(\hat{R}_n^{jj} - R^{jj})^2] = G_n^{jj} + (F_n^{jj})^2 \tag{4.4.16}$$

In this expression, $(F_n^{jj})^2$ is due to the bias of the estimator and G_n^{jj} is due to possible deviations from this bias. The unconditional mean square estimation error follows from the above.

$$E[(\hat{R}_n^{jj} - R^{jj})^2] = E(G_n^{jj}) + E[(F_n^{jj})^2]$$

Evaluation of $E(G_n^{jj})$ and $E[(F_n^{jj})^2]$ is extremely complicated so the details of their evaluation are given in Appendix B. Only the results of that evaluation are given here.

Under the assumptions that $R_o = \hat{\bar{R}}$ and $Q_o = \hat{\bar{Q}}$, then

$$E(G_n^{jj}) = \left(\frac{n-1}{n}\right)^2 E(G_{n-1}^{jj}) + \frac{2}{n^2} [(\bar{R}^{jj} - a_{n|n}^{*j})^2 + (C_{n|1} \textstyle\sum_R C_{n|1}^T \tag{4.4.17}$$

$$+ L_{n|1} \textstyle\sum_Q L_{n|1}^T)^{jj}]$$

where $\sum_R$ is a diagonal $\gamma x \gamma$ matrix whose diagonal elements are $E[(R^{jj} - \bar{R}^{jj})^2]$

$\sum_Q$ is a diagonal $\eta x \eta$ matrix whose diagonal elements are $E[(Q^{jj} - \bar{Q}^{jj})^2]$

$C_{n|1}$ is a $\gamma x \gamma$ matrix defined by

$$(C_{n|1})^{j\ell} = \sum_{k=1}^{n} ((H_n \lambda_{n|k} A_k^*)^{j\ell})^2 + (I - 2H_n A_n^*)^{jj} \delta_{j\ell}$$

$\lambda_{n|k}$ is a $\beta x \beta$ matrix defined by

$$\lambda_{n|k} = \prod_{i=k+1}^{n} (I - A_i^* H_i)\ \Phi(i,i-1), \qquad \lambda_{k|k} \triangleq I$$

$L_{n|1}$ is a $\gamma x \eta$ matrix defined by

$$(L_{n|1})^{j\ell} = \sum_{k=1}^{n} ((H_n \lambda_{n|k} D_k \Gamma_k)^{j\ell})^2$$

$$D_k = I - A_k^* H_k$$

$$a^{*j}_{n|n} = (H_n P^*_{n|n} H^T_n)^{jj}$$

Also $E[(F^{jj}_n)^2] = \left(\frac{n-1}{n}\right)^2 E[(F^{jj}_{n-1})^2] + \frac{1}{n^2}(2\ n\ C'_{n|1} \sum_R \bar{C}'^T_{n|1} - C'_{n|1} \sum_R C'^T_{n|1})^{jj}$

$$+ (2\ n\ L_{n|1} \sum_Q \bar{L}^T_{n|1} - L_{n|1} \sum_Q L^T_{n|1})^{jj}$$

where $$C'_{n|1} \triangleq C_{n|1} - I$$

$$\bar{C}'_{n|1} \triangleq \frac{1}{n} \sum_{k=1}^{n} C'_{k|1}$$

$$\bar{L}_{n|1} \triangleq \frac{1}{n} \sum_{k=1}^{n} L_{k|1}$$

Evaluation of the no feedback Q estimator is similar to that of the R estimator. From (4.4.13)

$$\hat{Q}^{jj}_n = \frac{n-1}{n} \hat{Q}^{jj}_{n-1} + \frac{1}{n} (\Gamma^{-1}_n (\Delta x^*_n \Delta x^{*T}_n + P^*_{n|n} - U^*_n) \Gamma^{T-1}_n)^{jj}$$

where $$\Delta x^*_n = \hat{x}^*_{n|n} - \hat{x}^*_{n|n-1}$$

and $$U^*_n = \Phi(n,n-1) P^*_{n-1|n-1}\ \Phi^T(n,n-1)$$

The conditional expected value of $\hat{Q}^{jj}_n$ is given by

$$\varepsilon(\hat{Q}^{jj}_n) = \frac{n-1}{n}\ \varepsilon(\hat{Q}^{jj}_{n-1}) + \frac{1}{n}\ (Q^{jj} + \Delta M^{jj}_n)$$

where $\Delta M_n = \Gamma^{-1}_n (P_{n|n} + P^*_{n|n} - P^*_{n|n} P^{*-1}_{n|n-1} P_{n|n-1} - P_{n|n-1} P^{*-1}_{n|n-1} P^*_{n|n}$

$$+ U_n - U^*_n) \Gamma^{T-1}_n$$

$$U_n = \Phi(n,n-1)P_{n-1|n-1}\,\Phi^T(n,n-1)$$

Define $$M_n = \frac{1}{n}\sum_{k=1}^{n} \Delta M_k$$

Then $$\varepsilon(\hat{Q}_n^{jj}) = Q^{jj} + M_n^{jj} \qquad (4.4.18)$$

If $\hat{R}_o = R$ and $\hat{Q}_o = Q$, it can be seen that $\Delta M_k = 0$ for all k. In this case $\hat{Q}_n$ is an unbiased estimator for Q. If $\hat{R}_o \neq R$ or $\hat{Q}_o \neq Q$, then $\hat{Q}_n$ is biased, the bias equalling M_n.

The unconditional expected value of $\hat{Q}_n$ follows from the above.

$$E(\hat{Q}_n^{jj}) = E(Q^{jj}) + E(M_n^{jj}) \qquad (4.4.19)$$

But $$E(Q^{jj}) = \overline{Q}^{jj}$$

and $$E(M_n^{jj}) = \frac{1}{n}\sum_{k=1}^{n} E(\Delta M_k^{jj})$$

If $\hat{R}_o = \overline{R}$ and $\hat{Q}_o = \overline{Q}$, the $E(\Delta M_k) = 0$ for all k and

$$E(Q_n^{jj}) = \overline{Q}^{jj}$$

If $\hat{R}_o \neq \overline{R}$ or $\hat{Q}_o \neq \overline{Q}$, $\hat{Q}_n$ is biased across the ensemble of all possible R and Q, the bias equalling $E(M_n)$.

As with the R estimator, the measures of estimator error are $\varepsilon[(\hat{Q}_n^{jj} - Q^{jj})^2]$ and $E[(\hat{Q}_n^{jj} - Q^{jj})^2]$.

$$\varepsilon[(\hat{Q}_n^{jj} - Q^{jj})^2] = \varepsilon[(\hat{Q}_n^{jj} - \varepsilon(\hat{Q}_n^{jj}))^2] + [\varepsilon(\hat{Q}_n^{jj} - Q^{jj})]^2$$

$\varepsilon[(\hat{Q}_n^{jj} - \varepsilon(\hat{Q}_n^{jj}))^2]$ can be computed recursively by noting that

$$\varepsilon[(\hat{Q}_n^{jj} - \varepsilon(\hat{Q}_n^{jj}))^2] = \varepsilon[(\hat{Q}_n^{jj})^2] - [\varepsilon(\hat{Q}_n^{jj})]^2$$

The diagonal elements of (4.4.13) are squared and the conditional expected value then evaluated. Use is made of the approximation

$$\varepsilon[(\hat{Q}_{n-1}^{jj}(\Gamma_n^{-1}\Delta x_n^*\Delta x_n^{*T}\Gamma_n^{T-1})^{jj}] \simeq \varepsilon(\hat{Q}_{n-1}^{jj})\ \varepsilon[(\Gamma_n^{-1}\Delta x_n^*\Delta x_n^{*T}\Gamma_n^{T-1})^{jj}]$$

Using this approximation and after extensive algebraic manipulation

$$\varepsilon[(\hat{Q}_n^{jj} - \varepsilon(\hat{Q}_n^{jj}))^2] = J_n^{jj} \qquad (4.4.20)$$

where $J_n^{jj} = \left(\frac{n-1}{n}\right)^2 J_{n-1}^{jj} + \frac{2}{n^2}[((Q + \Delta M_n - T_n^*)^{jj})^2]$

$$T_n^* = \Gamma_n^{-1}(P_{n|n}^* - U_n^*)\Gamma_n^{T-1}$$

So $$\varepsilon[(\hat{Q}_n^{jj} - Q^{jj})^2] = J_n^{jj} + (M_n^{jj})^2 \qquad (4.4.21)$$

In this expression, M_n^{jj} is due to the bias of the estimator and J_n^{jj} is due to possible deviations from this bias. The unconditional mean square Q estimation error follows from the above.

$$E[(\hat{Q}_n^{jj} - Q^{jj})^2] = E(J_n^{jj}) + E[(M_n^{jj})^2]$$

As before, evaluation of $E(J_n^{jj})$ and $E[(M_n^{jj})^2]$ is complicated so only the results of such evaluations are given here.

Under the assumptions that $\hat{R}_o = \overline{R}$ and $\hat{Q}_o = \overline{Q}$, then

$$E(J_n^{jj}) = \left(\frac{n-1}{n}\right)^2 E(J_{n-1}^{jj}) + \frac{2}{n^2}[((\overline{Q} - T_n^*)^{jj})^2 + (U'_{n|1} \textstyle\sum_R U'^T_{n|1}$$

$$+ W'_{n|1} \textstyle\sum_Q W'^T_{n|1})^{jj}] \qquad (4.4.22)$$

where $U'_{n|1}$ is a $\eta x \gamma$ matrix defined by

$$(U'_{n|1})^{j\ell} = \sum_{k=1}^{n} ((g_n^{jT} D_n^{-1} \lambda_{n|k} A_k^*)^{\ell})^2 - ((g_n^{jT} D_n^{-1} A_n^*)^{\ell})^2 + (m_n^j)^{\ell}$$

$$g_n^j = H_n^T f_n^j$$

f_n^j is defined as the j^{th} column of a matrix f_n^T

$$f_n = \Gamma_n^{-1} A_n^*$$

$(m_n^j)^{\ell} = ((f_n^j)^{\ell})^2$ the square of the ℓ^{th} element of the vector f_n^j

$W'_{n|1}$ is a $\eta x \eta$ matrix defined by

$$(W'_{n|1})^{j\ell} = \sum_{k=1}^{n} ((g_n^{jT} D_n^{-1} \lambda_{n|k} D_k \Gamma_k)^{\ell})^2$$

Also from Appendix C,

$$E[(M_n^{jj})^2] = \left(\frac{n-1}{n}\right)^2 E[(M_{n-1}^{jj})^2] + \frac{1}{n^2}[(2\ n\ U'_{n|1} \Sigma_R \overline{U}'^T_{n|1} - U'_{n|1} \Sigma_R U'^T_{n|1})^{jj}$$

$$+ (2\ n\ W'_{n|1}\ \Sigma_Q\ \overline{W}'^T_{n|1} - W'_{n|1}\ \Sigma_Q\ W'^T_{n|1})^{jj}] \qquad (4.4.23)$$

where

$$\overline{U}'_{n|1} \triangleq \frac{1}{n} \sum_{k=1}^{n} U'_{k|1}$$

$$\overline{W}'_{n|1} \triangleq \frac{1}{n} \sum_{k=1}^{n} W'_{k|1}$$

When feedback is used, the estimates of R and Q are used to compute filter gains, so that these filter gains become random variables under both the conditional and unconditional expectation operators. Evaluation of the expected values of nonlinear functions of the R and Q estimates becomes impractical unless approximations are made. The nature of these approximations is $\varepsilon[f(x)] \simeq f(\varepsilon(x))$, where $f(x)$ is a nonlinear function of a random variable x. As before

$$\hat{R}_n^{jj} = \frac{n-1}{n} \hat{R}_{n-1}^{jj} + \frac{1}{n}(\Delta z_n'^{*} \Delta z_n'^{*T} + H_n P^*_{n|n} H_n^T)^{jj}$$

However, now $\Delta z_n'^{*}$ and $P^*_{n|n}$ are computed using the past estimates of R and Q. The conditional expected value of R_n^{jj} is then

$$\varepsilon(\hat{R}_n^{jj}) = \frac{n-1}{n}\varepsilon(\hat{R}_{n-1}^{jj}) + \frac{1}{n}(\varepsilon(\Delta z_n'^{*} \Delta z_n'^{*T}) + H_n \varepsilon(P^*_{n|n}) H_n^T)^{jj}$$

$P^*_{n|n}$ is a function of $\hat{R}_{n-1},\ldots,\hat{R}_o,\hat{Q}_{n-1},\ldots,\hat{Q}_o$, so that it now becomes a random variable and $\varepsilon(P^*_{n|n}) \neq P^*_{n|n}$ as was true in the case of the no feedback case.

$$\varepsilon(\Delta z'^*_n \Delta z'^{*T}_n) = \varepsilon(v_n v_n^T) + H_n \varepsilon(\tilde{x}^*_{n|n} \tilde{x}^{*T}_{n|n}) H_n^T - H_n \varepsilon(\tilde{x}^*_{n|n} v_n^T) - \varepsilon(v_n \tilde{x}^{*T}_{n|n}) H_n^T$$

where $\varepsilon(v_n v_n^T) = R$ as before

$$\varepsilon(\tilde{x}^*_{n|n} \tilde{x}^{*T}_{n|n}) \triangleq \varepsilon(P_{n|n})$$

$$\varepsilon(\tilde{x}^*_{n|n} v_n^T) = \varepsilon(A^*_n v_n v_n^T)$$

In the feedback case, A^*_n is a random variable so it cannot be taken outside the expectation operator.

$$A^*_n = P^*_{n|n} H_n^T \hat{R}^{-1}_{n-1}$$

$$P^*_{n|n} = (I - A^*_n H_n) P^*_{n|n-1} (I - A^*_n H_n)^T + A^*_n \hat{R}_{n-1} A^{*T}_n$$

$$P^*_{n|n-1} = \Phi(n,n-1) P^*_{n-1|n-1} \Phi^T(n,n-1) + \Gamma_n \hat{Q}_{n-1} \Gamma_n^T$$

v_n is independent of A^*_n as past values of $\hat{R}$ and $\hat{Q}$ are used to compute A^*_n. So

$$\varepsilon(A^*_n v_n v_n^T) = \varepsilon(A^*_n)\, \varepsilon(v_n v_n^T) = \varepsilon(A^*_n)\, R$$

So $\varepsilon(\Delta z'^*_n \Delta z'^{*T}_n) = R + H_n \varepsilon(P_{n|n}) H_n^T - H_n \varepsilon(A^*_n)\, R - R\varepsilon(A^{*T}_n) H_n^T$

Define $\Delta F_n = H_n \varepsilon(P_{n|n}) H_n^T + H_n \varepsilon(P^*_{n|n}) H_n^T - H_n \varepsilon(A_n^*) R - R\varepsilon(A_n^{*T}) H_n^T$

Then
$$\varepsilon(\hat{R}_n^{jj}) = \frac{n-1}{n} \varepsilon(\hat{R}_{n-1}^{jj}) + \frac{1}{n}(R + \Delta F_n)^{jj} \qquad (4.4.24)$$

$$= R^{jj} + F_n^{jj}$$

where
$$F_n = \frac{1}{n} \sum_{k=1}^{n} \Delta F_k$$

Approximations must be used to evaluate $\varepsilon(P_{n|n})$, $\varepsilon(P^*_{n|n})$, and $\varepsilon(A_n^*)$.

$$P_{n|n} = (I-A_n^* H_n) P_{n|n-1} (I-A_n^* H_n)^T + A_n^* R A_n^{*T}$$

So
$$\varepsilon(P_{n|n}) = \varepsilon[(I-A_n^* H_n) P_{n|n-1} (I-A_n^* H_n)^T] + \varepsilon(A_n^* R A_n^{*T})$$

The following approximations are used.

$$\varepsilon[(I-A_n^* H_n) P_{n|n-1} (I-A_n^* H_n)^T] \simeq \varepsilon(I-A_n^* H_n)\ \varepsilon(P_{n|n-1})\ \varepsilon[(I-A_n^* H_n)^T]$$

$$\varepsilon(A_n^* R A_n^{*T}) \simeq \varepsilon(A_n^*)\ R\varepsilon(A_n^{*T})$$

$$\varepsilon(A_n^*) \simeq \varepsilon(P^*_{n|n}) H_n \cdot [\varepsilon(\hat{R}_{n-1})]^{-1}$$

Using these approximations, $\varepsilon(P_{n|n})$ can be evaluated recursively.

$$\varepsilon(P_{n|n}) = (I - \varepsilon(A_n^*) H_n)\ \varepsilon(P_{n|n-1}) (I - \varepsilon(A_n^*) H_n)^T + \varepsilon(A_n^*) R\varepsilon(A_n^{*T}) \qquad (4.4.25)$$

with $\varepsilon(P_{n|n-1}) = \Phi(n,n-1)\, \varepsilon(P_{n-1|n-1})\, \Phi^T(n,n-1) + \Gamma_n Q \Gamma_n^T$

$$(4.4.26)$$

Using the same approximations, $\varepsilon(P^*_{n|n})$ can be computed recursively.

$$\varepsilon(P^*_{n|n}) \simeq \varepsilon(I-A^*_n H_n)\, \varepsilon(P^*_{n|n-1})\, \varepsilon(I-A^*_n H_n)^T + \varepsilon(A^*_n)\varepsilon(\hat{R}_{n-1})\varepsilon(A_n^{*T}) \quad (4.4.27)$$

where $\varepsilon(P^*_{n|n-1}) = \Phi(n,n-1)\varepsilon(P^*_{n-1|n-1})\Phi^T(n,n-1) + \Gamma_n \varepsilon(\hat{Q}_{n-1})\Gamma_n^T$

$$(4.4.28)$$

The unconditional expected value of $\hat{R}_n$ follows from the above.

$$E(\hat{R}_n^{jj}) = E(R^{jj}) + E(F_n^{jj}) = \overline{R}^{jj} + \frac{1}{n}\sum_{k=1}^{n} E(\Delta F_k^{jj}) \qquad (4.4.29)$$

where $E(\Delta F_k) = H_k(E(P_{k|k}) + E(P^*_{k|k}))H_k^T - H_k E(\varepsilon(A^*_k)R)$

$$- E(R\varepsilon(A_k^{*T}))H_k^T$$

Additional approximations must be made to evaluate $E(P_{k|k})$, $E(P^*_{k|k})$, and $E(\varepsilon(A^*_k)R)$, namely

$$E(P_{k|k}) \simeq E(I-A^*_k H_k)E(P_{k|k-1})E(I-A^*_k H_k)^T + E(A^*_k)\overline{R}\, E(A_k^{*T}) \qquad (4.4.30)$$

where $E(P_{k|k-1}) = \Phi(k,k-1)E(P_{k-1|k-1})\Phi^T(k,k-1) + \Gamma_k \overline{Q}\Gamma_k^T \qquad (4.4.31)$

Similarly,

$$E(P^*_{k|k}) \simeq E(I-A^*_k H_k)E(P^*_{k|k-1})E(I-A^*_k H_k)^T + E(A^*_k)E(\hat{R}_{k-1})E(A_k^{*T}) \quad (4.4.32)$$

where $E(P^*_{k|k-1}) = \Phi(k,k-1)E(P^*_{k-1|k-1})\Phi(k,k-1) + \Gamma_k E(\hat{Q}_{k-1})\Gamma_k^T$

$$(4.4.33)$$

and $$E(A^*_k) \simeq E(P^*_{k|k})\ H_k^T\ [E(\hat{R}_{k-1})]^{-1}$$

In a similar fashion, the conditional and unconditional expected value of the estimate of Q can be obtained.

$$\varepsilon(\hat{Q}^{jj}_n) = \frac{n-1}{n}\ \varepsilon(\hat{Q}^{jj}_{n-1}) + \frac{1}{n}\ (Q + \Delta M_n)^{jj} \qquad (4.4.34)$$

where after algebraic manipulation, ΔM_n can be expressed in the following form

$$\Delta M_n \simeq \overline{Q} - Q + f_n[(R - \overline{R}) + H_n(\varepsilon(P_{n|n-1}) - \varepsilon(P^*_{n|n-1}))H_n^T]f_n^T \qquad (4.4.35)$$

where here $$f_n = \Gamma_n^{-1}\ \varepsilon(A^*_n)$$

Similarly,

$$E(\hat{Q}^{jj}_n) = \frac{n-1}{n}\ E(\hat{Q}^{jj}_{n-1}) + \frac{1}{n}\ (\overline{Q} + E(\Delta M_n))^{jj}$$

where $E(\Delta M_n)$ is evaluated approximately by using unconditional expected values instead of conditional expected values in (4.4.35).

$$E(\Delta M_n) \simeq f_n H_n(E(P_{n|n-1}) - E(P^*_{n|n-1}))H_n^T f_n^T$$

where here $$f_n = \Gamma_n^{-1}\ E(A^*_n)$$

The mean squared estimation error of the R and Q estimators can be found approximately by using the results of the no feedback estimators replacing $P_{n|n}$ by $\varepsilon(P_{n|n})$ or $E(P_{n|n})$, $P^*_{n|n}$ by $\varepsilon(P^*_{n|n})$ or $E(P^*_{n|n})$, etc., the conditional or unconditional expected values being used depending upon whether the conditional or unconditional mean squared error is being evaluated.

Once estimates of R and Q have been found by the above procedures, some way must be found for incorporating the a priori estimates of R and Q into a combined estimate of these quantities. Presumably, along with the a priori estimates of R and Q there is available some measure of the quality of these estimates, say the variances of the estimates. Expressions have been developed for evaluating the quality (mean square error) of the estimates based upon measurements alone. Then a reasonable, but not necessarily optimal, technique for combining these two estimates would be in some inverse variance fashion. In the case of the R estimate,

$$\hat{R}^*_n = \left(\sum_{R^*_n}\right)\left[\left(\sum_{R_n}\right)^{-1}\hat{R}_n + \left(\sum_{R_o}\right)^{-1}\hat{R}_o\right]$$

where $\hat{R}^*_n$ is the combined estimate, $\hat{R}_n$ is the estimate based upon the measurements alone, and $\hat{R}_o$ is the a priori estimate. $\sum_{R_O}$ is the unconditional mean square R estimation error matrix $E[(\hat{R}_n-R)^{jj})^2]\delta_{jk}$ and $\sum_{R_O}$ is the variance of the a priori estimate of R, and

$$\Sigma_{R_n^*}^{-1} = \Sigma_{R_n}^{-1} + \Sigma_{R_o}^{-1}$$

If $E(\hat{R}_n) = \overline{R}$ and $\hat{R}_o = \overline{R}$, then

$$\Sigma_{R_n^*}^{ij} = E[(\hat{R}_n^{*jj} - R^{jj})^2]\delta_{ij}$$

Similar expressions can be developed for a combined Q estimate.

4.5 Review of Procedures Suggested by Others

Several authors have studied the problem of optimal filtering when the parameters describing the statistics of the measurement and driving noises are not accurately known. A brief summary of the results of their work is included in this chapter. As will be seen, the estimators are simple to use but are suboptimal, that is, no optimality condition is satisfied by the solution. In many applications the resulting estimators may be biased or may not actually exist. However, in some applications, such estimators may provide useful solutions to the problem.

A technique of estimating R and Q developed by Shellenbarger (Ref. 31) utilizes the theory of maximum likelihood estimation outlined in Chapter 2, but applies the theory only to obtain an approximate solution. His technique is based upon estimating the parameters of R and Q using the information obtained at one measurement time and then performing some average over current and past estimates to obtain a combined estimate. If there is insufficient information available at each measurement time to estimate all of the unknown elements of R and Q, then his technique cannot be used. Unfortunately, many interesting applications of optimal filtering have a small dimension measurement compared to the dimension of the state so that there may be little information at each measurement time upon which to base estimates of R and Q. However, if it is assumed that the driving noise statistics are precisely known a priori, then there is always sufficient information in the measurements to estimate the statistics of the measurement noise.

Given n measurements $z_1,..,z_n$, and the conditonal maximum likelihood estimate of the state prior to each measurement (conditioned upon the assumed values of R and Q), the joint probability density function of the n measurements can be written as

$$f(z_1,..,z_n) = f(z_n|z_{n-1},..,z_1)f(z_{n-1}|z_{n-2},..,z_1)...f(z_1) \quad (4.5.1)$$

where $$z_k = H_k x_k + v_k$$

Given that all the assumptions used in deriving the maximum likelihood state estimator of Chapter 2 are valid, then

$$f(z_k|z_{k-1},..,z_1) = \frac{1}{(2\pi)^{\gamma/2}|B_k|^{1/2}} e^{-\frac{1}{2}(\Delta z_k^T B_k^{-1} \Delta z_k)}$$

where $$\Delta z_k = z_k - H_k \hat{x}_{k|k-1}$$

$$B_k = R_k + H_k P_{k|k-1} H_k^T$$

Therefore

$$f(z_1,..,z_n) = \prod_{k=1}^{n} \frac{1}{(2\pi)^{\gamma/2}|B_k|^{1/2}} e^{-\frac{1}{2}(\Delta z_k^T B_k^{-1} \Delta z_k)}$$

For estimation of R_n with $Q_1,\ldots,Q_n$ known, Shellenbarger suggests maximizing $f(z_1,..,z_n)$ with respect to R_n and solving for $\hat{R}_n$. However, he realizes that the solution depends upon the unknown $R_1,..,R_{n-1}$. To solve for all R_i, $f(z_1,..,z_n)$ would have to be maximized with respect to $R_i (1 \leq i \leq n)$ and the resulting equations solved for $\hat{R}_i$. Shellenbarger dismisses this approach as being infeasible for any nontrivial system.

Rather than simultaneously estimating all R_i, he suggests that the single measurement conditional likelihood function $f(z_n|z_{n-1},..,z_1)$ be maximized with respect to R_n, using past estimates of $R_1,..,R_{n-1}$ to compute the necessary quantities appearing in this likelihood function.

$$\frac{\partial \ln f(z_n|z_{n-1},..,z_1)}{\partial R_n} = B_n^{-1} - B_n^{-1} \Delta z_n \Delta z_n^T B_n^{-1} \qquad (4.5.2)$$

(4.5.2) is set to zero and the resulting equation solved for R_n. $\hat{x}_{n|n-1}$ and $P_{n|n-1}$ are not precisely known as they are defined as the maximum likelihood estimate of the state and its covariance conditioned upon the true values of $R_1,\ldots,R_{n-1}$. For this solution they must be evaluated using some average of the estimates of the past R_i. The results of such a procedure are

$$B_n^{*-1}(B_n^* - \Delta z_n^* \Delta z_n^{*T})B_n^{*-1} = 0 \qquad (4.5.3)$$

where
$$B_n^* \triangleq R_n + H_n P^*_{n|n-1} H_n^T$$

$$z_n^* \triangleq z_n - H_n \, x^*_{n|n-1}$$

$\hat{x}^*_{n|n-1}$ and $P^*_{n|n-1}$ are the values of $\hat{x}_{n|n-1}$ and $P_{n|n-1}$ evaluated recursively using some average of past estimates of R at each updating time. Then the estimate of R_n is defined by

$$\hat{R}_n = \Delta z_n^* \, \Delta z_n^{*T} - H_n P^*_{n|n-1} H_n^T \tag{4.5.4}$$

The conditional expected value of $\hat{R}_n$ is

$$\varepsilon(\hat{R}_n) = \varepsilon[(v_n - H_n \tilde{x}^*_{n|n-1})(v_n - H_n \tilde{x}^*_{n|n-1})^T] - H_n \, \varepsilon(P^*_{n|n-1}) H_n^T$$

$$= R_n + H_n (\varepsilon(\tilde{x}^*_{n|n-1} \tilde{x}^{*T}_{n|n-1}) - \varepsilon(P^*_{n|n-1})) H_n^T$$

$\varepsilon(\tilde{x}^*_{n|n-1} \tilde{x}^{*T}_{n|n-1})$ represents the conditional covariance of the state estimation error, conditioned upon the true values of R and Q and the fact that estimates of the past values of the measurement noise covariance matrices have been used in computing filter gains. $\varepsilon(P^*_{n|n-1})$ represents the average (over the ensemble of all measurement and driving noises) computed state error covariance matrix. As was shown in the previous section, when past values of estimates of R or Q are used to compute filter gains, evaluation of these two quantities is exceedingly difficult and in general cannot be performed without approximations. Shellenbarger states without proof that

$$\varepsilon(\tilde{x}^*_{n|n-1}\tilde{x}^{*T}_{n|n-1}) = \varepsilon(P^*_{n|n-1}) \qquad (4.5.5)$$

and thus concludes that

$$\varepsilon(\hat{R}_n) = R_n \qquad (4.5.6)$$

This demonstration of unbiasedness depends upon the validity of (4.5.5), something that Shellenbarger does not adequately discuss.

The case of estimating Q_n with $R_1,..,R_n$ known is considerably more complicated that the previous case. The solution depends upon the rank of the measurement matrix H_n. The forcing function matrix Γ_n is presumed to be the identity matrix. The single measurement conditional likelihood function is maximized with respect to Q_n, using past estimates of $Q_1,..,Q_{n-1}$ to compute the necessary quantities appearing in this likelihood function.

$$\frac{\partial \ln f(z_n|z_{n-1},..,z_1)}{\partial Q_n} = H_n^T B_n^{-1}(B_n - \Delta z_n \Delta z_n^T)B_n^{-1}H_n^T \qquad (4.5.7)$$

where $B_n = R_n + H_n(\Phi(n,n-1)P_{n-1|n-1}\ \Phi^T(n,n-1) + Q_n)H_n^T$

(4.5.7) is set to zero and the resulting equation solved for $\hat{Q}_n$. As in the estimation of R_n, $\hat{x}_{n|n-1}$ and $P_{n-1|n-1}$ are evaluated using some average of past estimates of Q_i.

$$H_n^T B_n^{*-1} (H_n Q_n H_n^T - C_n^*) B_n^{*-1} H_n = 0 \tag{4.5.8}$$

where $C_n^* \triangleq \Delta z_n^* \, \Delta z_n^{*T} - R_n - H_n \, \Phi(n,n-1) P^*_{n-1|n-1} \, \Phi^T(n,n-1) H_n^T$

If H_n is square and possesses an inverse, then

$$\hat{Q}_n \triangleq H_n^{-1} \, C_n^* \, H_n^{T-1} \tag{4.5.9}$$

The conditional expected value of $\hat{Q}_n$ is then

$$\varepsilon(\hat{Q}_n) = Q_n + \Phi(n,n-1)[\varepsilon(\tilde{x}^*_{n-1|n-1}\tilde{x}^{*T}_{n-1|n-1}) - \varepsilon(P^*_{n-1|n-1})]\Phi^T(n,n-1) \tag{4.5.10}$$

Again, Shellenbarger assumes that

$$\varepsilon(\tilde{x}^*_{n-1|n-1}\tilde{x}^{*T}_{n-1|n-1}) = \varepsilon(P^*_{n-1|n-1}) \tag{4.5.11}$$

and thus concludes that

$$\varepsilon(\hat{Q}_n) = Q_n \tag{4.5.12}$$

The same comments apply here as before concerning the validity of (4.5.11).

If H_n^{-1} does not exist, but $(H_n^T Y_n^{-1} H_n)^{-1}$ exists, where

$$Y_n = R_n + H_n \, \Phi(n,n-1) P^*_{n-1|n-1} \, \Phi^T(n,n-1) H_n^T$$

then a solution for $\hat{Q}_n$ can be obtained from (4.5.8) by finding B_n^{*-1} using the matrix inversion lemma and carrying out some matrix manipulation.

$$\hat{Q}_n \triangleq (H_n^T Y_n^{-1} H_n)^{-1} H_n^T Y_n^{-1} C_n^* Y_n^{-1} H_n (H_n^T Y_n^{-1} H_n)^{-1} \tag{4.5.13}$$

The conditional expected value of this estimate of $\hat{Q}_n$ is equal to (4.5.10).

If neither H_n^{-1} nor $(H_n^T Y_n^{-1} H_n)^{-1}$ exists, then a unique solution of (4.5.7) does not exist. However, by use of the generalized inverse of H_n a particular solution can be defined which satisfies (4.5.8).

$$\hat{Q}_n \triangleq H_n^{\#} \; C_n^* \; H_n^{T\#} \tag{4.5.14}$$

where $H_n^{\#}$ is the generalized inverse of H_n. If $(H_n H_n^T)^{-1}$ exists, then

$$H_n^{\#} = H_n^T (H_n H_n^T)^{-1} \tag{4.5.15}$$

The conditional expected value of (4.5.14) using (4.5.15) is

$$\varepsilon(\hat{Q}_n) = H_n^T (H_n H_n^T)^{-1} H_n Q_n H_n^T (H_n H_n^T)^{-1} H_n \tag{4.5.16}$$

$$+ H_n^T (H_n H_n^T)^{-1} H_n \Phi(n,n-1) [\varepsilon(\tilde{x}^*_{n-1|n-1} \tilde{x}^{*T}_{n-1|n-1})$$

$$- \varepsilon(P^*_{n-1|n-1})] \; \Phi^T(n,n-1) H_n^T (H_n H_n^T)^{-1} H_n$$

As with the explicit suboptimal estimator, the real difficulty in the use of Shellenbarger's method is when the elements of R_n and Q_n are to be estimated simultaneously. (4.5.3) and (4.5.8) must be solved for R_n and Q_n. However, there is no possibility that these equations can be solved uniquely for these two quantities since the equations are not independent. In essence, Shellenbarger suggests that the number of unknown elements of R_n and Q_n be reduced until the number of unknowns is equal to the number of independent equations. If R_n is assumed to be diagonal, the number of unknown elements in R_n is reduced from $\gamma(\gamma+1)/2$ to γ. However, a solution for these diagonal elements and Q_n is possible only when there are redundant measurements, or $(H_n^T H_n)^{-1}$ exists. In such a case,

$$c(\hat{R}_n) \triangleq (I-N_n{*}N_n)^{-1} \; c(\Delta z_n^* \Delta z_n^{*T} - N_n \Delta z_n^* \Delta z_n^{*T} N_n) \qquad (4.5.17)$$

where c( ) represents a column vector whose elements are the diagonal elements of the matrix argument and

$$N_n \triangleq H_n (H_n^T H_n)^{-1} H_n^T$$

$N_n{*}N_n$ is a matrix whose elements are the squares of the corresponding elements of N_n.

Once the diagonal elements of $\hat{R}_n$ are obtained, $\hat{Q}_n$ can be obtained from (4.5.13) using estimates of R_n in place of the unknown R_n. Clearly this technique has applicability in only those cases when redundant measurements are taken at each measurement time. In most applications the dimension of the state is large compared with that of the measurement, in which case the unknown elements of R_n and Q_n cannot be simultaneously estimated by Shellenbarger's method.

Dennis (Ref. 5) uses an entirely different technique for obtaining estimators for R and Q, but as in the case of Shellenbarger's method, he essentially relies upon having sufficient information in each measurement to define estimates of R and Q based upon a single residual. If sufficient information is not available, or if some components of the driving noise are not observable from one residual alone, then Dennis suggests lagging the driving noise variance estimation with respect to the measurement noise variance estimation, that is, use past as well as current residuals to obtain some estimate of the driving noise covariance.

Dennis obtains a functional relationship between certain residuals and the measurement and driving noises. From this relationship he postulates the form of the estimators. No criterion of optimality is used, and his proof of unbiasedness and stability of the resulting estimation loop is questionable.

At each measurement time, the existence of a minimum variance or maximum likelihood state estimator is presumed, with estimates of R and Q used to compute the proper residual weighting matrices. From the recursive state estimate updating equation (2.3.38),

$$\hat{x}^*_{n|n} = \Phi(n,n-1)\hat{x}^*_{n-1|n-1} + A^*_n(z_n - H_n\ \Phi(n,n-1)\hat{x}^*_{n-1|n-1})$$

The assumed models for the state and measurement are

$$x_n = \Phi(n,n-1)x_{n-1} + \Gamma_n w_n$$

$$z_n = H_n\, x_n + v_n$$

Then the following expression can be obtained for the estimation error.

$$\overset{\sim *}{x}_{n|n} = \overset{\wedge *}{x}_{n|n} - x_n = (I-A_n^* H_n)\ \Phi(n,n-1)\overset{\sim *}{x}_{n-1|n-1}$$

$$+ A_n^* v_n + (A_n^* H_n - I)\Gamma_n w_n$$

Consider the three residuals

$r_n^m \triangleq z_n - H_n \overset{\wedge *}{x}_{n|n}$ a γ x 1 vector

$\hat{r}_n^m \triangleq z_n - H_n \overset{\wedge *}{x}_{n|n-1}$ a γ x 1 vector

$r_n^s \triangleq \overset{\wedge *}{x}_{n|n} - \overset{\wedge *}{x}_{n|n-1}$ a β x 1 vector

It can be shown that

$$r_n^m = (I-H_n A_n^*)v_n + H_n(I-A_n^* H_n)\Gamma_n w_n + H_n(A_n^* H_n - I)\Phi(n,n-1)\overset{\sim *}{x}_{n-1|n-1} \quad (4.5.18)$$

$$\hat{r}_n^m = v_n + H_n\Gamma_n w_n - H_n\Phi(n,n-1)\overset{\sim *}{x}_{n-1|n-1} \quad (4.5.19)$$

$$r_n^s = A_n^* v_n + A_n^* H_n \Gamma_n w_n - A_n^* H_n \Phi(n,n-1)\overset{\sim *}{x}_{n-1|n-1} \quad (4.5.20)$$

These equations are singular in the sense that v_n and w_n can never be exactly determined from the residuals alone. However, in terms of squared residuals, some nonsingular mappings of averages can be obtained.

Begin by considering the two residuals r_n^m and r_n^s in the following form.

$$\begin{bmatrix} \hat{r}_n^m \\ \\ r_n^s \end{bmatrix} = \begin{bmatrix} I & H_n\Gamma_n \\ \\ A_n^* & A_n^*H_n\Gamma_n \end{bmatrix} \begin{bmatrix} v_n \\ \\ w_n \end{bmatrix} - \begin{bmatrix} H_n\Phi(n,n-1) \\ \\ A_n^*H_n\Phi(n,n-1) \end{bmatrix} \tilde{x}^*_{n-1|n-1} \qquad (4.5.21)$$

Or
$$\bar{r}_n = K_n\,\phi_n - C_n\,\tilde{x}^*_{n-1|n-1} \qquad (4.5.22)$$

where
$$\bar{r}_n^T = (\hat{r}_n^{mT},\ r_n^{sT})$$

$$\phi_n^T = (v_n^T,\ w_n^T)$$

$$K_n = \begin{bmatrix} I & H_n\Gamma_n \\ \\ A_n^* & A_n^*H_n\Gamma_n \end{bmatrix}$$

$$C_n = \begin{bmatrix} H_n\Phi(n,n-1) \\ \\ A_n^*H_n\Phi(n,n-1) \end{bmatrix}$$

Consider the i^{th} element of $\bar{r}_n$.

$$r_n^i = \sum_j K_n^{ij}\,\phi_n^j - C_n^{ij}\,\tilde{x}^{*j}_{n-1|n-1}$$

Squaring r_n^i leads to

$$(r_n^i)^2 = \sum_j (K_n^{ij})^2 (\phi_n^j)^2 + \sum_j \sum_{k\neq j} K_n^{ij} K_n^{ik} \phi_n^j \phi_n^k \tag{4.5.23}$$

$$- 2 \sum_j \sum_k K_n^{ij} C_n^{ik} \phi_n^j \tilde{x}^{*k}_{n-1|n-1} + \sum_j \sum_k C_n^{ij} C_n^{ik} \tilde{x}^{*k}_{n-1|n-1} \tilde{x}^{*j}_{n-1|n-1}$$

Assuming that v_n and w_n are mutually independent zero mean normal random vectors with time independent statistics, then the only terms of interest in (4.5.23) are the first and last because the average values of the others are zero. Therefore

$$(r_n^i)^2 = \sum_j (K_n^{ij})^2 (\phi_n^j)^2 + \sum_j \sum_k C_n^{ij} C_n^{ik} \tilde{x}^{*k}_{n-1|n-1} \tilde{x}^{*j}_{n-1|n-1} + \alpha_n^i \tag{4.5.24}$$

where α_n^i is the sum of all other terms in (4.5.23) and is by definition zero mean.

Next Dennis assumes that $(\phi_n^j)^2$ is a Rayleigh variable having mean $\bar{R}_n$ or $\bar{Q}_n$ as appropriate, where $\bar{R}_n$ and $\bar{Q}_n$ are vectors of the diagonal elements of R_n and Q_n. Thus

$$\begin{bmatrix} (\hat{r}_n^m)^2 \\ (r_n^s)^2 \end{bmatrix} = \tilde{K}_n \begin{bmatrix} \bar{R}_n + \zeta_n \\ \bar{Q}_n + \xi_n \end{bmatrix} + S_n + \zeta_n \tag{4.5.25}$$

where $\tilde{K}_n$ is a matrix composed of the squared elements of K_n, ζ_n and ξ_n are zero mean random vectors,

$$S_n = \begin{bmatrix} C_n^{1T}\tilde{x}^*_{n-1|n-1}\tilde{x}^{*T}_{n-1|n-1}C_n^1 \\ \vdots \\ C_n^{\gamma+\beta\,T}\tilde{x}^*_{n-1|n-1}\tilde{x}^{*T}_{n-1|n-1}C_n^{\gamma+\beta} \end{bmatrix}$$

where C_n^j is the j^{th} column of C_n^T.

$$\alpha_n^T = (\alpha_n^1,..,\alpha_n^{\gamma+\beta})$$

$(\hat{r}_n^m)^2$ and $(r_n^s)^2$ denote vectors whose elements are the squares of $\hat{r}_n^m$ and r_n^s respectively.

(4.5.25) is central to Dennis's development of noise variance estimation. In particular, it can be seen that if $\tilde{K}_n$ is of full rank then

$$\begin{bmatrix} \bar{R}_n + \zeta_n \\ \bar{Q}_n + \xi_n \end{bmatrix} = \tilde{K}_n^{-1} \begin{bmatrix} (\hat{r}_n^m)^2 \\ (r_n^s)^2 \end{bmatrix} - \tilde{K}_n^{-1}S_n - \tilde{K}_n^{-1}\alpha_n$$

Or

$$\begin{bmatrix} \bar{R}_n \\ \bar{Q}_n \end{bmatrix} = \tilde{K}_n^{-1} \begin{bmatrix} (\hat{r}_n^m)^2 \\ (r_n^s)^2 \end{bmatrix} - \tilde{K}_n^{-1}S_n - \tilde{K}_n^{-1}\alpha_n - \begin{bmatrix} \zeta_n \\ \xi_n \end{bmatrix} \qquad (4.5.26)$$

From an examination of (4.5.26) Dennis postulates the form of the estimator for R_n and Q_n based upon one set of residuals.

$$\begin{bmatrix} \hat{\bar{R}}_n \\ \\ \hat{\bar{Q}}_n \end{bmatrix} \triangleq \tilde{K}_n^{-1} \begin{bmatrix} (\hat{r}_n^m)^2 \\ \\ (r_n^s)^2 \end{bmatrix} - \tilde{K}_n^{-1} S_n^*$$

where S_n^* is the matrix S_n with $P^*_{n-1|n-1}$ substituted for $(\tilde{x}^*_{n-1|n-1}\tilde{x}^{*T}_{n-1|n-1})$.

The above estimates are not those used for the computation of the filter gain matrices but rather some average of past estimates. This allows for some "smoothing" of the single residual set estimates.

$$\hat{\bar{R}}_n^* \triangleq \sum_{j=1}^{n} \omega_j^R \hat{\bar{R}}_j; \qquad \sum_{j=1}^{n} \omega_j^R = I$$

$$\hat{\bar{Q}}_n^* \triangleq \sum_{j=1}^{n} \omega_j^Q \hat{\bar{Q}}_j; \qquad \sum_{j=1}^{n} \omega_j^Q = I$$

where ω_j^R and ω_j^Q are weighting factors that can be arbitrarily chosen.

The conditional expected value of such an estimate is most difficult to obtain since the matrix K_n is a random function of the previous estimates of R and Q. Dennis does show that for scalar measurements and no driving noise the estimate $\hat{\bar{R}}_n^*$ is to first order independent of variations in the value of R used to compute the gain matrix K_n. He states that this is true whether or not driving noise is present but does not show that the estimate $\hat{\bar{R}}_n^*$ is independent of variations in the value of Q used to compute K_n. He also states that the estimate $\hat{\bar{Q}}_n^*$ is independent of variations in Q used to compute K_n for a scalar state variable. What all these statements mean with respect to the biasedness of the estimates in realistic situations is not clear.

If the matrix K_n is singular, then a slightly different procedure must be used. It is assumed that R and Q are constant or slowly varying in time. A time average of (4.5.25) is taken prior to inversion.

$$\sum_{j=m}^{n} \Omega_j \begin{bmatrix} (\hat{r}_j^m)^2 \\ (r_j^s)^2 \end{bmatrix} = \sum_{j=m}^{n} \left[\Omega_j K_j \begin{bmatrix} \bar{R}_j + \zeta_j \\ \bar{Q}_j + \xi_j \end{bmatrix} + \Omega_j S_j + \Omega_j \alpha_j \right]$$

$$\simeq \sum_{j=m}^{n} \Omega_j \tilde{K}_j \begin{bmatrix} \bar{R} \\ \bar{Q} \end{bmatrix} + \Omega_j S_j^*$$

where Ω_j are arbitrary weighting factors with

$$\sum_{j=m}^{n} \Omega_j = I$$

If $\left[\sum_{j=m}^{n} \Omega_j K_j \right]^{-1}$ exists,

$$\begin{bmatrix} \hat{\bar{R}} \\ \hat{\bar{Q}} \end{bmatrix} = \sum_{j=m}^{n} \Omega_j K_j^{\ -1} \left[\sum_{j=m}^{n} \Omega_j \begin{bmatrix} (\hat{r}_j^m)^2 \\ (r_j^s)^2 \end{bmatrix} - \Omega_j S_j^* \right]$$

Dennis attempts to show that for some n, the weighted $\tilde{K}_j$ matrix is nonsingular. However, using his own analysis, if the measurement matrix is time invariant, the weighted $\tilde{K}_j$ matrix is always singular if $\tilde{K}_j$ itself is, thus limiting the applicability of the solution to cases when this is not true.

Smith (Ref. 33) has studied the problem of real time estimation of the state and the measurement noise covariance but only obtains a suboptimal solution of the problem. In his dynamical model of the state, there may be noise driving the state but it is assumed that the statistics of this noise are precisely known.

The state obeys the recursive relationship

$$x_n = \Phi(n,n-1)\, x_{n-1} + w_n \tag{4.5.27}$$

The measurements of the system state have the usual form

$$z_n = H_n\, x_n + v_n \tag{4.5.28}$$

where z_n is a γ x 1 vector. This single vector measurement is equivalent to γ scalar measurements when the measurements are independent, or equivalently, when the measurement noise covariance matrix is diagonal. In this case, the j^{th} scalar measurement at time n is given by

$$z_n^j = h_n^{jT}\, x_n + v_n^j \tag{4.5.29}$$

here h_n^j is the j^{th} column of the matrix H_n^T.

It is assumed that the initial value of the state is normally distributed and that w_n is also normal. The distribution of each v_n^j is also normal with zero mean and variance

R_n^{jj}. It is further assumed that R_n^{jj} can be represented as

$$R_n^{jj} = k^j \, R_{nom_n}^{jj} \tag{4.5.30}$$

where k^j is a time invariant but unknown precision factor associated with each component of the measurement. Smith assumes that each k^j has an inverted Gamma distribution which describes the a priori uncertainty in the value of k^j. The form (4.5.30) is used for R_n so that deterministic time-varying characteristics of R_n can be easily modeled. The probability density function of each k^j can be represented as

$$\begin{aligned} f(k) &= C \left(\frac{a\ b}{k}\right)^{\frac{a}{2}+1} e^{-\frac{1}{2}\frac{a\ b}{k}} & k \geq 0 \\ &= 0 & k < 0 \end{aligned} \tag{4.5.31}$$

where $$C = \frac{1}{2^{(\frac{a}{2}+1)}\ \Gamma(\frac{a}{2}+1)\ b}$$

and a and b are parameters of the distribution of k. The mean of this distribution is proportional to b.

$$E(k) \triangleq \int_0^\infty k\ f(k)\ dk = \frac{a}{a-2}\ b \tag{4.5.32}$$

The joint conditional probability density function of the state x_n and the parameter k is given by

$$f(x_n,k|Z_n) = \frac{f(x_n,k|Z_{n-1})\ f(z_n|x_n,k,Z_{n-1})}{f(z_n|Z_{n-1})} \tag{4.5.33}$$

where Z_n represents the vector of n measurements, Z_{n-1} represents the vector of n-1 measurements, and

$$f(z_n|Z_{n-1}) = \int_0^\infty f(z_n|Z_{n-1},k)\ f(k)\ dk \qquad (4.5.34)$$

The conditional probability density function of the measurement z_n given k, x_n, and Z_{n-1} is

$$f(z_n|x_n,k,Z_{n-1}) = \frac{1}{(2\pi)^{\gamma/2}|R_n|^{1/2}}\ e^{-\frac{1}{2}(z_n-H_nx_n)^T R_n^{-1}(z_n-H_nx_n)} \qquad (4.5.35)$$

Since the components of the vector measurement error v_n are independent, Smith considers z_n as a scalar since a vector z_n can be thought of as a sequence of scalar measurements as mentioned before. Thus all subsequent expressions involving the measurement z_n can be thought of as expressions involving a component of a vector measurement. For notational convenience the superscript j denoting the component is dropped. Then with $R_n = k\ R_{nom_n}$,

$$f(z_n|x_n,k,Z_{n-1}) = \frac{k^{-1/2}}{(2\pi R_{nom_n})^{1/2}}\ e^{-\frac{1}{2}(z_n-h_n^Tx_n)^2/\ k\ R_{nom_n}} \qquad (4.5.36)$$

By Bayes' rule

$$f(x_n,k|Z_{n-1}) = f(x_n|k,Z_{n-1})\ f(k|Z_{n-1}) \qquad (4.5.37)$$

In Chapter 2 it was shown that

$$f(x_n|k,Z_{n-1}) = \frac{1}{(2\pi)^{\beta/2}|P_{n|n-1}|^{1/2}} e^{-\frac{1}{2}(x_n-\hat{x}_{n|n-1})^T P_{n|n-1}^{-1}(x_n-\hat{x}_{n|n-1})} \tag{4.5.38}$$

where $\hat{x}_{n|n-1}$ is the maximum likelihood estimate of x_n before the n^{th} measurement, $P_{n|n-1}$ is the conditional covariance of x_n about its conditional mean $\hat{x}_{n|n-1}$, and β is the dimension of the state x_n. Both the state estimate and its conditional covariance are functions of the unknown k.

It will now be shown that $f(x_n,k|Z_n)$ has a particular form and that this form is preserved after repeated measurements.

It is assumed that the distribution of the initial state x_o is a normal distribution with mean $\hat{x}_{o|o}$ and covariance $P_{o|o}$.

$$f(x_o) = \frac{1}{(2\pi)^{\beta/2}|P_{o|o}|^{1/2}} e^{-\frac{1}{2}(x_o-\hat{x}_{o|o})^T P_{o|o}^{-1}(x_o-\hat{x}_{o|o})} \tag{4.5.39}$$

Initially, x_o is independent of the parameter k so the joint probability density function of x_o and k is

$$f(x_o,k) = f(x_o)\ f(k) \tag{4.5.40}$$

The joint probability density function of the state and the parameter k immediately before the first measurement is given by

$$f(x_1,k) = f(x_1|k)\ f(k) \tag{4.5.41}$$

It is easy to show that

$$f(x_1|k) = \frac{1}{(2\pi)^{\beta/2}|P_{1|o}|^{1/2}} e^{-\frac{1}{2}(x_1-\hat{x}_{1|o})^T P_{1|o}^{-1}(x_1-\hat{x}_{1|o})} \qquad (4.5.42)$$

where $\hat{x}_{1|o} = \Phi(1,0)\ \hat{x}_{o|o}$

$$P_{1|o} = \Phi(1,0)\ P_{o|o}\ \Phi^T(1,0) + Q_1$$

where Q_1 is the covariance of the driving noise w_1. Then using (4.5.31) and (4.5.42), (4.5.41) becomes

$$f(x_1,k) = C_1\ k^{-(\frac{a}{2}+1)}\ e^{-\frac{1}{2}\left[\frac{a\ b}{k} + (x_1-\hat{x}_{1|o})^T P_{1|o}^{-1}(x_1-\hat{x}_{1|o})\right]} \qquad (4.5.43)$$

where $C_1 = \dfrac{C\ (a\ b)^{\frac{a}{2}+1}}{(2\pi)^{\beta/2}|P_{1|o}|^{1/2}}$

Because of the form of (4.5.43), $f(x_1,k)$ is termed a normal inverted gamma probability density function. Then by (4.5.33)

$$f(x_1,k|Z_1) = C_2\ k^{-\frac{1}{2}(a+3)}\ e^{-\frac{1}{2}\left[\frac{1}{k\ R_{nom_1}}(z_1-h_1^T x_1)^2 + \frac{a\ b}{k} + (x_1-\hat{x}_{1|o})^T P_{1|o}^{-1}(x_1-\hat{x}_{1|o})\right]} \qquad (4.5.44)$$

where $C_2 = \dfrac{C_1}{(2\pi R_{nom_1})^{1/2}\ f(z_1)}$

is the normalizing coefficient.

After extensive manipulation, (4.5.44) can be written in terms of new parameters $\hat{x}_{1|1}$, $P_{1|1}$, a', and b'.

$$f(x_1,k|Z_1) = C_2\, k^{-(\frac{a'}{2}+1)}\, e^{-\frac{1}{2}\left[\frac{a'b'}{k} + (x_1-\hat{x}_{1|1})^T P_{1|1}^{-1}(x_1-\hat{x}_{1|1})\right]} \tag{4.5.45}$$

where
$$\hat{x}_{1|1} = \hat{x}_{1|o} + A_1(z_1 - h_1^T\hat{x}_{1|o}) \tag{4.5.46}$$

$$P_{1|1} = P_{1|o} - A_1(k\, R_{nom_1} + h_1^T P_{1|o} h_1) A_1^T \tag{4.5.47}$$

$$A_1 = P_{1|o} h_1 / (h_1^T P_{1|o} h_1 + k\, R_{nom_1}) \tag{4.5.48}$$

$$b' = \frac{1}{a+1}\left[a\, b + \frac{k(z_1 - h_1^T\hat{x}_{1|o})^2}{(h_1^T P_{1|o} h_1 + k\, R_{nom_1})}\right] \tag{4.5.49}$$

$$a' = a + 1 \tag{4.5.50}$$

Thus, the joint conditional density function after the first measurement also has a normal inverted gamma form.

Using the same procedures as above, it can be shown that $f(x_n,k|Z_n)$ has a normal inverted gamma form for any n. The appropriate parameters of the density function can be computed using recursive relationships of the form shown in (4.5.46) - (4.5.50). Each component of the measurement has associated with it its own a, b, R_{nom_n}, and h_n, which are used in these recursive relationships when that particular type of observation is being considered. The resulting a' and b' are not updated again until another observation of the same type is considered. On the other hand, $\hat{x}_{n|n-1}$ and $P_{n|n-1}$, being associated with the state x_n, which is common to all observation types, are updated at each and every data processing stage.

Unfortunately, Eqs. (4.5.46) - (4.5.50) cannot be computed in a real problem because they involve k, which is unknown. Thus in order to compute $\hat{x}_{n|n}$, $P_{n|n}$, and b', an estimate of k is required. Smith dismisses the question of strict optimality and observes that for large a the parameter b is almost equal to the mean of the k distribution. An estimate of k is then defined to be equal to the parameter b, and the following estimation equations are obtained.

$$\hat{x}^*_{n|n} = \hat{x}^*_{n|n-1} + A^*_n(z_n - h^T_n \hat{x}^*_{n|n-1}) \tag{4.5.51}$$

$$P^*_{n|n} = P^*_{n|n-1} - A^*_n(h^T_n P^*_{n|n-1} h_n + \hat{k}\, R_{nom_n}) A^{*T}_n \tag{4.5.52}$$

$$A^*_n = P^*_{n|n-1} h_n / (h^T_n P^*_{n|n-1} h_n + \hat{k}\, R_{nom_n}) \tag{4.5.53}$$

$$\hat{k}' = \frac{a}{a+1}\,\hat{k} + \frac{1}{a+1}\,\frac{\hat{k}\,(z_n - h^T_n \hat{x}^*_{n|n-1})^2}{(h^T_n P^*_{n|n-1} h_n + \hat{k}\, R_{nom_n})} \tag{4.5.54}$$

$$a' = a + 1 \tag{4.5.55}$$

It can be seen that (4.5.51) and (4.5.52) are just the maximum likelihood filter equations, except that $\hat{k}$ is used in place of the unknown k. The state estimate and its "computed" covariance matrix are propagated between measurements using the relationships

$$\hat{x}^*_{n|n-1} = \Phi(n,n-1)\ \hat{x}^*_{n-1|n-1} \qquad (4.5.56)$$

$$P^*_{n|n-1} = \Phi(n,n-1)\ P^*_{n-1|n-1}\ \Phi^T(n,n-1) + Q_n \qquad (4.5.57)$$

It should be noted that unless $\hat{k} = k$, the "computed" covariance matrix $P^*_{n|n}$ does not accurately represent the covariance of the estimation error. Smith attempts to show that the estimator for k as given by (4.5.54) is an unbiased estimator but makes several unrecognized approximations in evaluating the expected value of k'. He first says that the expected value of the second term in (4.5.54) is given by

$$\frac{\varepsilon(\hat{k})\ \varepsilon[(z_n - h_n\hat{x}^*_{n|n-1})^2]}{h_n^T\ \varepsilon(P^*_{n|n-1})h_n + \varepsilon(\hat{k})\ R_{nom_n}} \qquad (4.5.58)$$

However, the expected value of a nonlinear function of the random variables $\hat{k}$, $P^*_{n|n-1}$, z_n, and $\hat{x}^*_{n|n-1}$ is not equal to the function evaluated at the expected values of these respective variables.

He then states that

$$\varepsilon[(z_n - h_n\hat{x}^*_{n|n-1})^2] = k\ R_{nom_n} + h_n^T\ \varepsilon(P^*_{n|n-1})h_n \qquad (4.5.59)$$

However, this is true only if on every trial $P^*_{n|n-1}$ is equal to the actual covariance of the state estimation error. This generally will be true only if $\hat{k} = k$ at every estimation stage.

The third approximation involved is in the computation of $\varepsilon(P^*_{n|n-1})$. He obtains this quantity recursively using the following equations.

$$\varepsilon(P^*_{n|n}) = \varepsilon(P^*_{n|n-1}) - \varepsilon(A^*_n)(h^T_n\ \varepsilon(P^*_{n|n-1})h_n + \varepsilon(\hat{k})R_{nom_n})\ \varepsilon(A^{*T}_n)$$

where $\varepsilon(A^*_n) = \varepsilon(P^*_{n|n-1})h_n/(h^T_n\ \varepsilon(P^*_{n|n-1})h_n + \varepsilon(\hat{k})\ R_{nom_n})$.

Here as before, Smith fails to realize that the expected value of a nonlinear function of a random variable is not equal to the function evaluated at the expected value of the random variable.

In testing the above theoretical results, Smith only simulates the equations for the mean of the estimate of k and the mean computed covariance matrix. This is unfortunate since many approximations were made in their derivation, namely the rather dubious use of the expectation operators above. So his results are somewhat open to question since he did not simulate the actual performance of the estimator of the state and the parameter k in a realistic situation.

Chapter 5

TESTING OF STATISTICAL HYPOTHESES

5.1 Introduction

In Chapters 3 and 4 techniques for estimating the state and noise variance parameters were discussed, and the necessary equations for the solution of the problem derived. As was seen, even in the simplest case, considerably more computation was needed for estimating the noise variance parameters as compared with estimation of the state alone. In those applications when estimation of the state is of primary importance, estimation of the noise parameters should not be undertaken unless there is reason to believe that the a priori estimates of these parameters are sufficiently in error to seriously affect the state estimation. The purpose of this chapter is to develop expressions and criteria which allow a decision to be made as to whether observed data are consistent with the assumptions about the values of the noise variance parameters. If it is concluded that the data are not consistent, then estimation of the parameters using the techniques of the previous chapters should be undertaken.

Testing of statistical hypotheses is an important part of statistical analysis but is perhaps one of the least understood and applied techniques in optimal estimation theory. Historically this is so because of a lack of a consistent theory which is generally applicable to a wide class of problems. But even today, long after the tools of hypothesis testing have been developed, often little use is made of such tools. This can result in major difficulties in applying optimal estimation theory to operational situations.

A statistical hypothesis is usually a statement about one or more population distributions, and specifically about one or more parameters of such population distributions. It is always a statement about the population, not about a finite sample taken from the population.

There are two types of hypotheses which are of interest, namely simple and composite hypotheses. Hypotheses that completely specify a population distribution are known as simple hypotheses. An example of such a hypothesis is: the population is normal with mean m_o and standard deviation σ_o, where m_o and σ_o are specified values. When the population is not determined completely by the hypothesis, the hypothesis is known as composite. An example of such a hypothesis is: the population is normal with mean m_o. Here the exact population distribution is not specified, since no requirement was put on σ, the population standard deviation.

Hypotheses may also be classified by whether they specify exact parameter values, or merely a range or interval of such values. For example, the hypothesis $m = m_o$ is an exact hypothesis, although $m \geq m_o$ is not exact.

Whatever procedure may be used for testing a hypothesis, that is, deciding on the basis of observed data whether to accept or reject the hypothesis, there are two possible errors involved: 1) rejecting the hypothesis when it is true, and 2) not rejecting the hypothesis when it is false. For any given situation, there may exist a family of different tests of the same hypothesis all of which give the same probability of rejecting the hypothesis when it is true but result in different probabilities of accepting the hypothesis when it is in fact false. It seems reasonable that the "best" test is the one which minimizes the probability of accepting a false hypothesis for a given probability of rejecting the true hypothesis.

All tests involve finding a test variable, or sample characteristic, which is a function of the observed data. One of the first problems to be faced in making a decision from the data is that of choosing the relevant and appropriate sample characteristic for the particular purpose. Different combinations of the sample data give different kinds and amounts of information about the population. Reaching a conclusion about some population characteristic requires effective use of the right information in the sample, and various sample characteristics differ in their relevance to different questions about the population.

Once the sample characteristic has been selected, a "critical region" of the test is defined such that if the characteristic lies within the critical region the hypothesis is accepted, and if it lies outside the critical region, the hypothesis is rejected.

Let S be the sample space of outcomes of an experiment and x denote an arbitrary element of S. Let H_o be the hypothesis being tested (called the null hypothesis), and let w denote the critical region. The probability of the first kind of error, rejecting H_o when it is true, is denoted by

$$P(x \text{ in } (S-w) | H_o) = \alpha \qquad (5.1.1)$$

where α is called the level of significance of the test. The probability of the second kind of error, accepting a false hypothesis, is denoted by

$$P(x \text{ in } w | H) = \beta(H) \qquad (5.1.2)$$

where H is a particular alternative hypothesis in the class of all possible alternative hypotheses. The function

$$\gamma(H) = 1 - \beta(H)$$

defined over all possible H is called the power function and for a particular value of H, is called the power of the test of H. The problem of statistical hypothesis testing is that of determining a critical region such that for a given level of significance, the power of the test is as large as possible.

The next sections of this chapter are devoted to discussion of certain sample characteristics and distributions upon which subsequent hypothesis tests are based.

5.2 Sampling Characteristics and Distributions

Let x be a random variable with probability density function f(x) and consider n independent repetitions of a random experiment to which x is attached. Performing the series of n repetitions, n observed values of x are obtained, denoted $x_1,..,x_n$. Any sample characteristic will be a function of the sample values, say $g(x_1,..,x_n)$ and accordingly the probability distribution of this latter variable will be called the sampling distribution of the characteristic $g(x_1,..,x_n)$.

The sample mean is defined by

$$\bar{x} \triangleq \frac{1}{n} \sum_{i=1}^{n} x_i \tag{5.2.1}$$

and the sample variance

$$s^2 \triangleq \frac{1}{n} \sum_{i=1}^{n} (x_i - \bar{x})^2 \tag{5.2.2}$$

Let the population characterictics of the random variable x be

$$\varepsilon(x_i) = m$$

$$\varepsilon[(x_i - m)^2] = \sigma^2$$

Then the expected value of the sample characteristic $\bar{x}$ is equal to the population characteristic, m. Moreover, the variance of $\bar{x}$ will be small for large values of n. Thus for a sufficiently large value of n, the sample mean $\bar{x}$ will be approximately equal to its expected value m. If m is unknown, $\bar{x}$ can be used as an estimate of m.

Consider the variance of the sample.

$$s^2 = \frac{1}{n} \sum_{i=1}^{n} (x_i - \bar{x})^2$$

$$\varepsilon(s^2) = \frac{n-1}{n} \sigma^2$$

Thus the expected value of the sampling characteristic s^2 is not equal to the population characteristic σ^2 but is equal to $((n-1)/n)\sigma^2$. This difference is insignificant for large n; but for moderate n, it will be preferable to consider the corrected sample variance

$$\frac{n}{n-1} s^2 = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x})^2$$

which has an expected value exactly equal to σ^2.

The variance of s^2 is given by the expression

$$\varepsilon[(s^2-\varepsilon(s^2))^2] = \frac{\mu_4-\mu_2^2}{n} - \frac{2(\mu_4-2\mu_2^2)}{n^2} + \frac{\mu_4-3\mu_2^2}{n^3} \tag{5.2.3}$$

where μ_2 and μ_4 are the second and fourth central moments of the distribution function of x. (Ref. 3, P. 183)

For large n, the variance of s^2 will be small and s^2 can be expected to agree approximately with the population variance since, as already pointed out, the expected value of s^2 is practically equal to σ^2 when n is large.

Thus far, the sample mean and variance and their first two moments were studied without reference to the density function of the random variable involved. In order to obtain more precise results about the properties of sampling distributions, it will be necessary to introduce further assumptions about f(x). The case of interest is when f(x) is a normal density function.

If x is an observation from a normal distribution with population mean m and variance σ^2, the probability density function of x is

$$f(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2}\left(\frac{x-m}{\sigma}\right)^2} \qquad (5.2.4)$$

It has been assumed that the n observations x_i are independent, so

$$f(x_1,..,x_n) = \prod_{i=1}^{n} f(x_i)$$

It can be shown that if the n observations x_i are independent normal random variables with population mean m and variance σ^2, then

1) $\bar{x}$ is a normal variable with mean m and variance σ^2/n.

2) $\frac{n\ s^2}{\sigma^2}$ is a central chi square distributed random variable with n-1 degrees of freedom.

3) $\bar{x}$ and s^2 are independently distributed.

4) $t = \sqrt{n-1}\,\frac{\bar{x} - m}{s}$ is a Students' distributed random variable with n-1 degrees of freedom.

5.3 Confidence Intervals

An understanding of confidence intervals is necessary before testing of simple hypotheses can be undertaken. When estimating the value of a parameter by observations on a random variable, it is usually desirable to obtain not only the estimated parameter value but also a measure of the precision of such an estimate. To obtain such a measure of an estimate $\hat{\xi}$ of an unknown parameter ξ, two positive numbers δ and α might be found such that the probability that the true value of ξ is included between the limits $\hat{\xi} \pm \delta$ is equal to $1 - \alpha$. Or

$$P(\hat{\xi} - \delta < \xi < \hat{\xi} + \delta) = 1 - \alpha \qquad (5.3.1)$$

For a given probability $1 - \alpha$, high precision of the estimate would be associated with small values of δ. More generally, to an unknown parameter ξ, two functions of the sample values ξ_1^* and ξ_2^* are found such that the probability that the interval (ξ_1^*, ξ_2^*) includes the true value ξ has a given value $1 - \alpha$, or

$$P(\xi_1^* < \xi < \xi_2^*) = 1 - \alpha \qquad (5.3.2)$$

Such an interval is called the confidence interval for the parameter and the probability $1 - \alpha$ is denoted as the confidence coefficient of the interval.

5.4 Tests on the Mean

Two situations will be treated here concerning confidence intervals and tests on the mean, one in which σ is presumed known, the other when σ is not known. The case when σ is known is considered first.

Let the variable x be normally distributed with mean m and variance σ^2, where m is unknown and σ is known precisely. Given n independent observed values $x_1, \ldots, x_n$, a confidence interval for the mean is sought.

In Section 5.2 it was stated that the variable

$$\bar{x} = \frac{1}{n} \sum_{i=1}^{n} x_i$$

has a normal distribution with mean m and variance σ^2/n. Therefore

$$t \triangleq \frac{\sqrt{n}\ (\bar{x} - m)}{\sigma} \qquad (5.4.1)$$

is a zero mean unit variance normally distributed variable.

Let α denote a given fraction and t_α be the α percent value of t found directly from a table of the normal distribution. By the definition of t_α,

$$P(-t_\alpha < t < t_\alpha) = 1 - \alpha \qquad (5.4.2)$$

By a simple transformation, (5.4.2) can be written as

$$P(\bar{x} - t_\alpha \frac{\sigma}{\sqrt{n}} < m < \bar{x} + t_\alpha \frac{\sigma}{\sqrt{n}}) = 1 - \alpha \qquad (5.4.3)$$

(5.4.3) is a relation of the type suggested by (5.3.2). Accordingly, the interval

$$(\bar{x} - t_\alpha \frac{\sigma}{\sqrt{n}}, \quad \bar{x} + t_\alpha \frac{\sigma}{\sqrt{n}}) \qquad (5.4.4)$$

is a $1 - \alpha$ confidence interval for m, the limits of the interval are confidence limits for m, and the corresponding confidence coefficient is $1 - \alpha$.

Thus the confidence interval (5.4.4) provides a rule for estimating the parameter m, which is associated with a constant risk of error equal to α, where α can be chosen arbitrarily.

Testing the hypothesis that m has some given value, say m_o, is related to the confidence interval deduced above. In this case a decision is made concerning which of the following hypotheses is true, based upon the observed data:

1) H_o: x is normal with mean $m = m_o$ and variance σ^2 (known)

2) H_1: x is normal with mean $m \neq m_o$ and variance σ^2 (known)

Working on a given level α, the confidence limits of m are computed according to (5.4.4). If the given value m_o falls outside the confidence interval, it is said that m differs significantly from m_o on the α level and accordingly H_o is rejected and H_1 is accepted. If the confidence interval includes the point m_o, it is said that no significant difference has been found and the hypothesis H_o is accepted.

In the case when H_o is in fact true, this test gives a probability $1 - \alpha$ of accepting the hypothesis and consequently the probability α of rejecting it. Thus the probability of committing an error by rejecting the hypothesis when it is true is equal to the level of the test α.

In order to apply this test, the sample characteristic $\bar{x}$ is found, and the quantity t computed, where

$$t = \frac{\sqrt{n}\ (\bar{x} - m_o)}{\sigma} \qquad (5.4.5)$$

Denoting by α the desired level of the test, the value t_α is found from a normal distribution table. If $|t| > t_\alpha$, the hypothesis H_o is rejected on the level α.

In the case when H_o is not true, but rather H_1 is true, the probability of accepting the incorrect hypothesis based upon the above test is not $1 - \alpha$. This is because (5.4.5) does not have a zero mean unit variance normal distribution when $m \neq m_o$. However, the variable

$$t' = \frac{\sqrt{n}\ (\bar{x} - m)}{\sigma} \qquad (5.4.6)$$

does have such a distribution. Define

$$\Delta = \frac{\sqrt{n}\ (m - m_o)}{\sigma} \tag{5.4.7}$$

Then $$t = t' + \Delta$$

The probability that the test variable t lies within the range $|t| < t_\alpha$ is

$$P(-t_\alpha < t < t_\alpha) = P(-t_\alpha - \Delta < t' < t_\alpha - \Delta)$$

Define $\beta = P(-t_\alpha < t < t_\alpha)$ = probability of accepting H_o when H_1 is true

Then $$\beta = \int_{-t_\alpha - \Delta}^{t_\alpha - \Delta} f(u)\ du$$

where f(u) is the normal probability density function with zero mean and unit variance. Since f(u) is symmetrical about u = 0,

$$\beta = \frac{1}{2}\left[\int_{-(t_\alpha+\Delta)}^{t_\alpha+\Delta} f(u)\ du + \int_{-(t_\alpha-\Delta)}^{t_\alpha-\Delta} f(u)\ du\right]$$

Define $t_{\beta_1} = t_\alpha - \Delta,\qquad t_{\beta_2} = t_\alpha + \Delta$

$$\beta_1 = \int_{-t_{\beta_1}}^{t_{\beta_1}} f(u)\ du, \qquad \beta_2 = \int_{-t_{\beta_2}}^{t_{\beta_2}} f(u)\ du$$

Then $$\beta = \tfrac{1}{2}(\beta_1 + \beta_2)$$

and $$\gamma = 1 - \beta = 1 - \tfrac{1}{2}(\beta_1 + \beta_2)$$

For $-t_\alpha < \Delta < t_\alpha$, both β_1 and β_2 will be positive, while if $\Delta > t_\alpha$, β_1 will be negative, and if $\Delta < -t_\alpha$, β_2 will be negative. It can be seen that

$\beta_1 = \pm P(|t| < |t_{\beta_1}|)$ where the + sign is used when t_{β_1} is positive

$\beta_2 = \pm P(|t| < |t_{\beta_2}|)$ with the same sign convention

Thus the power of the test, γ, is a function of m, m_o, α, σ, and n. However, if γ is plotted as a function of the nondimensional parameter Δ, the only free variable is α, the level of the test. Such a plot is shown in Figure 5.1 for α = .05, .10, and .20.

Note that

$$\beta = \int_{-t_\alpha}^{t_\alpha} f(u)\,du + \int_{-t_\alpha-\Delta}^{-t_\alpha} f(u)\,du + \int_{t_\alpha}^{t_\alpha-\Delta} f(u)\,du$$

and $$1 - \alpha = \int_{-t_\alpha}^{t_\alpha} f(u)\,du$$

so $$\beta = 1 - \alpha + \int_{-t_\alpha-\Delta}^{-t_\alpha} f(u)\,du + \int_{t_\alpha}^{t_\alpha-\Delta} f(u)\,du$$

It can be shown that for any $\Delta \neq 0$,

$$\int_{-t_\alpha-\Delta}^{-t_\alpha} f(u)\,du + \int_{t_\alpha}^{t_\alpha-\Delta} f(u)\,du < 0$$

so $$\beta < 1 - \alpha \qquad \text{for } \Delta \neq 0$$

$$= 1 - \alpha \qquad \text{for } \Delta = 0$$

In other words, the probability of accepting a false hypothesis H_1 is less than the probability of accepting a true hypothesis H_o. As $|\Delta|$ increases, β decreases with the limits $\beta = 1 - \alpha$ for $\Delta = 0$ and $\beta \to 0$ as $|\Delta| \to \infty$.

Since β will ordinarily be small for large α, it follows that setting α larger will make for relatively more powerful tests of H_o. The power curves shown in Figure 5.1 indicate that if α is set at .10 rather than .05, the test with $\alpha = .10$ is more powerful than that for $\alpha = .05$ over all possible values of m under H_1. Making the probability of error in rejecting a true hypothesis larger has the effect of making the test more powerful. The proper value of α for any particular application depends upon the relative penalty paid for overlooking a true departure from H_o versus rejecting H_o falsely.

For a given m, m_o, and σ, increasing the sample size, n, has the effect of increasing $|\Delta|$, so that the power of the test is increased with increasing n. A similar increase in the test power could be achieved by reducing σ; but in the application under study, σ is not a variable which is easily reduced, so that the only effective way to increase the test power is to increase α or n.

Establishing confidence intervals in tests on the mean m with σ unknown is similar to the previous work, except that the sampling characteristic and its distribution are somewhat different.

In Section 5.2, it was stated that the variable

$$t = \frac{\sqrt{n-1}\ (\bar{x} - m)}{s} \tag{5.4.8}$$

has a Students' distribution with n-1 degrees of freedom, where $\bar{x}$ is the sample mean and s is the sample standard deviation. Let α denote a given fraction and t_α be the α percent value of t for n-1 degrees of freedom found directly from a table of the Students' distribution. By the definition of t_α,

$$P(-t_\alpha < \frac{\sqrt{n-1}\ (\bar{x} - m)}{s} < t_\alpha) = 1 - \alpha \tag{5.4.9}$$

In the same fashion as before, the interval

$$(\bar{x} - t_\alpha \frac{s}{\sqrt{n-1}},\ \ \bar{x} + t_\alpha \frac{s}{\sqrt{n-1}}) \tag{5.4.10}$$

is a 1 - α confidence interval for m, the limits of the interval are confidence limits for m, and the corresponding confidence coefficient is 1 - α.

Testing the hypothesis that m has some given value, say m_o, for σ unknown is quite similar to the test for σ known. A decision is made concerning which of the following hypotheses is true based upon the observed data:

1) H_o: x is normal with mean $m = m_o$
2) H_1: x is normal with mean $m \neq m_o$

In order to apply this test, the sample characteristics $\bar{x}$ and s^2 are found and the quantity t computed, where

$$t = \frac{\sqrt{n-1}\ (\bar{x} - m_o)}{s} \tag{5.4.11}$$

Denoting by α the desired level of the test, the value t_α is found from a Students' distribution table. If $|t| > t_\alpha$, the hypothesis H_o is rejected on the level α.

Define

$$t' \triangleq \frac{\sqrt{n-1}\ (\bar{x} - m)}{s} \tag{5.4.12}$$

and
$$\Delta \triangleq \frac{\sqrt{n-1}\ (m - m_o)}{s} \tag{5.4.13}$$

Then
$$t = t' + \Delta$$

If H_o is false (H_1 is true), t defined by (5.4.11) does not have a Students' distribution, but t' does have such a distribution. So

$$P(-t_\alpha < t < t_\alpha) = P(-t_\alpha - \Delta < t' < t_\alpha - \Delta)$$

Define $\beta = P(-t_\alpha < t < t_\alpha)$ = probability of accepting H_o when H_1 is true

Then $$\beta = \frac{1}{2}(\beta_1 + \beta_2)$$

where β_1 and β_2 are as defined before except f(u) appearing there is now the Students' distribution with n-1 degrees of freedom.

As before, it is possible to construct a power curve versus Δ for a given α. Now, however, the curve is also a function of n, the number of sample values.

As in the case with σ known, $\beta < 1 - \alpha$ for any $|\Delta| > 0$, so that the probability of accepting a false hypothesis is less than the probability of accepting a true hypothesis, with $\beta \rightarrow 0$ as $|\Delta| \rightarrow \infty$. Increasing α results in a more powerful test of H_o but also increases the risk of rejecting a true hypothesis.

Figure 5.2 shows the power of the above test versus the nondimensional parameter Δ for α = .05 and .10, with n = 10. Figure 5.3 shows the power versus Δ for n = 5, 10, 20, with α = .10.

In this section, the normal distribution and the Students' statistic have been used for drawing inferences on the unknown mean of a population from which observations are obtained. The distribution of the t-statistic defined by (5.4.11) is obtained after making the following assumptions:

1) the distribution of the random variable x is normal

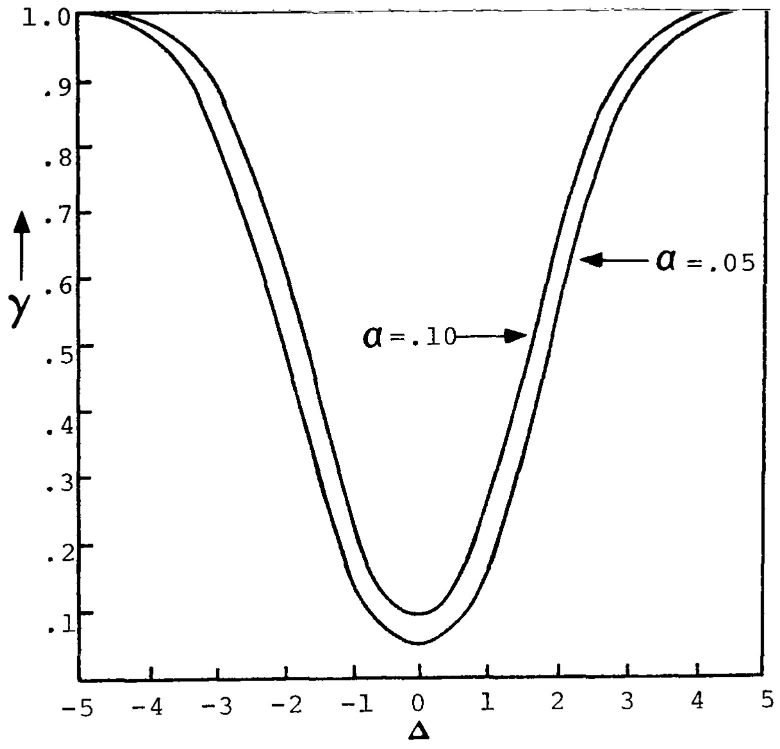

Fig. 5.1 Test Power vs Δ and α

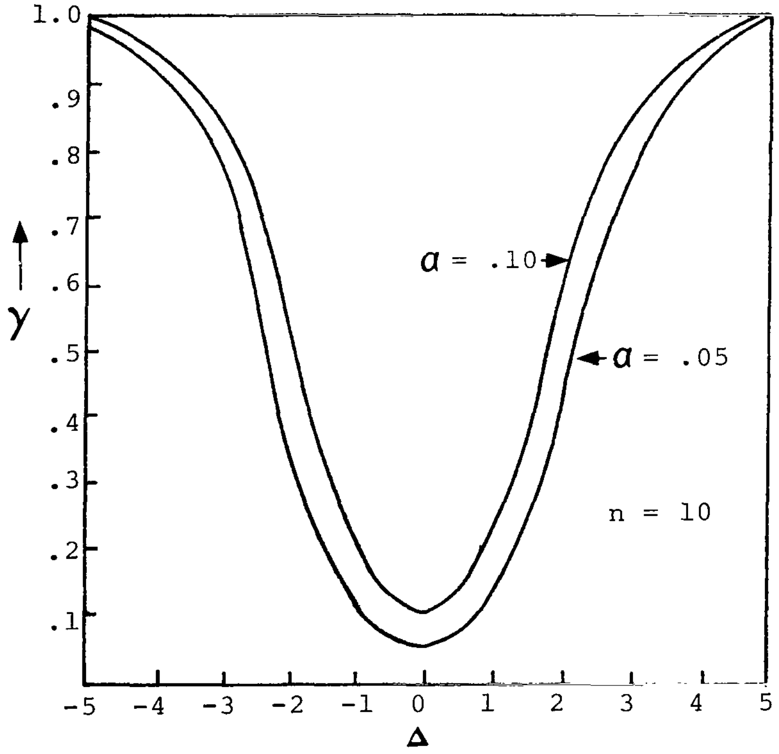

Fig. 5.2 Test Power vs Δ and α for Fixed n

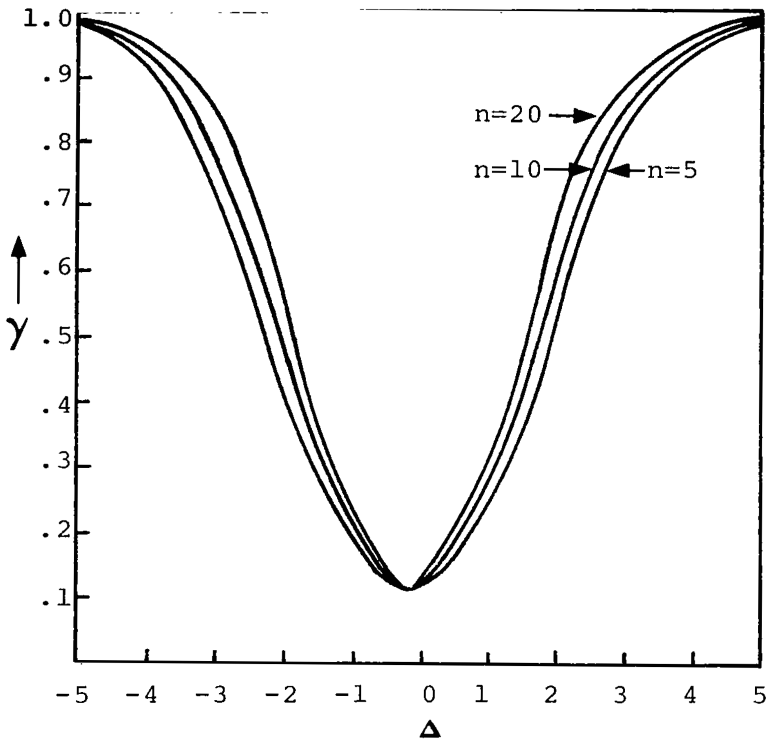

Figure 5.3 Test Power vs Δ and n for Fixed α

2) the observations are mutually independent

3) the mean of the population is exactly m_o

From the theoretical and empirical studies it is known that the t distribution is not sensitive to moderate departures from normality so that its application is not strictly governed by the normality assumptions. A significant t may not, therefore, be interpreted as indicating departure from the normality of the observations.

Suppose that all the observations are mutually correlated with a common positive correlation ρ for any two. Then

$$\varepsilon[(\bar{x} - m_o)^2] = \frac{\sigma^2}{n}(1 + (n-1)\rho) \qquad (5.4.14)$$

$$\varepsilon(s^2) = \frac{n-1}{n}\sigma^2(1 - \rho) \qquad (5.4.15)$$

Instead of the t-statistic (5.4.11) consider

$$t^2 = \frac{(n-1)(\bar{x} - m_o)^2}{s^2} \qquad (5.4.16)$$

which can be shown to have a F distribution on 1 and n-1 degrees of freedom. From (5.4.14) and (5.4.15), the expected values of the numerator and denominator of t^2 are

$$\frac{n-1}{n}\sigma^2(1 - (n-1)\rho) \text{ and } \frac{n-1}{n}\sigma^2(1 - \rho) \qquad (5.4.17)$$

The ratio of the expectations is unity when $\rho = 0$, but is greater than unity when $\rho > 0$ and $\to \infty$ as $\rho \to 1$. Thus a large value of t is expected to occur when ρ is positively large, even when m_o is exactly equal to m. A significant t may therefore be due to a departure in assumption 2).

Finally, when the assumptions 1) and 2) are true and $m \neq m_o$, the ratio of the expected values of the numerator and denominator of t^2 of (5.4.16) is

$$\frac{n(m - m_o)^2}{\sigma^2} + 1 \qquad (5.4.18)$$

compared with 1 when $m = m_o$, so that large values of t occur when assumption 3) is wrong. This is exactly the reason why the t-test is used to test the null hypothesis concerning the mean of a distribution.

In computing (5.4.17) the extreme case of mutual dependence with a common correlation ρ was considered. But in general, any dependence giving positive correlation to pairs of variables will increase the significance of t, so that the test will indicate any significant departure from assumption 2).

5.5 Tests on the Variance

Let the variable x be normal with mean m and variance σ^2, where m and σ are both unknown. Given n independent observed values $x_1,..,x_n$, a confidence interval for the variance σ^2 is sought. In Section 5.2 it was stated that the variable

$$\chi^2 = \frac{n\ s^2}{\sigma^2}$$

has a chi square distribution with n-1 degrees of freedom. For any given level of test α, infinitely many intervals can be found, each of which contains exactly the area $1 - \alpha$ in this distribtuion. Among all these intervals, the particular interval $(\chi^2_{\alpha_1}, \chi^2_{\alpha_2})$ is chosen, where $\chi^2_{\alpha_1}$ and $\chi^2_{\alpha_2}$ are the α_1 and α_2 values of the χ^2 distribution for n-1 degrees of freedom, where

$$\alpha_1 = 1 - \tfrac{1}{2}\alpha, \quad \alpha_2 = \tfrac{1}{2}\alpha$$

Each of the tails $\chi^2 < \chi^2_{\alpha_1}$ and $\chi^2 > \chi^2_{\alpha_2}$ contain equal area $\frac{1}{2}\alpha$, and thus

$$P(\chi^2_{\alpha_1} < \frac{ns^2}{\sigma^2} < \chi^2_{\alpha_2}) = 1 - \alpha \qquad (5.5.1)$$

By a simple transformation, (5.5.1) can be written as

$$P\left(\frac{ns^2}{\chi^2_{\alpha_2}} < \sigma^2 < \frac{ns^2}{\chi^2_{\alpha_1}}\right) = 1 - \alpha \qquad (5.5.2)$$

Thus the interval

$$\left(\frac{ns^2}{\chi^2_{\alpha_2}}, \qquad \frac{ns^2}{\chi^2_{\alpha_1}}\right) \qquad (5.5.3)$$

is a $1 - \alpha$ confidence interval for σ^2, the limits of the interval are the confidence limits for σ^2, and the corresponding confidence coefficient is $1 - \alpha$. The confidence interval (5.5.3) provides a rule for estimating the parameter σ^2, which is associated with a constant risk of error equal to α.

Testing the hypothesis that σ^2 has some given value, say σ_o^2, is analogous to the tests of the mean given in the previous section. In this case a decision is made concerning which of the following hypotheses is true:

1) H_o: x is normal with variance $\sigma^2 = \sigma_o^2$
2) H_1: x is normal with variance $\sigma^2 \neq \sigma_o^2$

In order to apply the test, the sample characteristic s^2 must be found and the quantity χ^2 computed, where

$$\chi^2 = \frac{n\ s^2}{\sigma_o^2} \tag{5.5.4}$$

Denoting by α the desired level of the test, the values $\chi^2_{\alpha_1}$ and $\chi^2_{\alpha_2}$ are found from a chi square distribution table with n-1 degrees of freedom. If $\chi^2_{\alpha_1} < \chi^2 < \chi^2_{\alpha_2}$, the hypothesis H_o is accepted on the level , otherwise H_o is rejected and H_1 accepted. In the case when H_o is in fact true, this test gives a probability of 1 - α of accepting the hypothesis and consequently a probability α of rejecting. Thus the probability of rejecting H_o when it is true is equal to the level of the test, α.

In the case when H_o is not true (thus H_1 is true), the probability of accepting the incorrect H_o is not 1 - α. This is because (5.5.4) does not have a chi square distribution when $\sigma \neq \sigma_o$. However, the variable

$$\chi'^2 = \frac{n\ s^2}{\sigma^2} \tag{5.5.6}$$

does have a chi square distribution with n-1 degrees of freedom. Define

$$\eta \triangleq \frac{\sigma_o^2}{\sigma^2} \tag{5.5.7}$$

Then $$\chi'^2 = \eta\chi^2$$

and the probability that the test variable lies within the range $\chi^2_{\alpha_1} < \chi^2 < \chi^2_{\alpha_2}$ is

$$P(\chi^2_{\alpha_1} < \chi^2 < \chi^2_{\alpha_2}) = P(\eta\chi^2_{\alpha_1} < \chi'^2 < \eta\chi^2_{\alpha_2})$$

Define $\beta = P(\chi^2_{\alpha_1} < \chi^2 < \chi^2_{\alpha_2})$ = probability of accepting H_o when H_1 is true

Then $$\beta = \int_{\eta\chi^2_{\alpha_1}}^{\eta\chi^2_{\alpha_2}} f(u)\, du$$

where f(u) is the chi square distribution with n-1 degrees of freedom. It should be noted that unless $\eta = 1$, the area under the tails of f(u) for $u < \eta\chi^2_{\alpha_1}$ and $u > \eta\chi^2_{\alpha_2}$ are not equal. Define

$$\beta_1 = \int_0^{\eta\chi^2_{\alpha_1}} f(u)\, du = P(\chi'^2 < \eta\chi^2_{\alpha_1})$$

$$\beta_2 = \int_0^{\eta\chi^2_{\alpha_2}} f(u)\, du = P(\chi'^2 < \eta\chi^2_{\alpha_2})$$

Then $$\beta = \beta_2 - \beta_1$$

and $$\gamma = 1 - \beta$$

It is again possible to construct a power curve versus η for a given α, the curve also being a function of n, the number of sample values. Such curves are shown in Figures 5.4 and 5.5.

5.6 Multidimensional Hypothesis Tests with Time Varying Population Parameters

In the preceding sections, hypothesis tests on the time invariant parameters of the distribution of a scalar random variable were discussed. The results can be generalized to include tests on vector random variables with time varying parameters. First the case of vector random variables with constant population parameters will be discussed.

Let X be a r x 1 random variable with density function f(X) and consider n independent repetitions of a random experiment to which X is attached. The resulting observed values of X are denoted $X_1,..,X_n$. The sample mean is defined by

$$\overline{X} = \frac{1}{n} \sum_{i=1}^{n} X_i \qquad (5.6.1)$$

and the sample covariance is defined by

$$s^2 = \frac{1}{n} \sum_{i=1}^{n} (X_i - \overline{X})(X_i - \overline{X})^T \qquad (5.6.2)$$

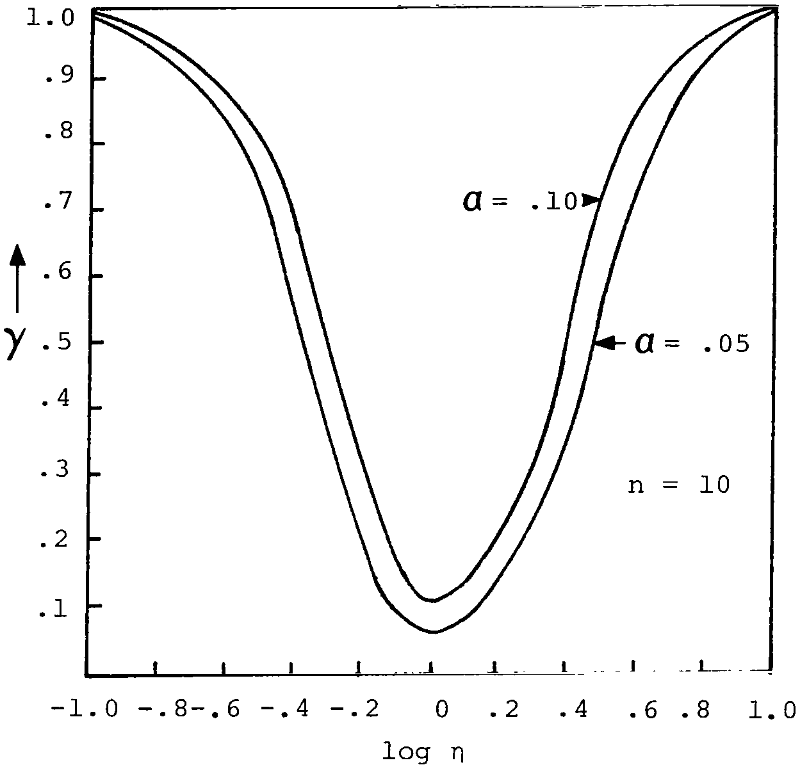

Fig. 5.4 Test Power vs η and α for Fixed n

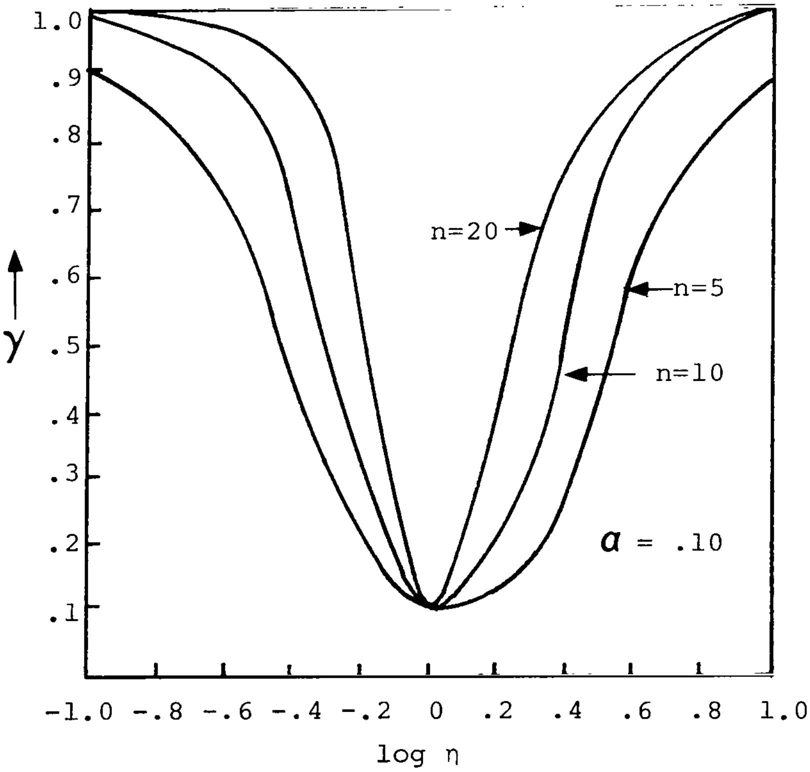

Fig. 5.5 Test Power vs η and n for Fixed α

Let the population characteristics of the random variable X be

$$\varepsilon(X_i) = M \qquad \text{for } i = 1,\ldots,n$$

$$\varepsilon[(X_i - M)(X_i - M)^T] = P$$

Then the sample mean is a random variable with

$$\varepsilon(\overline{X}) = M$$

$$\varepsilon[(\overline{X} - M)(\overline{X} - M)^T] = \frac{1}{n} P$$

and the sample covariance is a random variable with

$$\varepsilon(s^2) = \frac{n-1}{n} P$$

As in the preceding sections, it will be necessary to introduce further assumptions about f(X) in order to obtain more precise results about the properties of the sampling distributions. The case of interest is when f(X) is a multidimensional normal distribution.

$$f(X) = \frac{1}{(2\pi)^{r/2}|P|^{1/2}} e^{-\frac{1}{2}[(X-M)^T P^{-1}(X-M)]}$$

It is also assumed that the n observations X_i are. independent, so

$$f(X_1,..,X_n) = \prod_{i=1}^{n} f(X_i)$$

It can be shown that if the n observations X_i are independent normal variables with population mean M and covariance P, then

1) $\overline{X}$ is a r x 1 dimensional normal variable with mean M and covariance P/n.

2) for <u>any</u> fixed vector L, $n \frac{L^T S^2 L}{L^T P\ L}$ is a central chi square distributed random variable with n-1 degrees of freedom.

3) $\overline{X}$ and S^2 are independently distributed.

4) $t \triangleq \frac{\sqrt{n-1}\ L^T(\overline{X} - M)}{\sqrt{L^T S^2 L}}$ for any fixed L, is a Students' distributed random variable with n-1 degrees of freedom.

Comparing these four results with those of Section 5.2, tests of hypotheses and confidence intervals on a vector random variable can be handled in the same fashion as a scalar random variable. If the mean and variance of each component of the random variable X are to be tested, the proper choice of L for each test is

$$L_j = \begin{bmatrix} 0 \\ \cdot \\ \cdot \\ 0 \\ 1 \\ 0 \\ \cdot \\ 0 \end{bmatrix} \leftarrow j^{th} \text{ compon}\_\text{nt not zero}$$

Instead of a single confidence interval and test of the mean and variance, there will now be r such intervals and r tests on the mean and variance. The power of such tests under deviations from the null hypothesis can be computed in a manner entirely analogous to that of the previous sections.

The sample mean and covariance are not the only sample characteristics which can be used to test hypotheses on the distribution of X. Below are discussed alternative sample characteristics which might be used and in many applications they will provide sufficiently powerful tests.

Consider the random variable

$$Y_i \triangleq C(X_i - M)$$

where $$C \triangleq \sqrt{P^{-1}}$$

such that $C^T C = P^{-1}$ and C^{-1} exists.

Such a C can always be found because P is positive definite. Then

$$\varepsilon(Y_i) = 0$$

$$\varepsilon(Y_i Y_i^T) = I \qquad \text{the identity matrix}$$

The elements of Y_i are independent, zero mean unit variance normally distributed variables, and

$$U_i \triangleq \frac{1}{r} \sum_{j=1}^{r} Y_i^j$$ where Y_i^j is the j^{th} element of the vector Y_i

is a zero mean normal variable with variance $\frac{1}{r}$. Since Y_i is independent of Y_k for $i \neq k$, U_i is independent of U_k for $i \neq k$, and

$$\overline{U} \triangleq \frac{1}{n} \sum_{i=1}^{n} U_i$$

is a zero mean normal variable with variance $\frac{1}{n\ r}$.

Define

$$W_i \triangleq Y_i^T Y_i \tag{5.6.3}$$

Since the Y_i^j are zero mean unit variance normal variables, W_i is a central chi square variable with r degrees of freedom, with W_i independent of W_k for $i \neq k$. Then

$$Z \triangleq \sum_{i=1}^{n} W_i \tag{5.6.4}$$

is a central chi square variable with n r degrees of freedom.

Now consider the variable

$$Y_i' = C(X_i - \overline{X}) = Y_i + C(M - \overline{X})$$

and define

$$W_i' = Y_i'^T Y_i'$$

$$= Y_i^T Y_i + (\overline{X}-M)^T P^{-1} (\overline{X}-M) - 2\, Y_i^T C(\overline{X}-M)$$

$$\overline{Y} = \frac{1}{n} \sum_{i=1}^{n} Y_i = C(\overline{X} - M)$$

Then $$W_i' = Y_i^T Y_i + Y^T \overline{Y} - 2\, Y_i^T \overline{Y} \qquad (5.6.5)$$

Define $$Z' = \sum_{i=1}^{n} W_i' = \sum_{i=1}^{n} (Y_i^T Y_i - \overline{Y}^T \overline{Y}) \qquad (5.6.6)$$

It can be shown that Z' is a central chi square variable with (n-1)r degrees of freedom, and that Z' is independent of $\overline{U}$.

Since n r $\overline{U}$ is a zero mean unit variance normal variable and Z' is a central chi square variable with (n-1)r degrees of freedom, and $\overline{U}$ is independent of Z', the variable

$$t = r \frac{\sqrt{n(n-1)}\ \overline{U}}{\sqrt{Z'}}$$

is a Students' distributed random variable with (n-1)r degrees of freedom. Define

$$s^2 \triangleq \frac{1}{n\ r} Z' = \frac{1}{n\ r} \sum_{i=1}^{n} (Y_i^T Y_i - \overline{Y}^T \overline{Y}) \qquad (5.6.7)$$

Then $$t = \frac{\sqrt{(n-1)r}\ \overline{U}}{s} \qquad (5.6.8)$$

After some manipulation, it can be shown that

$$s^2 = \frac{1}{n\ r} \sum_{i=1}^{n} (X_i - \overline{X})^T P^{-1} (X_i - \overline{X}) \qquad (5.6.9)$$

By use of the sampling characteristics (5.6.8) and (5.6.9), the null hypothesis that $M = M_o$ and $P = P_o$ can be tested. The proper test variables for this test are

$$t = \frac{\sqrt{(n-1)r}\ \overline{U}}{s} \qquad (5.6.10)$$

and

$$nr\ s^2 = \sum_{i=1}^{n} (X_i - \overline{X})^T P_o^{-1} (X_i - \overline{X}) \qquad (5.6.11)$$

where

$$\overline{U} = \frac{1}{r} \sum_{j=1}^{r} \overline{Y}^j$$

$$\overline{Y} = C_o(\overline{X} - M_o), \qquad C_o = \sqrt{P_o^{-1}}$$

Under the null hypothesis, it has been shown that t has a Students' distribution and nr s^2 has a chi square distribution. It should be noted that unlike the tests of hypotheses about the distribution parameters of scalar normal variables, a mean test using (5.6.10) does depend upon the hypothesized value of the covariance P_o. Unless $M = M_o$ <u>and</u> $P = P_o$, t does not have a Students' distribution, and a significant t could arise from a departure from the hypothesis $M = M_o$ or $P = P_o$ or both. However, it can be shown that t is not highly sensitive to departures from the hypothesis $P = P_o$, so that a significant t can be used to reject the hypothesis $M = M_o$ alone, especially if the covariance test either accepts or does not strongly reject the hypothesis $P = P_o$. While the mean test depends somewhat upon the covariance hypothesis, it can be seen that the covariance test does not depend upon the mean hypothesis. The covariance test variable (5.6.11) has a chi square distribution if $P = P_o$, regardless of whether $M = M_o$.

The mean and covariance test variables (5.6.10) and (5.6.11) can be used to test the hypotheses outlined above in a fashion similar to that of Sections 5.4 and 5.5, as long as caution is employed in interpreting the results of such tests.

Now consider the case of vector random variables with time varying population parameters. The case of interest is when the population mean is time invariant, but the population covariance varies with time. Then the population characteristics of the random variable X are

$$\varepsilon(X_i) = M$$

$$\varepsilon[(X_i - M)(X_i - M)^T] = P_i$$

The sample mean is a random variable with

$$\varepsilon(\overline{X}) = M$$

$$\varepsilon[(\overline{X} - M)(\overline{X} - M)^T] = \frac{1}{n}\,\overline{P}$$

where $$\overline{P} \triangleq \frac{1}{n}\sum_{i=1}^{n} P_i$$

and the sample covariance is a random variable with

$$\varepsilon(S^2) = \frac{n-1}{n} \overline{P}$$

As before it will be assumed that the X_i are independent normal variables with the population parameters given above.

Because of the time varying population parameters, it will be necessary to utilize normalized variables in order to obtain the sampling distribution of certain sampling characteristics. Consider the variable

$$Y_i = C_i (X_i - M) \tag{5.6.12}$$

where $$C_i = \sqrt{P_i^{-1}}$$

such that $$C_i^T C_i = P_i^{-1} \text{ and } C_i^{-1} \text{ exists.}$$

Then $$\varepsilon(Y^i) = 0$$

$$\varepsilon(Y_i Y_i^T) = I$$

The elements of Y_i are independent zero mean unit variance normal variables, with Y_i independent of Y_j for $i \neq j$.

Define

$$\overline{Y} = \frac{1}{n} \sum_{i=1}^{n} Y_i$$

$$S'^2 = \frac{1}{n} \sum_{i=1}^{n} (Y_i - \overline{Y})(Y_i - \overline{Y})^T$$

After some manipulation, it can be shown that

$$S'^2 = \frac{1}{n} \sum_{i=1}^{n} [(C_i X_i - \overline{CX}) + M(\overline{C} - C_i)][(C_i X_i - \overline{CX}) + M(\overline{C} - C_i)]^T \qquad (5.6.13)$$

where

$$\overline{C} \triangleq \frac{1}{n} \sum_{i=1}^{n} C_i$$

$$\overline{CX} \triangleq \frac{1}{n} \sum_{i=1}^{n} C_i X_i$$

It can be shown that

1) $\overline{Y}$ is a r dimensional zero mean normal variable with covariance I/n.

2) for any fixed L, $\frac{n\, L^T S'^2 L}{L^T L}$ is a central chi square distributed variable with n-1 degrees of freedom.

3) $\overline{Y}$ and S'^2 are independently distributed.

4) for any fixed vector L, $\frac{\sqrt{n-1}\; L^T \overline{Y}}{\sqrt{L^T S'^2 L}}$ is a Students' distributed variable with n-1 degrees of freedom.

The hypothesis that $M = M_o$ and $P_i = P_{oi}$ can be tested in a fashion analogous to the tests outlined in this section for time invariant distribution parameters, except that now the test variables are

$$t_j = \frac{\sqrt{n-1} \quad L_j^T \overline{Y}}{\sqrt{L_j^T S'^2 L_j}} \qquad j=1,..,r \qquad (5.6.14)$$

$$\chi_j^2 = \frac{n\, L_j^T S'^2 L_j}{L_j^T L_j} \tag{5.6.15}$$

where $$S'^2 = \frac{1}{n} \sum_{i=1}^{n} (Y_i - \overline{Y})(Y_i - \overline{Y})^T$$

$$\overline{Y} = \frac{1}{n} \sum_{i=1}^{n} C_{oi}(X_i - M_o)$$

$$C_{oi} = \sqrt{P_{oi}^{-1}}$$

It can be seen that both (5.6.14) and (5.6.15) are functions of the sample values X_i and the values of M_o and and P_{io} unless P_{io} is time invariant. Therefore tests using these test variables do not provide independent tests of the mean and covariance. However, both variables can be used for testing the hypothesis $M = M_o$ and $P_i = P_{io}$, where M_o and P_{io} are specified values. Rejection of the hypothesis by either test can imply that $M \neq M_o$ or $P_i \neq P_{io}$ or both. However, even though the tests are not independent, it can be shown that the mean test is more sensitive to departures from the mean hypothesis than from the covariance hypothesis, and conversely for the covariance test.

As in the case of time invariant population parameters, the sample characteristic (5.6.14) and (5.6.15) are not the only characteristics which might be used to test hypotheses on the mean and covariance of X. In a manner analogous to the previous work, define

$$s^2 = \frac{1}{n\ r} \sum_{i=1}^{n} (Y_i^T Y_i - \overline{Y}^T \overline{Y}) \qquad (5.6.16)$$

$$t = \frac{\sqrt{(n-1)r}\ \overline{U}}{s} \qquad (5.6.17)$$

where here $Y_i = C_i (X_i - M)$

$$\overline{Y} = \frac{1}{n} \sum_{i=1}^{n} Y_i$$

$$U_i = \frac{1}{r} \sum_{j=1}^{r} Y_i^j$$

$$\overline{U} = \frac{1}{n} \sum_{i=1}^{n} U_i$$

As before, nr s^2 has a chi square distribution with n-1 degrees of freedom, and t has a Students' distribution with n-1 degrees of freedom. By use of these sampling characteristics, the null hypothesis $M = M_o$ and $P_i = P_{io}$ can be tested. The proper test variables for this test are (5.6.16) and (5.6.17) with M_o replacing M and P_{io} replacing P_i. The two tests are not independent tests of the mean and covariance so caution should be employed in interpreting the results of such tests.

5.7 Application of Hypothesis Tests to Maximum Likelihood State Estimation

In Chapter 2, the recursive maximum likelihood state estimation equations were derived for a linear dynamical system. These equations were derived under the following assumptions:

1) the measurement and driving noises are independent, zero mean normal variables.

2) the covariance of the noises are known precisely.

3) no computational errors are made.

4) all of the parameters describing the dynamical system and the linear measurement are known precisely.

If all of these assumptions are valid, it can be shown that the measurement residual at any time k

$$\Delta z_k \triangleq z_k - H_k \hat{x}_{k|k-1}$$

is a zero mean normal variable, independent of the residuals at times other than k, with conditional covariance

$$B_k = R_k + H_k P_{k|k-1} H_k^T$$

where Δz_k and B_k are computed using values of the noise covariance parameters assumed known by the previous assumptions.

It can be seen that the variable Δz_k is just such a variable upon which the tests of the mean and covariance given previously can be applied. Which set of tests is applied depends upon:

1) computational limitations

2) desired power of the tests

3) the dimension of the residual

4) whether the estimation equations have reached a steady state such that B_k is approximately constant, so that the population parameters of the residuals are time invariant

5) the need to isolate which component of the residual satisfies or violates the underlying assumptions

If the residuals fail the hypothesis tests, the tests themselves do not tell why, but merely indicate that one or more of the assumptions is probably in error. It is up to the analyst to isolate which of the assumptions is likely to be in error and make adjustments in the assumptions until the residuals pass the required hypothesis tests.

If all four of the previous assumptions are considered as the null hypothesis to be tested, it is most difficult to compute the power of the tests under deviations from the null hypothesis. In order to compute the test power, the distribution of the sample characteristics under deviations from the null hypothesis must be found. This is very difficult to do for very general deviations from the null hypothesis. Only when possible deviations from the null hypothesis are relatively simple, say errors in the covariances of the noises, can power of the test be computed. Even then, when the residuals are vector valued with time varying conditional covariance, the computation of the test power is most difficult. However, even if the test power is not known accurately, it can be expected that the tests will indicate significant deviations from the null hypothesis, which is the primary purpose of such tests.

The measurement residual is not the only observable random variable upon which hypothesis tests can be based. Consider the situation when the values of R and Q used to compute the state estimation weighting matrices may be in error. It is desired to test a hypothesis concerning the values of these parameters. In Chapter 3, it was shown that the score

$$S_n(Z_n,\alpha) = \frac{\partial L_n(Z_n,\alpha)}{\partial \alpha}$$

evaluated at the true value of the parameters α is asymptotically normal with zero mean and covariance $J_n(\alpha_o)$, where α_o is the true value of α. In the case of state and noise variance estimation, the parameter set α consists of the state x_n and the vector $\xi^T = (R^{11},..,R^{\gamma\gamma},Q^{11},..,Q^{\eta\eta})$. If the score is computed as a function of the measurements and the a priori values of R and Q, a large score will indicate that the a priori values of R and Q are probably in error.

Only those components of the score corresponding to differentiation by ξ are useful in testing the hypothesis on R and Q because it was seen that

$$S_n^x \triangleq \frac{\partial L_n(Z_n,\alpha)}{\partial x_n} \tag{5.7.1}$$

evaluated at the estimated value of x_n is identically zero, regardless of the true value of R and Q. However, the quantity

$$S_n^{\xi} \triangleq \frac{\partial L_n(Z_n,\alpha)}{\partial \xi} \qquad (5.7.2)$$

is a useful indicator for testing the hypothesis. If the null hypothesis $R = \hat{R}_o$ and $Q = \hat{Q}_o$ is true, then S_n^{ξ} is asymptotically a zero mean normal variable with covariance

$$\begin{aligned} \operatorname{cov}(S_n^{\xi}) &= \varepsilon\left[\left(\frac{\partial L_n}{\partial \xi}\right)^T \frac{\partial L_n}{\partial \xi} \,|x_n,\xi\right] \\ &\triangleq J_n^{\xi}\ (\alpha) \end{aligned}$$

There are two functions of the score (5.7.2) which might be used for hypothesis testing. Define

$$t = C_n\ S_n^{\xi}\ (Z_n,\hat{\xi})$$

where $$C_n = \sqrt{J_n^{\xi}(\hat{\xi})^{-1}}$$

Then each component of the vector t is asymptotically an independent zero mean unit variance normal variable. Tests on t can be conducted using the results of tests on the mean of a random variable with known variance. A significant t will indicate that one or more elements of the a priori values of R and Q are in error.

Another possible test variable is defined by

$$\chi^2 = S_n^\xi \quad (J_n^\xi)^{-1} \; S_n^{\xi T}$$

It can be shown that under the null hypothesis, χ^2 is asymptotically a central chi square random variable with r degrees of freedom, where r is the dimension of the vector ξ. Tests on this variable can be conducted using the results of tests on the variance previously outlined.

It is difficult to assess the relative power of these two tests under deviations from the null hypothesis. Even if the distribution of the score under deviations from the null hypothesis could be found, the power of the tests could be found only after great computational expense. However these tests do have the distinct advantage of using test parameters which allow a determination of the first linear correction in the a priori values of R and Q if the hypothesis test fails, using the results of the linearized maximum likelihood solution discussed in Chapter 4.

Chapter 6

NUMERICAL RESULTS

6.1 Introduction

Theoretical results about various techniques for estimating noise covariance parameters and testing statistical hypotheses have been developed in the preceding chapters. This chapter is devoted to a discussion of the results of a digital computer simulation of the equations derived. The purpose of this simulation is twofold. The theoretical results must be checked to ensure that they accurately portray the situation. Once the validity of these results is established, a numerical comparison of the various techniques for estimating the noise covariance parameters will be made to determine the trade-offs involved in using simpler but less accurate methods of estimation.

The principal theoretical results that are to be checked are:

1) convergence of the iterative maximum likelihood solution
2) the unbiasedness of the maximum likelihood solution
3) comparison of the actual mean squared estimation error of the maximum likelihood solution with the inverse information matrix
4) the range of applicability of the linearized maximum likelihood solution

5) convergence of the near maximum likelihood solution

6) comparison of actual mean and mean squared estimation error of the explicit suboptimal solution with the theoretical expressions for the quantities

7) the sensitivity and power of hypothesis testing in realistic situations

The system simulated was purposely made simple. Many of the estimation equations are very complex and require iterative solutions. Only by limiting the complexity of the system and the number of parameters to be estimated could the required computations be kept within reasonable limits.

In checking the above theoretical results, Monte Carlo simulations are required. Many trials are required in which actual noises and realistic parameter values are simulated, this being a time consuming and expensive procedure. However, once the theoretical results are established, then Monte Carlo simulations are not required, thus allowing statistical simulation in which only the expressions for the mean and mean squared error of the estimates are computed, resulting in the ensemble average of the results that would be obtained if a large series of Monte Carlo simulations were performed.

6.2 Description of System and Measurement

The system simulated is a second order damped oscillator with time invariant damping ratio and natural frequency, driven by stationary zero mean uncorrelated normally distributed noise. The state of the system is defined as a two component column vector of the position and velocity coordinates of the system.

$$x_n = \Phi(n,n-1)x_{n-1} + \Gamma_n w_n$$

where x_n is the state at time "n"

x_{n-1} is the state at time "n-1"

$\Phi(n,n-1)$ is the 2 x 2 state transition matrix

w_n is either a scalar or 2 x 1 column vector driving noise

Γ_n is either a 2 x 1 or 2 x 2 forcing function matrix

Q is the driving noise covariance matrix

The state transition matrix obeys the differential equation

$$\frac{d\Phi(t,t_o)}{dt} = F(t)\ \Phi(t,t_o), \qquad \Phi(t_o,t_o) = I$$

For a second order oscillator with time invariant parameters,

$$F = \begin{bmatrix} 0 & 1 \\ -\Omega^2 & -2\zeta\Omega \end{bmatrix}$$

where ζ is the damping ratio and Ω is the system natural frequency.

The measurement of the state is either a scalar or a 2 x 1 column vector defined by

$$z_n = H_n x_n + v_n$$

where H_n is a 1 x 2 or 2 x 2 time invariant measurement matrix

v_n is a scalar or 2 x 1 column vector measurement noise

R is the measurement noise covariance matrix

In the simulation of variance estimation, the values of the diagonal elements of the measurement and driving noise covariance matrices are chosen from a Gamma distribution as described in Chapter 3.

6.3 Effect of Incorrect Noise Covariance Parameters Upon Maximum Likelihood State Estimation

In Section 2.3 equations were derived for the evaluation of the performance of a maximum likelihood state estimator when incorrect values of the measurement and driving noise covariance matrices are used in the computation of the measurement residual weighting matrices. It was shown that even if incorrect values of the noise parameters are used, the maximum likelihood estimator remains unbaised. However, the covariance of the estimation error is a function of the errors in the noise parameters. In this simulation the "true" and "computed" state covariance matrices are calculated as functions of the true and assumed values of the measurement and driving noise covariance matrices.

From (2.3.39) and (2.3.42), the computed state covariance matrix obeys the recursive relationships

$$P^*_{n|n} = (I-A^*_n H_n) P^*_{n|n-1} (I-A^*_n H_n)^T + A^*_n R^* A^{*T}_n$$

$$P^*_{n|n-1} = \Phi(n,n-1) P^*_{n-1|n-1} \Phi^T(n,n-1) + \Gamma_n Q^* \Gamma^T_n$$

$$A^*_n = P^*_{n|n-1} H^T_n (R^* + H_n P^*_{n|n-1} H^T_n)^{-1}$$

where R^* and Q^* are the assumed values of the measurement and driving noise covariance matrices.

From (2.3.43) and (2.3.44), the true state covariance matrix obeys the recursive relationships

$$P_{n|n} = (I-A^*_n H_n) P_{n|n-1} (I-A^*_n H_n)^T + A^*_n R A^{*T}_n$$

$$P_{n|n-1} = \Phi(n,n-1) P_{n-1|n-1} \Phi^T(n,n-1) + \Gamma_n Q \Gamma^T_n$$

where R and Q are the true values of the measurement and driving noise covariance matrices. It is assumed that

$$P^*_{o|o} = P_{o|o}$$

The following graphs show the variation in the trace of the true and computed covariance matrices after the last measurement as a function of the estimated values of R and Q. For simplicity the measurement and driving noises are scalar random variables and the time interval between measurements is constant. The parameters of the system and the measurements are:

$$\zeta = .05, \qquad \Omega = .1 \text{ rad/sec}$$

$$H_n = (1,0), \qquad \Gamma_n^T = (0,1)$$

Time between measurements = 1 sec
Total number of measurements = 200

$$P_{o|o} = \begin{bmatrix} 10 & 0 \\ 0 & 10 \end{bmatrix}$$

For the values of the system and noise parameters chosen, the covariance equations reach a steady state after approximately 10 measurements.

It should be noted that each of the two diagonal elements of $P^*_{n|n}$ had the same general variation with R^* and Q^* as the trace of $P^*_{n|n}$. For simplicity, only graphs of the trace of $P^*_{n|n}$ versus R^* and Q^* are shown.

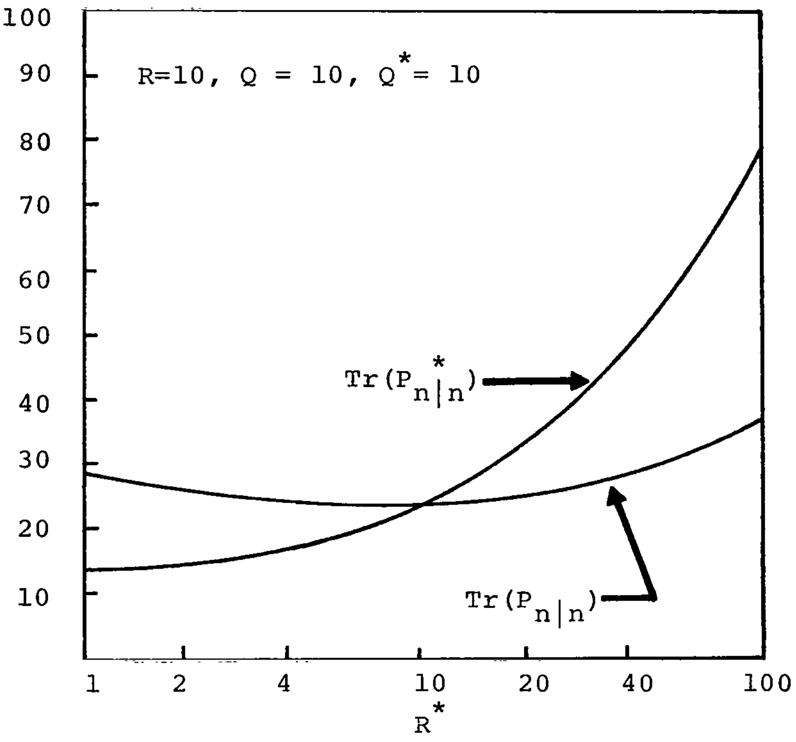

Fig. 6.1 Covariance vs Estimated R

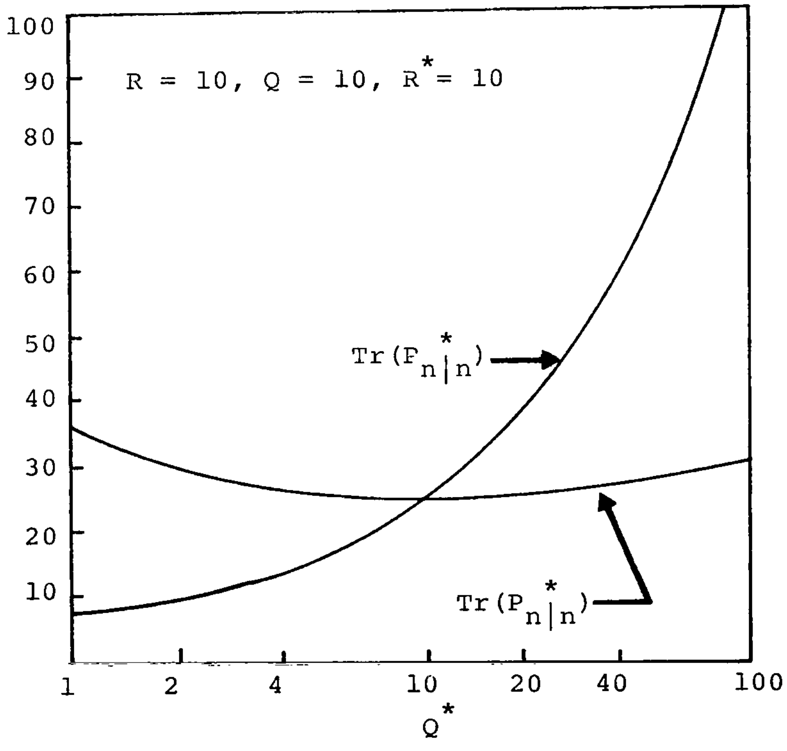

Fig. 6.2 Covariance vs Estimated Q

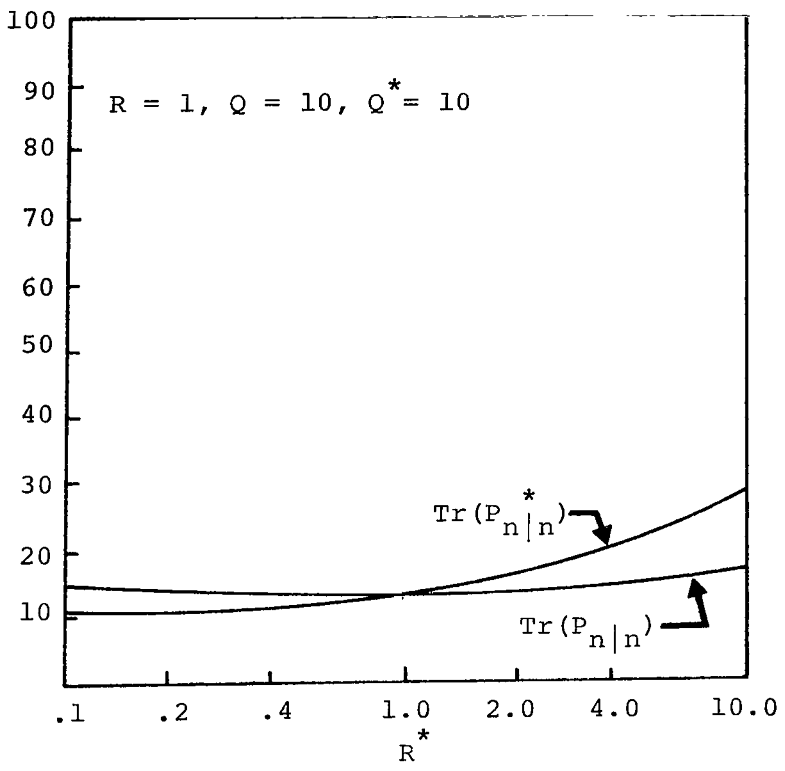

Fig. 6.3 Covariance vs Estimated R

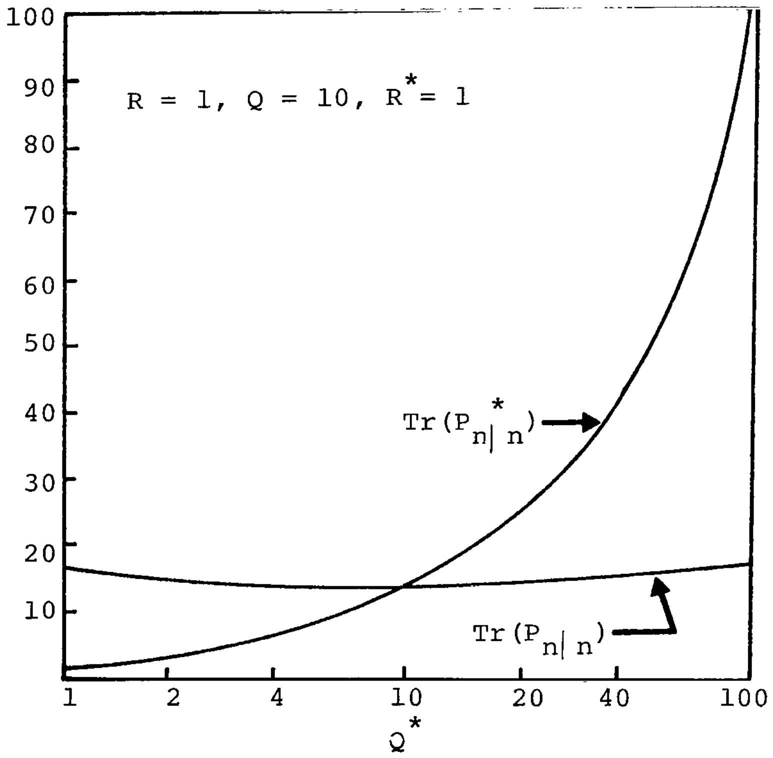

Fig. 6.4 Covariance vs Estimated Q

As can be seen, the trace of the true state estimation error covariance matrix $P_{n|n}$ is not highly sensitive to errors in the estimated values of R or Q. This means that the estimation error is not highly sensitive to use of incorrect values of R and Q in the computation of the measurement residual weighting matrices. However, the trace of the computed state estimation error covariance matrix $P^*_{n|n}$ is a strong function of errors in R and Q. This means that for moderate errors in R and Q, the computed covariance matrix is a poor representation of the actual state estimation error covariance. Although the actual error covariance may be small, there is no way to know this unless R^* and Q^* are very near the true values of R and Q. Therefore any decision made about the probable state estimation error using $P^*_{n|n}$ may be incorrect due to the large difference between $P_{n|n}$ and $P^*_{n|n}$.

6.4 Comparison of State and Noise Covariance Estimation Procedures

Four procedures for estimating the state and noise covariance parameters are simulated and compared:

1) maximum likelihood

2) linearized maximum likelihood

3) near maximum likelihood

4) explicit suboptimal

The simulations are divided into two parts, Monte Carlo simulations and statistical simulations.

Maximum Likelihood and Linearized Maximum Likelihood

The equations for the simultaneous estimation of the system state and noise covariance parameters with a priori information about the state and noise covariance were solved by the iterative procedures of Section 3.6. In Chapter 3, it was shown that the asymptotic distribution of the R and Q estimation error is a zero mean normal distribution with conditional covariance $W_n(\xi)$, where $W_n^{-1}(\xi)$ is a submatrix of the conditional information matrix, and

$$\xi^T = (R^{11},..,R^{\gamma\gamma},Q^{11},..,Q^{\eta\eta})$$

$$W_n^{-1}(\xi) = \varepsilon\left[\left(\frac{\partial L_n^A}{\partial \xi}\right)^T \frac{\partial L_n^A}{\partial \xi} \,|\, \xi\right]$$

In an actual situation, the above matrix cannot be computed because the true value of ξ is unknown. However, it is usually a good approximation to compute W_n^{-1} at the estimated value of ξ if a measure of the R and Q estimation error covariance is desired. All evaluations of the conditional information matrices in this section are at the <u>true</u> value of ξ.

In the case of scalar R and Q, $W_n(\xi)$ is a 2 x 2 matrix with elements

$$W_n(\xi) = \begin{bmatrix} \varepsilon[(\Delta R)^2] & \varepsilon[(\Delta R\Delta Q)] \\ \varepsilon[(\Delta R\Delta Q)] & \varepsilon[(\Delta Q)^2] \end{bmatrix}$$

where ΔR and ΔQ represent the R and Q estimation error. The square root of the appropriate diagonal element of $W_n(\xi)$ is the standard deviation of the corresponding noise covariance parameter asymptotic estimation error. The normalized estimation error can then be defined by

$$e_R = \frac{\hat{R} - R}{\sigma_{\hat{R}}} \tag{6.4.1}$$

where $\hat{R}$ is the estimate of R on a given trial, R is the true value of R on that trial, and $\sigma_{\hat{R}}$ is the standard deviation of the error as given above. A similar expression is used to define the normalized Q estimation error.

In order to check the theoretical unbiasedness and covariance of the estimates, the mean and mean squared error of the estimates over the ensemble of trials is computed and compared with the mean of the true values of R and Q and the average conditional information matrix. The average R is defined by

$$\text{ave}(R) = \frac{1}{K} \sum_{j=1}^{K} R_j \tag{6.4.2}$$

where R_j is the value of R on the j^{th} trial and K is the number of trials. A similar expression is used to compute ave(Q). The average of $\hat{R}$ is defined by

$$\text{ave}(\hat{R}) = \frac{1}{K} \sum_{j=1}^{K} \hat{R}_j \tag{6.4.3}$$

where $\hat{R}_j$ is the value of $\hat{R}$ on the j^{th} trial. A similar expression is used to compute ave$(\hat{Q})$.

The theoretical mean squared estimation error, averaged over the ensemble of all possible R and Q, was given in Chapter 3 by

$$E[(\hat{\xi} - \xi)(\hat{\xi} - \xi)^T] = \overline{W}_n$$

$$= \int_{\Omega} W_n(\xi)\ f(\xi)\ d\xi$$

As was noted, the above integral is difficult to compute. An easier to compute and possibly better measure of the average conditional covariance over the ensemble of values of ξ actually experienced in the trials would be

$$\overline{W}_n \simeq \frac{1}{K} \sum_{j=1}^{K} W_n(\xi_j) \qquad (6.4.4)$$

where ξ_j is the value of ξ on the j^{th} trial. The actual mean squared estimation error matrix is defined by

$$\frac{1}{K} \sum_{j=1}^{K} (\hat{\xi}_j - \xi_j)(\hat{\xi}_j - \xi_j)^T \qquad (6.4.5)$$

Similar expressions are used to compute the mean and mean squared estimation error of the linearized solution of the likelihood equations. The conditional information matrix associated with the linearized solution is computed at the <u>a priori values</u> of R and Q. If these values are not close to the true values, the information matrix computed at the a priori values may not accurately represent the inverse of the estimation error covariance. However, it represents the best measure of the estimation error covariance that would be available for a real time solution without having to recompute the information matrix at the linearized estimates of R and Q.

Figure 6.5 and Table 6.1 give the results of a ten sample Monte Carlo simulation. The system and measurement parameter values are those previously given, while the true values of R and Q are different on each trial. The values are selected from a Gamma distribution with

$$E(R) = \overline{R} = 1.0, \quad E(Q) = \overline{Q} = 0.5$$

$$E[(R - \overline{R})^2] = 1.0, \quad E[(Q - \overline{Q})^2] = 0.25$$

If there was no estimation error, the estimates of R and Q would lie along the diagonal line $\hat{R} = R$ and $\hat{Q} = Q$. The dispersion about this line is a measure of the estimation error.

Shown in Table 6.1 are the standard deviation of the R and Q estimation error and the normalized estimation error defined by (6.4.1). Also shown are the results of the linearized maximum likelihood solution. The estimates of R and Q are those obtained on the first iteration of the optimal solution. As described in Section 4.2, the linearized solution represents an estimate of R and Q that can be obtained in real time. The initial estimates of R and Q were equal to the means of their respective distributions. As can be seen, the linearized solution is quite close to the iterative solution even for large departures of the true values of R and Q from the initial estimates which were used to compute the score and conditional information matrices.

Figure 6.6 and Table 6.2 show the results of another set of ten Monte Carlo trials with a different set of random numbers used to simulate the noises and a different set of values of R and Q chosen from a Gamma distribution with

$$E(R) = 10, \quad E(Q) = 1$$

$$E[(R - \overline{R})^2] = 100, \quad E[(Q - \overline{Q})^2] = 1$$

Again the actual mean and mean squared estimation error over the ensemble of ten trials were computed and compared with the theoretical results. It can be seen that the mean of the estimates compares quite well with the mean of the actual values of R and Q. However, there is a rather large difference between the theoretical and actual mean squared estimation error matrices. This is a hazard of trying to compute ensemble statistics on the basis of ten samples. Almost all of the actual mean squared R estimation error comes from Sample 7, the error being nonrepresentative of the expected error. The actual error was a 2.57 sigma error based upon the standard deviation obtained from the conditional information matrix. Omitting this sample from the ensemble averages results in good agreement between theoretical and actual mean squared errors.

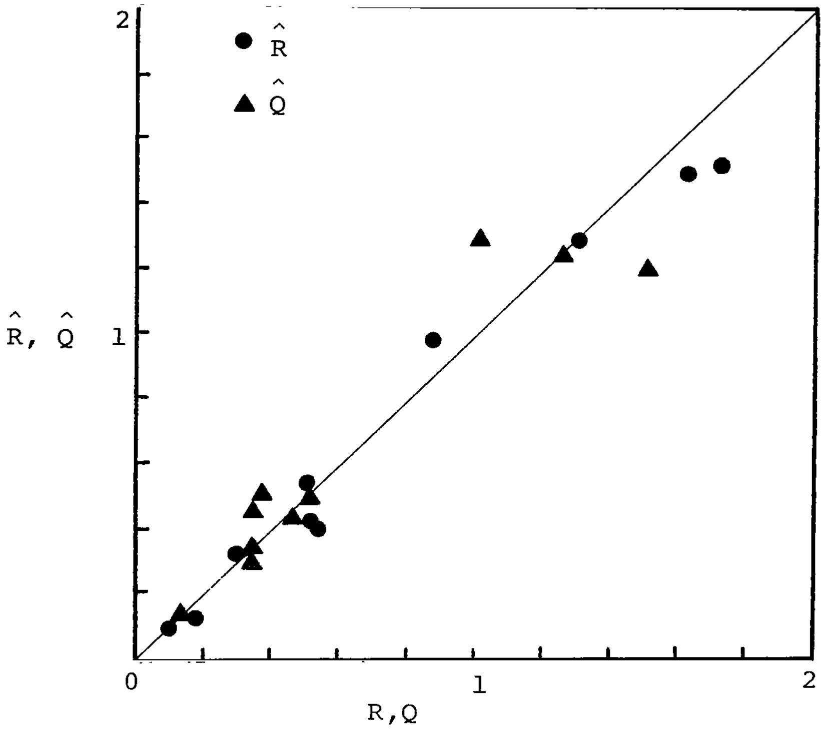

Fig. 6.5 Maximum Likelihood Solution Run 1

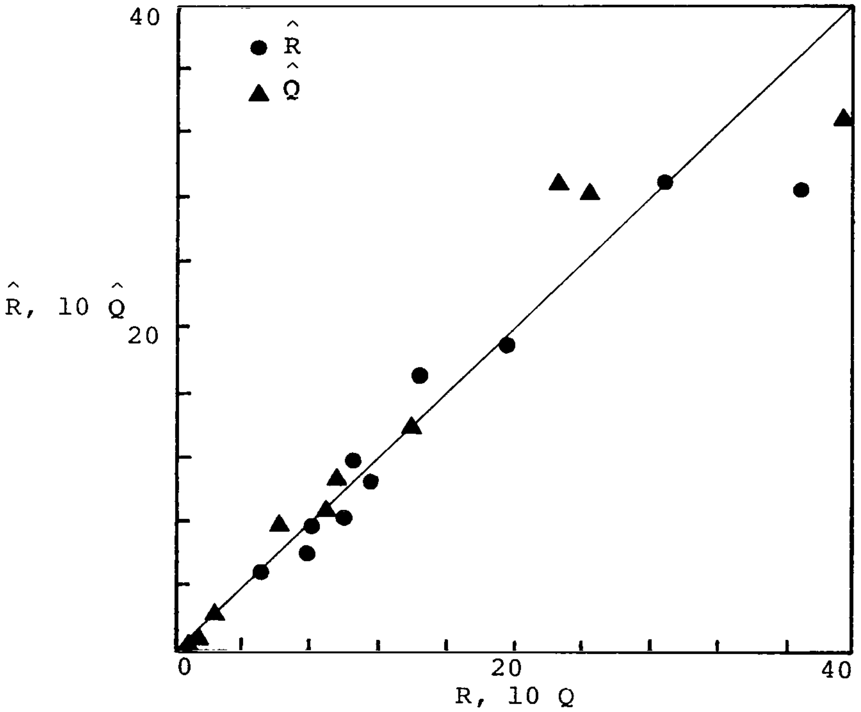

Fig. 6.6 Maximum Likelihood Solution Run 2

Table 6.1 Monte Carlo Run 1: Maximum Likelihood and Linearized Maximum Likelihood Solutions

Sample	R	M.L. $\hat{R}$	$\sigma_{\hat{R}}$	$e_{\hat{R}}$	Lin. $\hat{R}$
1	0.860	0.979	0.118	+1.014	0.978
2	1.599	1.461	0.207	-0.670	1.449
3	0.522	0.384	0.074	-1.876	0.391
4	1.288	1.276	0.165	-0.074	1.286
5	0.526	0.404	0.092	-1.322	0.394
6	0.102	0.096	0.021	-0.276	0.096
7	0.304	0.291	0.047	-0.288	0.317
8	1.718	1.470	0.200	-1.386	1.489
9	0.160	0.188	0.052	-0.807	0.128
10	0.484	0.523	0.092	+0.425	0.514
Average	.756	.700	0.107	-0.526	.704

Theoretical Mean Squared Estimation Error

Linearized

$$\begin{bmatrix} 0.0184 & -0.0039 \\ -0.0039 & 0.0101 \end{bmatrix}$$

Iterative

$$\begin{bmatrix} 0.0152 & -0.0035 \\ -0.0035 & 0.0163 \end{bmatrix}$$

Actual Mean Squared Estimation Error

Linearized

$$\begin{bmatrix} 0.0121 & -0.0034 \\ -0.0034 & 0.0233 \end{bmatrix}$$

Iterative

$$\begin{bmatrix} 0.0133 & -0.0034 \\ -0.0034 & 0.0204 \end{bmatrix}$$

Table 6.1 (Continued) Monte Carlo Run 1

Sample	Q	M.L. $\hat{Q}$	$\sigma_{\hat{Q}}$	$e_{\hat{Q}}$	Lin. $\hat{Q}$
1	0.456	0.445	0.093	-0.115	.450
2	0.512	0.484	0.111	-0.260	.483
3	0.330	0.447	0.066	+1.770	.463
4	0.360	0.498	0.079	+1.740	.483
5	0.997	1.286	0.174	+1.665	1.264
6	0.327	0.308	0.056	-0.345	.362
7	0.329	0.328	0.062	-0.009	.336
8	0.145	0.128	0.037	-0.480	.120
9	1.487	1.194	0.221	-1.325	1.133
10	1.249	1.242	0.209	-0.004	1.214
Average	.619	.641	0.111	+0.264	.636

Table 6.2 Monte Carlo Run 2: Maximum Likelihood and Linearized Maximum Likelihood Solutions

Sample	R	M.L. $\hat{R}$	$\sigma_{\hat{R}}$	$e_{\hat{R}}$	Lin. $\hat{R}$
1	8.231	7.584	1.120	-0.606	8.882
2	8.104	6.176	1.013	-1.900	6.303
3	10.359	8.369	1.123	-1.770	8.200
4	5.362	5.838	0.558	+0.855	5.283
5	14.526	17.012	1.780	+1.400	16.907
6	11.753	10.514	1.375	-0.900	10.522
7	37.523	27.297	3.980	-2.570	28.399
8	29.982	28.551	3.290	+0.436	30.378
9	10.499	11.929	1.316	+1.090	11.942
10	19.663	20.205	2.135	+0.254	18.914
Average	15.599	14.349	1.769	-0.371	14.543

Theoretical Mean Squared Estimation Error

Linearized

$$\begin{bmatrix} 1.420 & -0.049 \\ -0.049 & 0.063 \end{bmatrix}$$

Iterative

$$\begin{bmatrix} 4.191 & -0.092 \\ -0.092 & 0.163 \end{bmatrix}$$

Actual Mean Squared Estimation Error

Linearized

$$\begin{bmatrix} 10.152 & 0.179 \\ 0.179 & 0.274 \end{bmatrix}$$

Iterative

$$\begin{bmatrix} 12.494 & 0.425 \\ 0.425 & 0.115 \end{bmatrix}$$

Table 6.2 (Continued) Monte Carlo Run 2

Sample	Q	M.L. $\hat{Q}$	$\sigma_{\hat{Q}}$	$e_{\hat{Q}}$	Lin. Q
1	3.985	3.237	0.828	-0.093	2.784
2	1.677	1.461	0.385	-0.562	1.422
3	0.089	0.044	0.029	-1.530	0.088
4	0.002	0.002	9×10^{-4}	+0.000	0.046
5	2.476	2.833	0.582	+0.614	3.164
6	0.923	1.047	0.750	+0.165	1.048
7	0.614	0.786	0.188	+0.915	0.637
8	0.906	0.878	0.260	-0.017	0.592
9	2.361	2.895	0.537	+0.995	3.134
10	0.255	0.211	0.080	-0.547	0.455
Average	1.329	1.339	0.334	-0.087	1.337

Runs 3 and 4 and the corresponding Tables 6.3 and 6.4 are the results of the above two runs repeated except that the values of R and Q are held fixed at the same values on each trial. Different random numbers were used to simulate the measurement and driving noises. These runs simulate an ensemble of trials with a fixed value of ξ_o, so that the conditional information matrix is the same for each trial. Then the theoretical mean squared estimation error is given by $W_n(\xi_o)$, where ξ_o is the value of ξ for every trial.

The agreement between the sample mean and mean squared estimation error and the theoretical results is quite good for both runs. A better correspondence between theoretical and actual results is expected in these runs than in the first two runs because the ten trials in each of these runs are samples from an ensemble of trials with different noises but with the same noise covariances. The first two runs were samples from an ensemble with different noises _and_ different noise covariances.

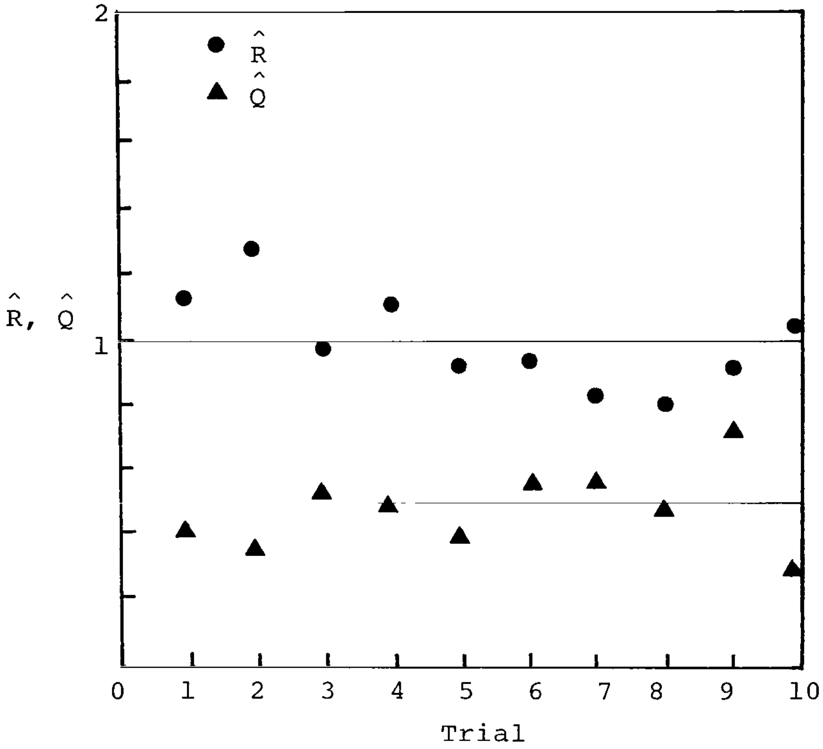

Fig. 6.7 Maximum Likelihood Solution Run 3

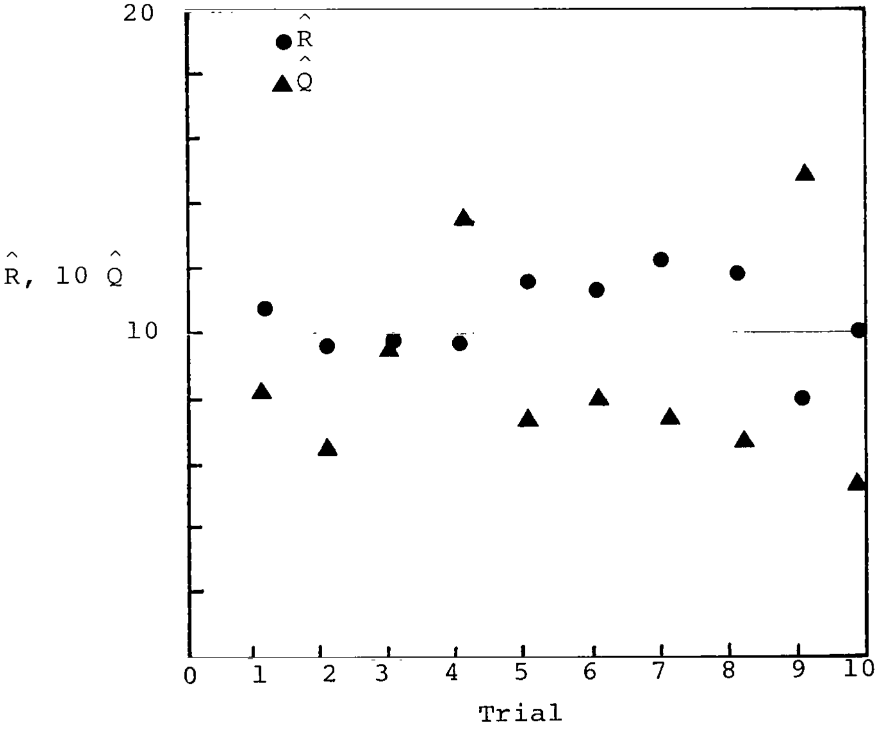

Fig. 6.8 Maximum Likelihood Solution Run 4

Table 6.3 Monte Carlo Run 3: Maximum Likelihood and Linearized Maximum Likelihood Solutions

Sample	R	M.L. $\hat{R}$	$\sigma_{\hat{R}}$	$e_{\hat{R}}$	Lin. $\hat{R}$
1	1.0	1.130	0.136	+0.96	1.134
2	1.0	1.295	0.136	+2.18	1.286
3	1.0	.982	0.136	-0.13	.977
4	1.0	1.106	0.136	+0.78	1.110
5	1.0	.921	0.136	-0.58	.935
6	1.0	.936	0.136	-0.47	.934
7	1.0	.821	0.136	-1.32	.810
8	1.0	.807	0.136	-1.42	.810
9	1.0	.908	0.136	-0.72	.939
10	1.0	1.043	0.136	+0.32	1.038
Average	1.0	.995	0.136	-0.04	.998

Theoretical Mean Squared Estimation Error

Linearized

$$\begin{bmatrix} .0184 & -.0039 \\ -.0039 & .101 \end{bmatrix}$$

Iterative

$$\begin{bmatrix} .0182 & -.0036 \\ -.0036 & .0096 \end{bmatrix}$$

Actual Mean Squared Estimation Error

Linearized

$$\begin{bmatrix} .0197 & -.0076 \\ -.0076 & .0118 \end{bmatrix}$$

Iterative

$$\begin{bmatrix} .0205 & -.0084 \\ -.0084 & .0140 \end{bmatrix}$$

Table 6.3 (Continued) Monte Carlo Run 3

Sample	Q	M.L. $\hat{Q}$	$\sigma_{\hat{Q}}$	$e_{\hat{Q}}$	Lin. $\hat{Q}$
1	.50	.421	.101	-0.79	.418
2	.50	.357	.101	-1.43	.358
3	.50	.534	.101	+0.34	.537
4	.50	.495	.101	-0.05	.490
5	.50	.392	.101	-1.08	.387
6	.50	.556	.101	+0.56	.549
7	.50	.575	.101	+0.75	.587
8	.50	.465	.101	-0.35	.471
9	.50	.712	.101	+2.12	.672
10	.50	.287	.101	-2.13	.304
Average	.50	.479	.101	-0.21	.478

Table 6.4 Monte Carlo Run 4: Maximum Likelihood and Linearized Maximum Likelihood Solutions

Sample	R	M.L. $\hat{R}$	$\sigma_{\hat{R}}$	$e_{\hat{R}}$	Lin. $\hat{R}$
1	10.0	10.812	1.185	+0.715	10.858
2	10.0	9.527	1.185	-0.291	9.658
3	10.0	9.872	1.185	-0.105	9.876
4	10.0	9.750	1.185	-0.236	9.721
5	10.0	11.624	1.185	+1.330	11.573
6	10.0	11.385	1.185	+1.210	11.437
7	10.0	12.195	1.185	+2.120	12.514
8	10.0	11.748	1.185	+1.510	11.781
9	10.0	7.757	1.185	-0.985	7.831
10	10.0	9.761	1.185	-0.905	9.893
Average	10.0	10.443	1.185	+0.374	10.514

Theoretical Mean Squared Estimation Error

Linearized

$$\begin{bmatrix} 1.405 & -.059 \\ -.059 & .058 \end{bmatrix}$$

Iterative

$$\begin{bmatrix} 1.503 & -.050 \\ -.050 & .011 \end{bmatrix}$$

Actual Mean Squared Estimation Error

Linearized

$$\begin{bmatrix} 1.970 & -.321 \\ -.321 & .106 \end{bmatrix}$$

Iterative

$$\begin{bmatrix} 1.848 & -.296 \\ -.296 & .107 \end{bmatrix}$$

Table 6.4 (Continued) Monte Carlo Run 4

Sample	Q	M.L. $\hat{Q}$	$\sigma_{\hat{Q}}$	$e_{\hat{Q}}$	Lin. $\hat{Q}$
1	1.0	.820	.238	-0.757	.814
2	1.0	.641	.238	-1.510	.637
3	1.0	.975	.238	-0.105	.976
4	1.0	1.341	.238	+1.430	1.329
5	1.0	.715	.238	-1.200	.737
6	1.0	.785	.238	-0.905	.776
7	1.0	.721	.238	-1.170	.645
8	1.0	.648	.238	-1.480	.654
9	1.0	1.461	.238	+1.930	1.413
10	1.0	.494	.238	-2.120	.502
Average	1.0	.860	.238	-0.590	.848

In Runs 1-4, the measurement and driving noise covariance matrices are scalars. In Run 5, both R and Q are 2 x 2 matrices, so that four noise covariance parameters are to be estimated, the diagonal elements of R and Q. Figures 6.9 and 6.10 and Table 6.5 show the results of a ten sample Monte Carlo simulation.

As before, the theoretical and actual mean and mean squared estimation error matrices were computed. Now the mean squared estimation errors (both theoretical and actual) are 4 x 4 matrices, with elements

$$\begin{bmatrix} \overline{(\Delta R_o)^2} & \overline{(\Delta R_o \Delta R_1)} & \overline{(\Delta R_o \Delta Q_o)} & \overline{(\Delta R_o \Delta Q_1)} \\ \overline{(\Delta R_1 \Delta R_o)} & \overline{(\Delta R_1)^2} & \overline{(\Delta R_1 \Delta Q_o)} & \overline{(\Delta R_1 \Delta Q_1)} \\ \overline{(\Delta Q_o \Delta R_o)} & \overline{(\Delta Q_o \Delta R_1)} & \overline{(\Delta Q_o)^2} & \overline{(\Delta Q_o \Delta Q_1)} \\ \overline{(\Delta Q_1 \Delta R_o)} & \overline{(\Delta Q_1 \Delta R_1)} & \overline{(\Delta Q_1 \Delta Q_o)} & \overline{(\Delta Q_1)^2} \end{bmatrix}$$

where ΔR and ΔQ represent the R or Q estimation error and the bar over these quantities indicates either the theoretical or actual mean, depending upon which matrix is given.

As can be seen, increasing the number of quantities to be estimated did not degrade the performance of the estimator. Of course, the number of computations needed to estimate four covariance parameters is considerably greater than that needed to estimate two covariance parameters.

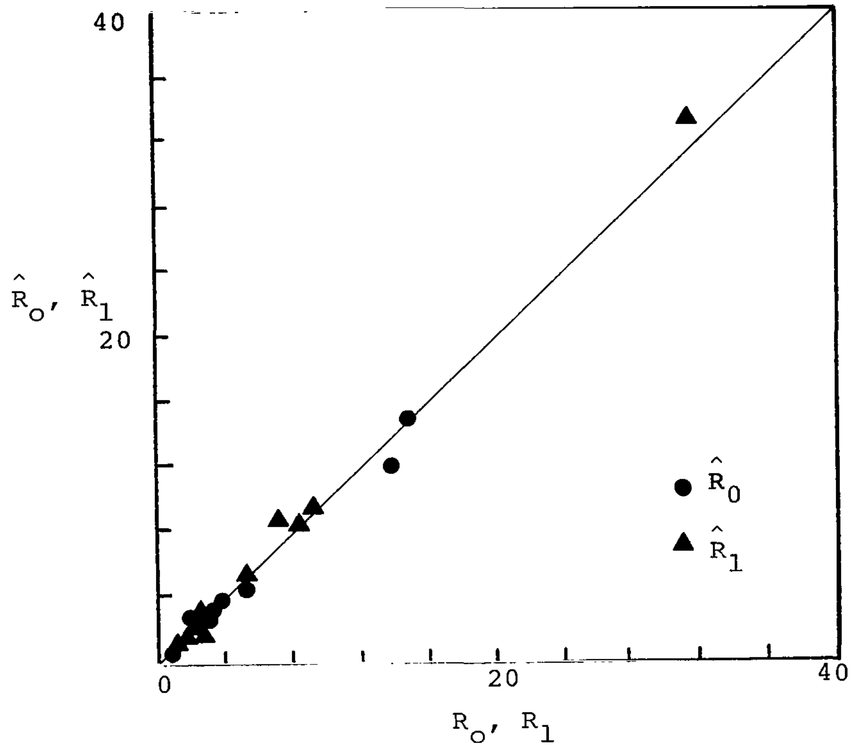

Fig. 6.9 Maximum Likelihood Solution Run 5

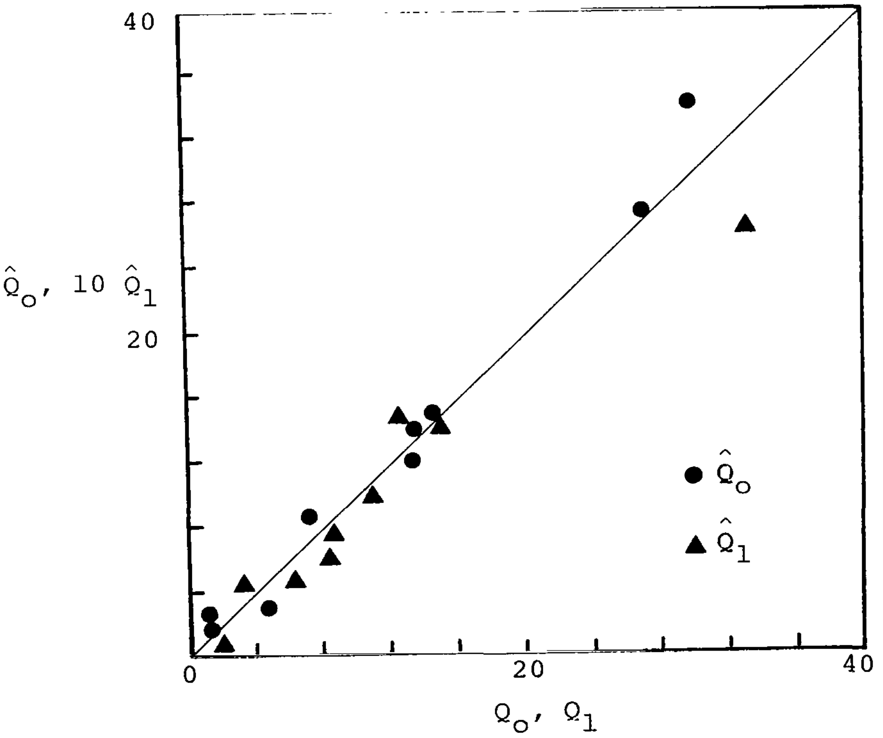

Fig. 6.10 Maximum Likelihood Solution Run 5 (cont'd.)

Table 6.5 Monte Carlo Run 5: Maximum Likelihood and Linearized Maximum Likelihood Solutions

Sample	R_o	M.L. $\hat{R}_o$	$\sigma_{\hat{R}_o}$	$e_{\hat{R}_o}$	Lin. $\hat{R}_o$
1	13.857	11.806	1.578	-1.300	11.836
2	2.137	1.887	0.598	-0.416	2.373
3	0.483	0.313	0.452	-0.376	5.000
4	14.717	15.512	2.013	+0.396	15.047
5	4.553	4.677	1.030	+0.120	4.243
6	44.597	47.657	4.610	+0.665	46.803
7	2.692	2.778	0.546	+0.157	2.917
8	1.963	2.865	0.636	+1.410	2.407
9	2.698	2.281	0.724	-0.575	2.475
10	3.972	4.106	0.467	+0.287	4.089
Average	9.167	9.358	1.265	+0.037	9.772

Theoretical Mean Square Estimation Error Matrix

$$\begin{bmatrix} 3.086 & -0.166 & -0.383 & 0.037 \\ -0.166 & 3.824 & -0.692 & 0.047 \\ -0.383 & -0.692 & 6.302 & -0.621 \\ 0.037 & 0.047 & -0.621 & 0.325 \end{bmatrix}$$

$E(R_o) = 10, \quad E[(R_o - \bar{R}_o)^2] = 100$

$E(R_1) = 10, \quad E[(R_1 - \bar{R}_1)^2] = 100$

$E(Q_o) = 10, \quad E[(Q_o - \bar{Q}_o)^2] = 100$

$E(Q_1) = 1, \quad E[(Q_1 - \bar{Q}_1)^2] = 1$

Table 6.5 (Continued) Monte Carlo Run 5

Sample	R_1	M.L. $\hat{R}_1$	$\sigma_{\hat{R}_1}$	$e_{\hat{R}_1}$	Lin. $\hat{R}_1$
1	2.937	3.115	0.752	+0.237	3.892
2	8.702	9.063	1.057	+0.340	8.619
3	8.325	8.690	0.965	+0.400	9.639
4	31.170	35.171	3.400	+1.180	34.572
5	44.397	40.871	4.420	-0.800	40.560
6	5.500	4.835	1.600	-0.415	6.464
7	0.953	0.654	0.476	-0.630	1.349
8	2.823	2.636	0.670	-0.280	1.924
9	7.598	8.785	1.032	+1.150	8.317
10	1.473	1.146	0.272	-1.200	1.146
Average	11.388	11.450	1.464	-0.002	11.648

Actual Mean Squared Estimation Error

$$\begin{bmatrix} 1.551 & -0.043 & 1.101 & -0.055 \\ -0.043 & 3.070 & -1.626 & 0.165 \\ 1.101 & -1.626 & 4.305 & -0.289 \\ -0.055 & 0.165 & -0.289 & 0.618 \end{bmatrix}$$

Table 6.5 (Continued) Monte Carlo Run 5

Sample	Q_o	M.L. $\hat{Q}_o$	$\sigma_{\hat{Q}_o}$	$e_{\hat{Q}_o}$	Lin. $\hat{Q}_o$
1	4.550	2.953	2.135	-0.750	5.000
2	7.568	8.962	1.732	+0.803	8.707
3	13.596	13.824	2.480	+0.311	17.180
4	1.551	0.735	2.450	-0.333	0.306
5	0.621	2.249	2.410	+0.675	4.077
6	14.583	15.668	3.660	+0.295	13.408
7	25.451	26.445	2.980	+0.334	24.907
8	29.335	34.589	3.250	+1.640	33.971
9	13.677	11.912	2.040	-0.864	11.433
10	1.052	1.424	0.610	+0.610	1.398
Average	11.198	11.876	2.375	+0.272	12.039

Table 6.5 (Continued) Monte Carlo Run 5

Sample	Q_1	M.L. $\hat{Q}_1$	$\sigma_{\hat{Q}_1}$	$e_{\hat{Q}_1}$	Lin. $\hat{Q}_1$
1	4.255	4.183	0.885	-0.083	5.641
2	1.568	1.488	0.473	-0.169	1.423
3	5.173	7.393	1.043	+2.100	6.742
4	1.355	1.050	0.411	-0.742	1.032
5	3.357	2.482	0.762	-1.150	2.223
6	1.218	1.523	0.451	+0.744	1.502
7	0.662	0.421	0.316	-0.764	0.196
8	0.180	0.021	0.123	-1.280	0.564
9	0.328	0.433	0.164	+0.641	0.656
10	0.864	0.741	0.228	-0.540	0.751
Average	1.896	1.978	0.487	-0.134	2.073

A further check was made concerning the hypothesis that the R and Q estimation errors are zero mean normally distributed random variables with conditional covariance $W_n(R,Q)$. Under this hypothesis, the normalized errors e_R and e_Q are zero mean unit variance and normally distributed. Define

$$\bar{e}_R = \frac{1}{K} \sum_{j=1}^{K} e_R^j$$

$$s_R^2 = \frac{1}{K} \sum_{j=1}^{K} (e_R^j - \bar{e}_R)^2$$

where e_R^j is the normalized R estimation error on a given trial and K is the number of trials. Similar expressions are defined for the Q estimation errors. From Chapter 5, under the above hypothesis, $\bar{e}_R$ is a zero mean normally distributed variable with variance (1/K) and K s_R^2 is a chi squared distributed variable with K-1 degrees of freedom, with mean (K-1) and variance 2K. $\bar{e}_R$, $\bar{e}_Q$, s_R^2, and s_Q^2 were computed for each of the ten sample Monte Carlo trials previously presented. In most cases, their computed values were within one standard deviation of their expected values under the above hypothesis. Therefore, the variations of the computed quantities about their means were within that which would be expected due to the relatively small sample size and the above hypothesis can be reasonably accepted as a valid hypothesis.

Near Maximum Likelihood

In Section 4.3 a near maximum likelihood solution for estimating the state and noise covariance parameters was given. In this solution certain terms in the likelihood equations and the conditional information matrix were omitted and the solution of the resulting "pseudo likelihood equations" sought.

The solution of these equations was attempted using the iterative procedure of Chapter 3, whereby the conditional information matrix was used as the negative gradient of the likelihood equations. Serious difficulty was encountered in implementing this solution. The information matrix given by (4.3.2) was nearly singular for the system and measurement schedule under study, resulting in an unstable iterative procedure. A different technique was then used to attempt to find a solution point of the pseudo likelihood equations. Essentially, the procedure was to evaluate the score as a function of the a priori values of R and Q for a given value of the true R and Q. The solutions for $\hat{R}_n$ and $\hat{Q}_n$ were the values of $\hat{R}_o$ and $\hat{Q}_o$ which produced the smallest magnitude of the score. A sufficient number of values of $\hat{R}_o$ and $\hat{Q}_o$ were chosen to reasonably ensure that $\hat{R}_n$ and $\hat{Q}_n$ produced the smallest or near the smallest magnitude of the score.

It was found that the solution point agreed quite well with the solution point of the full likelihood equations just given. In other words, the solution of the pseudo likelihood equations is a good estimate of the noise covariance parameters but a different technique of solution must be used when the information matrix associated with the pseudo likelihood equations is nearly singular. So the computational simplification obtained by omitting certain terms in the likelihood function and information matrix is offset by the need for a more complicated algorithm for finding the solution to the likelihood equations.

Whenever the information matrix is singular or nearly singular, there is a real question as to whether there is a unique solution of the likelihood equations. In the case mentioned above, instead of a single point where the magnitude of the score is minimized, there may exist a line in $\hat{R}_o$ and $\hat{Q}_o$ space along which the magnitude of the score is small and remains essentially constant. In such situations it is impossible to distinguish between errors in the estimates of R and errors in the estimates of Q.

From the limited simulation of the near maximum likelihood solution it is felt that for the system and measurement schedule used, a unique solution of the pseudo likelihood equations does exist. However, finding the solution point requires considerable trial and error. Because of this complication, no series of runs was conducted in which the pseudo likelihood equations were solved. From the few trials that were conducted, it is felt that the solutions are quite close to the solutions of the full likelihood equations.

Explicit Suboptimal

The explicit suboptimal solution of Section 4.4 was simulated so that it could be compared with the maximum likelihood solution. A series of runs was made that corresponds to the series made of the maximum likelihood solution.

Identical random numbers were used in the simulation of the noises so that a meaningful comparison could be made.

In Section 4.4, expressions for the theoretical conditional and unconditional mean and mean squared R and Q estimation error were developed. The conditional mean of the R estimation error was given by

$$\varepsilon(\hat{R}_n^{jj} - R^{jj}) = F_n^{jj}$$

and the conditional mean of the Q estimation was given by

$$\varepsilon(\hat{Q}_n^{jj} - Q^{jj}) = M_n^{jj}$$

where F_n and M_n are defined in Section 4.4. The conditional mean squared R and Q estimation errors were given by

$$\varepsilon[(R_n^{jj} - R^{jj})^2] = \varepsilon[(R_n^{jj} - \varepsilon(R_n^{jj}))^2] + [\varepsilon(R_n^{jj} - R^{jj})]^2$$

$$= G_n^{jj} + (F_n^{jj})^2$$

$$\varepsilon[(Q_n^{jj} - Q^{jj})^2] = \varepsilon[(Q_n^{jj} - \varepsilon(Q_n^{jj}))^2] + [\varepsilon(Q_n^{jj} - Q^{jj})]^2$$

$$= J_n^{jj} + (M_n^{jj})^2$$

where G_n and J_n are defined in Section 4.4. Note that J_n given here is not the conditional information matrix of the maximum likelihood solution. F_n and M_n represent the bias of the estimators and G_n and J_n represent the variance of the estimators about the biased values.

One of the purposes of this simulation is to check the validity of the above expressions. To do this, the following variables are defined.

$$\sigma^2_{\hat{R}^{jj}} = \varepsilon[(\hat{R}^{jj}_n - \varepsilon(\hat{R}^{jj}_n))^2]$$

$$\sigma^2_{\hat{Q}^{jj}} = \varepsilon[(\hat{Q}^{jj}_n - \varepsilon(\hat{Q}^{jj}_n))^2]$$

$$e_{R^{jj}} = \frac{\hat{R}^{jj}_n - \varepsilon(\hat{R}^{jj}_n)}{\sigma_{\hat{R}^{jj}}}$$

$$e_{Q^{jj}} = \frac{\hat{Q}^{jj}_n - \varepsilon(\hat{Q}^{jj}_n)}{\sigma_{\hat{Q}^{jj}}}$$

If the expressions for the conditional means of $\hat{R}^{jj}_n$ and $\hat{Q}^{jj}_n$ are accurate, the normalized errors $e_{R^{jj}}$ and $e_{Q^{jj}}$ should be zero mean unit variance random variables.

In maximum likelihood estimation, the unconditional mean squared error is usually a nonanalytic function. However, in the case of explicit suboptimal estimation, an analytic expression for the unconditional mean squared error was found. Because of the relatively small sample size, this expression is not used. The theoretical mean squared estimation error for R and Q that is shown in the tables is defined as the average conditional mean squared estimation error, averaged over the ensemble of values of R and Q actually encountered. This is the same definition of the theoretical mean squared estimation error that was used in the evaluation of the maximum likelihood estimator. When R and Q are scalars, the theoretical mean squared estimation error is defined by

$$\frac{1}{K} \sum_{j=1}^{K} \varepsilon[(\hat{R}_j - R_j)^2]$$

where here $\hat{R}_j$ is the estimate of R_j on the j^{th} trial and R_j is the true value of R on that trial. A similar expression is used for the theoretical mean squared Q estimation error. The actual mean squared R estimation error is defined by

$$\frac{1}{K} \sum_{j=1}^{K} (\hat{R}_j - R_j)^2$$

with a similar expression for the actual mean squared Q estimation error.

In runs 6 and 7 the values of R and Q were selected from a Gamma distribution with the same population characteristics as runs 1 and 2 respectively. The a priori estimates of R and Q were chosen to be the theoretical means of the appropriate Gamma distribution.

Several things can be seen from an examination of Figures 6.11 and 6.12 and Tables 6.6 and 6.7. First, there is good agreement between the actual values of $\hat{R}$ and $\hat{Q}$ and their conditional means. There is also good agreement between the theoretical and actual mean squared estimation error as defined above. This tends to substantiate the validity of the expressions developed in Chapter 4.

The second thing to notice is that the estimates of R and Q are biased towards the a priori values of $\hat{R}_o$ and $\hat{Q}_o$. The estimates are to a large degree independent of the actual values of R and Q on any given trial. Unlike the maximum likelihood estimator, the explicit suboptimal estimator remains biased even when the number of measurements becomes large. This will become even clearer later when the conditional means of $\hat{R}$ and $\hat{Q}$ are computed as functions of $\hat{R}_o$ and $\hat{Q}_o$ for fixed true values of R and Q.

In runs 8 and 9, the true values of R and Q were held fixed on each sample at the means of their respective distributions. As can be seen from Figures 6.13 and 6.14, when the a priori values of R and Q are equal to the true values of R and Q, the estimates $\hat{R}$ and $\hat{Q}$ are quite closely grouped about the true values. Runs 10 and 11 are repeats of runs 6 and 9, except that the a priori values of R and Q were not equal to the means of the respective distributions of R and Q. For run 10, $\hat{R}_o = 2.0$, $\hat{Q}_o = 1.0$ whereas $\overline{R} = 1.0$, $\overline{Q} = 0.5$, and for run 11, $\hat{R}_o = 20$, $\hat{Q}_o = 2$ whereas $\overline{R} = 10$, $\overline{Q} = 1$. Again it can be seen that the estimators for R and Q are biased if the a priori values of $\hat{R}_o$ and $\hat{Q}_o$ are not equal to the means of their respective distributions, exactly as predicted.

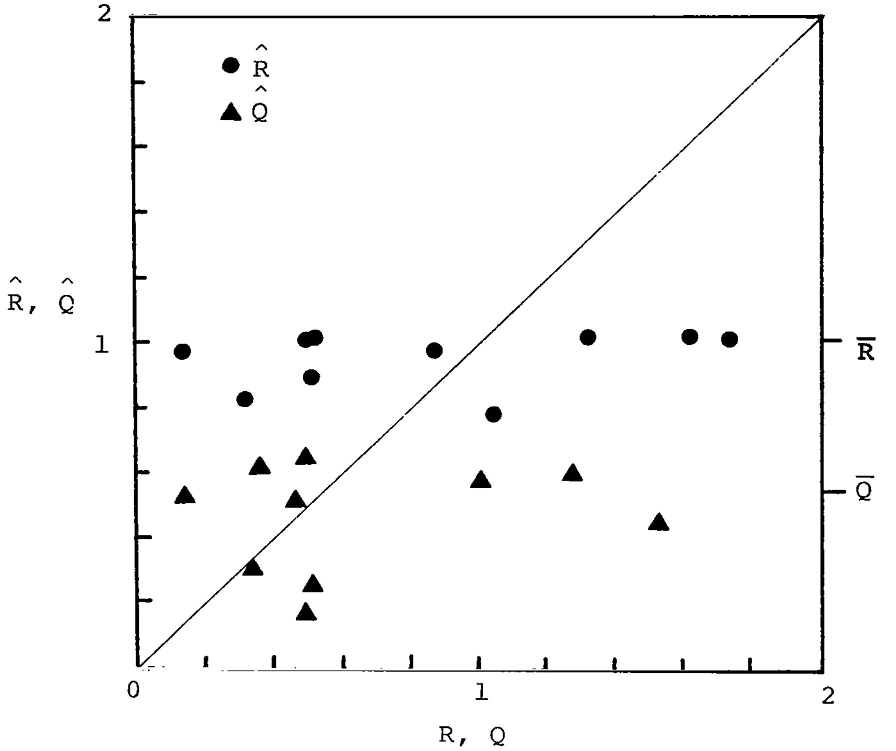

Fig. 6.11 Explicit Suboptimal Solution Run 6

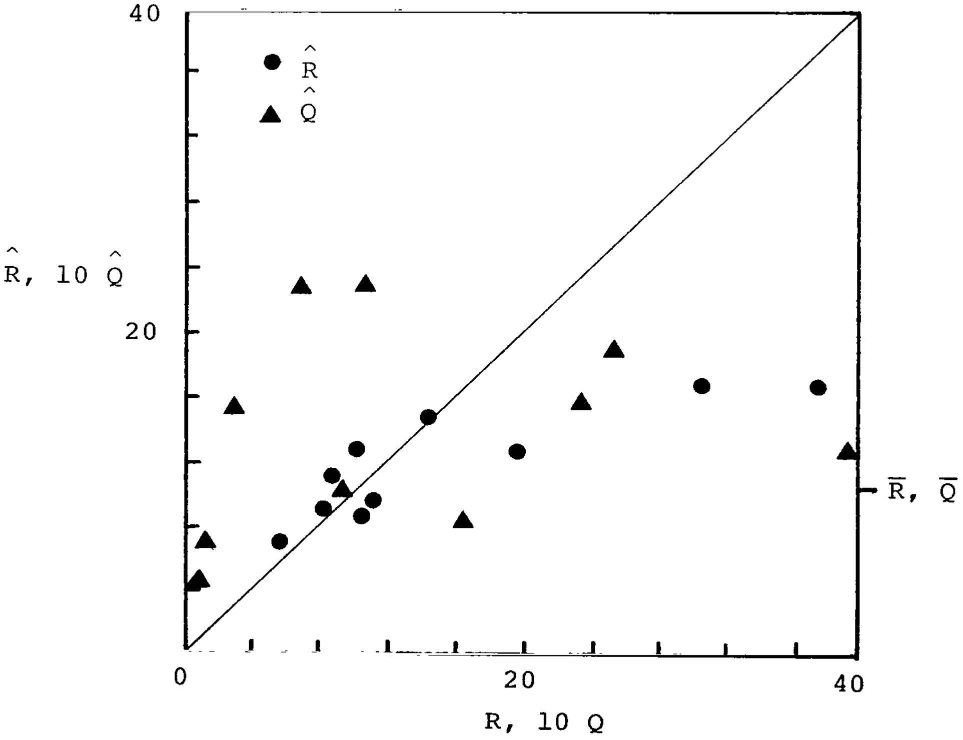

Fig. 6.12 Explicit Suboptimal Solution Run 7

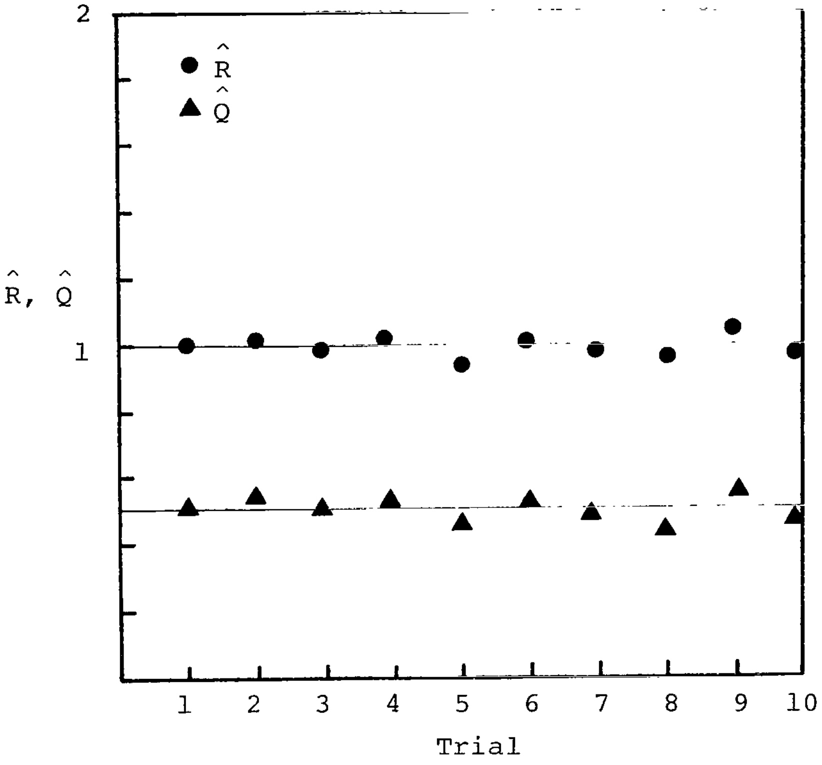

Fig. 6.13 Explicit Suboptimal Solution Run 8

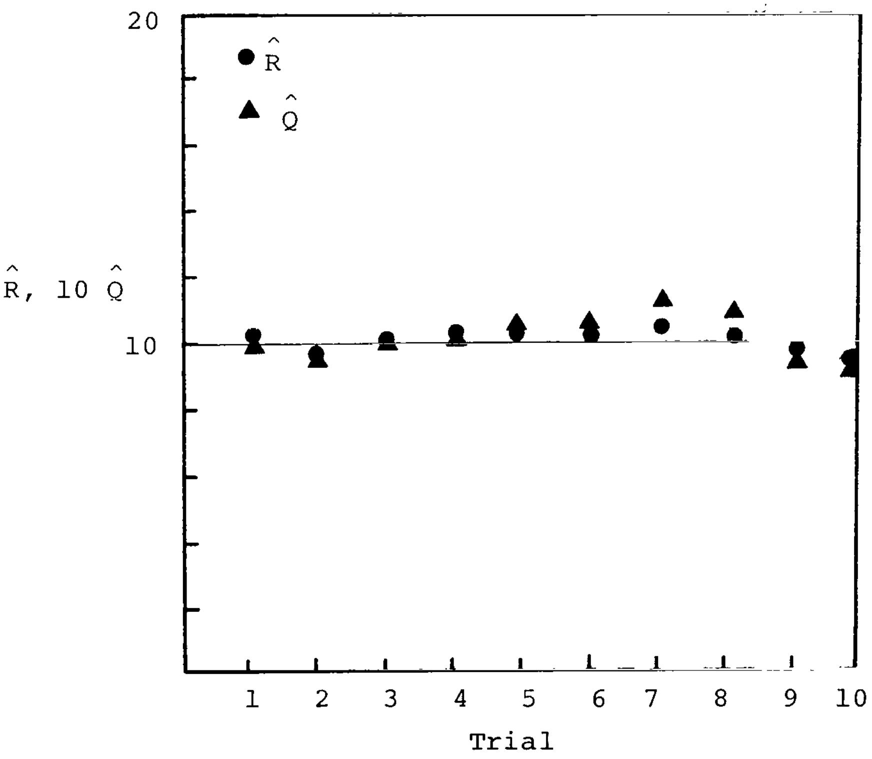

Fig. 6.14 Explicit Suboptimal Solution Run 9

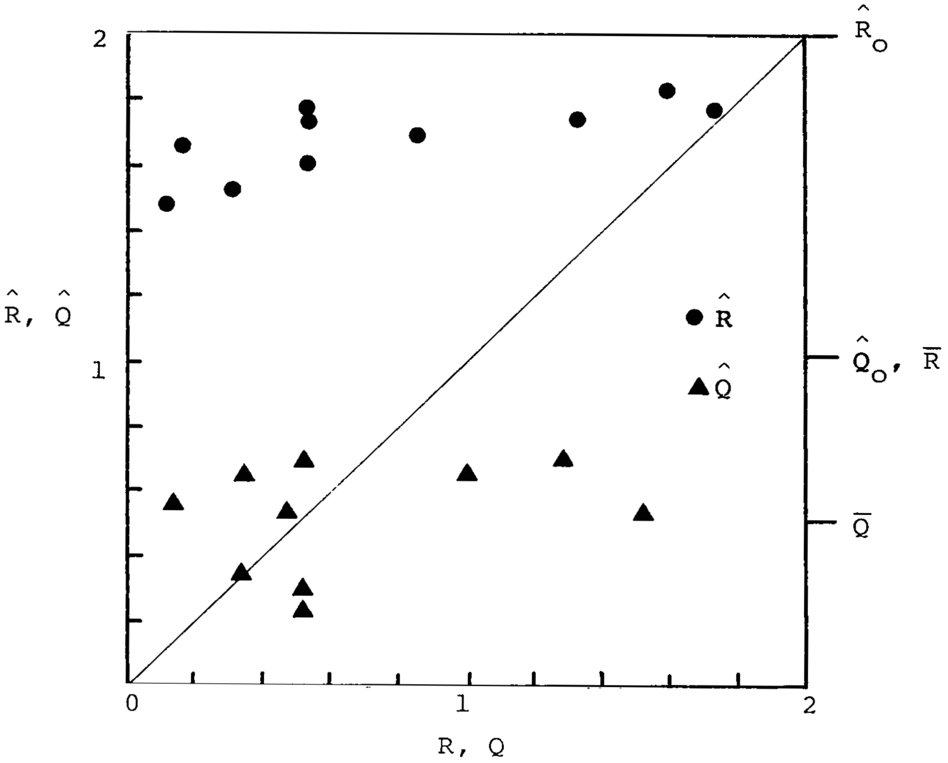

Fig. 6.15 Explicit Suboptimal Solution Run 10

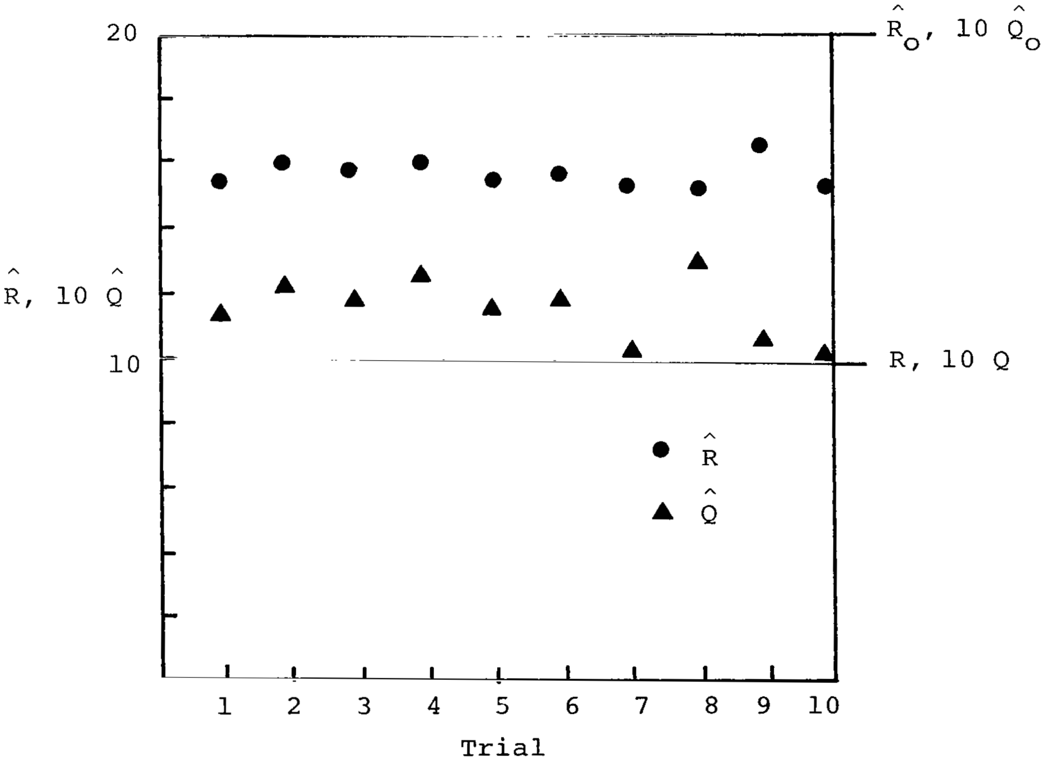

Fig. 6.16 Explicit Suboptimal Solution Run 11

Table 6.6 Monte Carlo Run 6: Explicit Suboptimal Solution

Sample	R	$\hat{R}$	$\varepsilon(\hat{R})$	$\sigma_{\hat{R}}$	$e_{\hat{R}}$
1	0.860	0.993	0.965	0.0252	+1.110
2	1.599	1.099	1.115	0.0403	-0.398
3	0.522	0.879	0.878	0.0164	+0.061
4	1.288	1.060	1.027	0.0315	+1.050
5	0.526	1.054	1.006	0.0291	+1.650
6	0.102	0.797	0.798	0.0084	-0.120
7	0.304	0.834	0.837	0.0123	-0.244
8	1.718	1.029	1.067	0.0355	-1.070
9	0.160	0.970	1.031	0.0317	-1.920
10	0.484	1.068	1.046	0.0332	+0.665
Average	0.756	0.978	0.977	0.0264	+0.079

Theoretical Mean Squared R Estimation Error: 0.294

Actual Mean Squared R Estimation Error: 0.296

Table 6.6 (Continued) Monte Carlo Run 6

Sample	Q	$\hat{Q}$	$\varepsilon(\hat{Q})$	$\sigma_{\hat{Q}}$	$e_{\hat{Q}}$
1	0.456	0.490	0.444	0.0405	+1.130
2	0.512	0.654	0.686	0.0648	-0.495
3	0.330	0.304	0.302	0.0263	+0.076
4	0.360	0.602	0.544	0.0504	+1.150
5	0.997	0.587	0.511	0.0472	+1.610
6	0.327	0.172	0.173	0.0134	-0.075
7	0.329	0.235	0.236	0.0196	-0.051
8	0.145	0.543	0.607	0.0568	-1.130
9	1.487	0.456	0.553	0.0514	-1.880
10	1.249	0.610	0.577	0.0538	+0.615
Average	0.619	0.465	0.463	0.0424	+0.095

Theoretical Mean Squared Q Estimation Error: 0.189

Actual Mean Squared Q Estimation Error: 0.191

Table 6.7 Monte Carlo Run 7: Explicit Suboptimal Solution

Sample	R	$\hat{R}$	$\varepsilon(\hat{R})$	$\sigma_{\hat{R}}$	$e_{\hat{R}}$
1	8.231	11.362	12.280	0.674	-1.360
2	8.104	9.237	9.997	0.447	-1.700
3	10.349	8.452	9.238	0.373	-2.105
4	5.362	7.442	7.419	0.190	+0.121
5	14.526	14.625	13.005	0.748	+2.160
6	11.753	10.269	10.534	0.510	-0.520
7	37.523	16.213	19.195	1.380	-2.160
8	29.982	16.658	16.856	1.140	-0.173
9	10.499	12.729	11.493	0.596	+2.080
10	19.663	12.767	12.637	0.716	+0.182
Average	15.599	11.975	12.166	0.671	-0.348

Theoretical Mean Squared R Estimation Error: 59.344

Actual Mean Squared R Estimation Error: 70.540

Table 6.7 (Continued) Monte Carlo Run 7

Sample	Q	$\hat{Q}$	$\varepsilon(\hat{Q})$	$\sigma_{\hat{Q}}$	$e_{\hat{Q}}$
1	3.985	1.285	1.452	0.576	-0.290
2	1.677	0.844	1.002	0.087	-0.182
3	0.089	0.707	0.849	0.072	-0.198
4	0.002	0.503	0.496	0.036	+0.193
5	2.476	1.895	1.588	0.146	+2.100
6	0.923	1.064	1.104	0.097	-0.410
7	0.614	2.183	2.783	0.266	-2.250
8	0.906	2.321	2.330	0.221	-0.046
9	2.361	1.556	1.294	0.117	+2.240
10	0.255	1.509	1.509	0.138	+0.000
Average	1.329	1.387	1.441	0.176	+0.116

Theoretical Mean Squared Q Estimation Error: 1.817

Actual Mean Squared Q Estimation Error: 1.566

Table 6.8 Monte Carlo Run 8: Explicit Suboptimal Solution

Sample	R	$\hat{R}$	$\varepsilon(\hat{R})$	$\sigma_{\hat{R}}$	$e_{\hat{R}}$
1	1.0	1.016	1.0	0.0286	+0.560
2	1.0	1.034	1.0	0.0286	+1.190
3	1.0	1.009	1.0	0.0286	+0.315
4	1.0	1.026	1.0	0.0286	+0.910
5	1.0	0.968	1.0	0.0286	-1.120
6	1.0	1.005	1.0	0.0286	+0.175
7	1.0	0.988	1.0	0.0286	-0.420
8	1.0	0.970	1.0	0.0286	-1.050
9	1.0	1.046	1.0	0.0286	+1.610
10	1.0	0.977	1.0	0.0286	-0.805
Average	1.0	1.004	1.0	0.0286	+0.137

Theoretical Mean Squared R Estimation Error: .000821

Actual Mean Squared R Estimation Error: .000694

Table 6.8 (Continued) Monte Carlo Run 8

Sample	Q	$\hat{Q}$	$\varepsilon(\hat{Q})$	$\sigma_{\hat{Q}}$	$e_{\hat{Q}}$
1	0.5	0.529	0.5	0.046	+0.630
2	0.5	0.559	0.5	0.046	+1.280
3	0.5	0.520	0.5	0.046	+0.435
4	0.5	0.546	0.5	0.046	+1.000
5	0.5	0.454	0.5	0.046	-1.000
6	0.5	0.513	0.5	0.046	+0.284
7	0.5	0.483	0.5	0.046	-0.370
8	0.5	0.442	0.5	0.046	-1.260
9	0.5	0.555	0.5	0.046	+1.200
10	0.5	0.458	0.5	0.046	-0.905
Average	0.5	0.506	0.5	0.046	+0.129

Theoretical Mean Squared Q Estimation Error: .002126

Actual Mean Squared Q Estimation Error: .001755

Table 6.9 Monte Carlo Run 9: Explicit Suboptimal Solution

Sample	R	$\hat{R}$	$\varepsilon(\hat{R})$	$\sigma_{\hat{R}}$	$e_{\hat{R}}$
1	10.0	10.218	10.0	0.447	+0.488
2	10.0	9.693	10.0	0.447	-0.687
3	10.0	10.013	10.0	0.447	+0.029
4	10.0	10.450	10.0	0.447	+1.010
5	10.0	10.355	10.0	0.447	+0.795
6	10.0	10.392	10.0	0.447	+0.875
7	10.0	10.600	10.0	0.447	+1.340
8	10.0	10.423	10.0	0.447	+0.950
9	10.0	9.850	10.0	0.447	-0.335
10	10.0	9.542	10.0	0.447	-1.020
Average	10.0	10.154	10.0	0.447	+0.365

Theoretical Mean Squared R Estimation Error: .201

Actual Mean Squared R Estimation Error: .139

Table 6.9 (Continued) Monte Carlo Run 9

Sample	Q	$\hat{Q}$	$\varepsilon(\hat{Q})$	$\sigma_{\hat{Q}}$	$e_{\hat{Q}}$
1	1.0	1.050	1.0	0.087	+0.575
2	1.0	0.935	1.0	0.087	-0.745
3	1.0	1.013	1.0	0.087	+0.150
4	1.0	1.085	1.0	0.087	+0.975
5	1.0	1.071	1.0	0.087	+0.815
6	1.0	1.069	1.0	0.087	+0.795
7	1.0	1.132	1.0	0.087	+1.520
8	1.0	1.076	1.0	0.087	+0.875
9	1.0	0.955	1.0	0.087	-0.517
10	1.0	0.902	1.0	0.087	-1.120
Average	1.0	1.029	1.0	0.087	+0.333

Theoretical Mean Squared Q Estimation Error: .00757

Actual Mean Squared Q Estimation Error: .00588

Table 6.10 Monte Carlo Run 10: Explicit Suboptimal Solution

Sample	R	$\hat{R}$	$\varepsilon(\hat{R})$	$\sigma_{\hat{R}}$	$e_{\hat{R}}$
1	0.859	1.708	1.680	0.0257	+1.090
2	1.599	1.819	1.831	0.0409	-0.294
3	0.533	1.593	1.592	0.0169	+0.059
4	1.288	1.775	1.743	0.0321	+1.000
5	0.526	1.770	1.721	0.0295	+1.660
6	0.102	1.511	1.512	0.0089	-0.112
7	0.304	1.547	1.551	0.0128	-0.031
8	1.718	1.746	1.783	0.0364	-1.020
9	0.160	1.685	1.745	0.0319	-1.880
10	0.484	1.785	1.761	0.0335	+0.715
Average	0.756	1.694	1.692	0.0269	+0.119

Theoretical Mean Squared R Estimation Error: 1.1203

Actual Mean Squared R Estimation Error: 1.1249

Table 6.10 (Continued) Monte Carlo Run 10

Sample	Q	$\hat{Q}$	$\varepsilon(\hat{Q})$	$\sigma_{\hat{Q}}$	$e_{\hat{Q}}$
1	0.456	0.531	0.485	0.0404	+1.140
2	0.512	0.695	0.726	0.0648	-0.480
3	0.330	0.346	0.343	0.0262	+0.115
4	0.360	0.643	0.585	0.0505	+1.150
5	0.997	0.628	0.552	0.0472	+1.610
6	0.327	0.214	0.215	0.0134	-0.075
7	0.329	0.276	0.277	0.0196	-0.051
8	0.145	0.585	0.648	0.0568	-1.110
9	1.487	0.498	0.594	0.0514	-1.860
10	1.249	0.652	0.618	0.0538	+0.633
Average	0.619	0.507	0.504	0.0427	+0.107

Theoretical Mean Squared Q Estimation Error: .1778

Actual Mean Squared Q Estimation Error: .1799

Table 6。11 Monte Carlo Run 11: Explicit Suboptimal Solution

Sample	R	$\hat{R}$	$\varepsilon(\hat{R})$	$\sigma_{\hat{R}}$	$e_{\hat{R}}$
1	10.0	15.697	15.543	0.087	+0.378
2	10.0	16.060	15.543	0.087	+1.010
3	10.0	15.729	15.543	0.087	+0.621
4	10.0	16.027	15.543	0.087	+1.340
5	10.0	15.363	15.543	0.087	-0.183
6	10.0	15.449	15.543	0.087	-0.046
7	10.0	15.150	15.543	0.087	-0.820
8	10.0	15.099	15.543	0.087	-0.172
9	10.0	16.248	15.543	0.087	+1.030
10	10.0	15.032	15.543	0.087	-1.150
Average	10.0	15.585	15。543	0.087	+0.201

Theoretical Mean Squared R Estimation Error: 30.931

Actual Mean Squared R Estimation Error: 31.366

Table 6.11 (Continued) Monte Carlo Run 11

Sample	Q	$\hat{Q}$	$\varepsilon(\hat{Q})$	$\sigma_{\hat{Q}}$	$e_{\hat{Q}}$
1	1.0	1.167	1.1342	0.450	+0.342
2	1.0	1.222	1.1342	0.450	+1.150
3	1.0	1.188	1.1342	0.450	+0.414
4	1.0	1.241	1.1342	0.450	+1.070
5	1.0	1.118	1.1342	0.450	-0.400
6	1.0	1.130	1.1342	0.450	-0.206
7	1.0	1.062	1.1342	0.450	-0.870
8	1.0	1.019	1.1342	0.450	-0.986
9	1.0	1.224	1.1342	0.450	+1.560
10	1.0	1.034	1.1342	0.450	-1.135
Average	1.0	1.141	1.1342	0.450	+0.120

Theoretical Mean Squared Q Estimation Error: .0256

Actual Mean Squared Q Estimation Error: .0257

The normalized differences between the R and Q estimates and their theoretical conditional means were studied using a procedure similar to that used in testing the normalized estimation errors of the maximum likelihood estimator. For each run presented, the mean and variance of these normalized differences across the ensemble of ten trials were computed. In most cases, the computed mean and variance of the differences were within one standard deviation of their expected values. From this it can reasonably be concluded that the theoretical expressions for the conditional mean and conditional variance of the estimate about the conditional mean are valid.

From the Monte Carlo runs presented, it can be seen that the theoretical results related to the maximum likelihood solution and the explicit suboptimal solution agree reasonably well with the actual results of the simulations. These theoretical results predict the ensemble averages of the estimation error and mean squared error. Therefore, to study the behavior of the various estimators, Monte Carlo simulations are not necessary. The following are the results of a statistical evaluation of the maximum likelihood and explicit suboptimal solutions.

It has been shown that the maximum likelihood estimator of the noise covariance parameters is unbiased for any values of R and Q that can be encountered. The conditional covariance of the estimates about the true values of R and Q was shown to be

$$\mathrm{cov}(\hat{R}, \hat{Q}|R, Q) = W_n(R,Q)$$

The conditional average of the estimation error is zero and is independent of the actual values of R and Q whereas the covariance of the estimation error is a strong function of R and Q.

Figure 6.17 shows the normalized variance of the R and Q estimator as a function of R for a fixed Q. Figure 6.18 shows the normalized variance as a function of Q for a fixed R. In both examples, the system and measurement schedule is that given previously.

Unlike the maximum likelihood estimator, the conditional average of the estimation error for the explicit suboptimal estimator is a strong function of the true and a priori values of R and Q. Figure 6.19 shows the variation of the conditional average of $\hat{R}_n$ and $\hat{Q}_n$ as a function of the a priori estimate $\hat{R}_o$, for a fixed $R = 1$, $Q = 0.5$, and $\hat{Q}_o = 0.5$. It can be seen that only when the a priori estimate of R is exactly equal to the true value of R are the conditional means of $\hat{R}_n$ and $\hat{Q}_n$ equal to the true values of R and Q. This means that if the a priori estimate of R is not equal to the true value of R, the explicit suboptimal estimators for R and Q are highly biased, with the amount of the bias obtained from this graph.

Figure 6.20 shows the variation in these conditional averages as a function of $\hat{Q}_o$, for a fixed $R = 1$, $Q = 0.5$, and $\hat{R}_o = 1$. The same general conclusions can be drawn from this graph concerning the bias of the estimator when the a priori value of Q is not equal to the true value.

The maximum likelihood estimator of the noise covariance parameters is slightly biased towards the a priori estimates of R and Q when the number of measurements is small. However, it was shown that as the number of measurements becomes large, the effect of this initial condition bias becomes arbitrarily small. The same is not true for the explicit suboptimal estimator. If the estimator is biased towards the initial estimates, this bias does not necessarily decrease as the number of measurements increases. The explicit estimator is often unable to distinguish between an error in R and an error in Q and resultingly the estimates of R and Q may be biased no matter how many measurements are taken.

As was mentioned in Chapter 4, just because the explicit suboptimal estimator is highly biased with respect to the a priori estimates of R and Q does not mean that no useful information can be obtained from them. In fact, the very fact that they are so highly biased if the a priori estimates are incorrect can be the basis for estimating the true values of R and Q. As will be shown, the variance of the estimators about the possibly biased values is quite small so that if the estimates obtained from the explicit suboptimal estimator differ by an appreciable amount from the a priori estimates, there is good justification for concluding that the a priori estimates are in error. Unfortunately, the explicit suboptimal estimators do not provide any information concerning how to correct the a priori values to make this discrepency smaller. In any actual situation, the a priori values of R and Q would have to be adjusted in a trial and error fashion to attempt to make the estimated values of R and Q equal to the a priori values. The values of $\hat{R}_o$ and $\hat{Q}_o$ which make $\hat{R}_n = \hat{R}_o$ and $\hat{Q}_n = \hat{Q}_o$ to the desired degree of accuracy are then the best estimates for R and Q.

In Section 4.4, expressions for the conditional mean squared estimation error of the estimators for R and Q were developed. As was mentioned, part of this error comes from possible bias in the estimator and part comes from possible variations about this bias. Figure 6.21 shows the variance of the estimator about the biased values as a function of the a priori estimates of $\hat{R}_o$ and $\hat{Q}_o$, for fixed values of R and Q. The bias can be found from the previous graphs 6.19 and 6.20. $\sigma^2_{\hat{Q}}(\hat{Q}_o)$ represents the variance of $\hat{Q}$ as a function of $\hat{Q}_o$ for a fixed $\hat{R}_o$, $\sigma^2_{\hat{Q}}(\hat{R}_o)$ represents the variance of $\hat{Q}$ as a function of $\hat{R}_o$ for a fixed $\hat{Q}_o$. Similarly, $\sigma^2_{\hat{R}}(\hat{R}_o)$ represents the variance of $\hat{R}$ as a function of $\hat{R}_o$ for a fixed $\hat{Q}_o$ and $\sigma^2_{\hat{R}}(\hat{Q}_o)$ represents the variance of $\hat{R}$ as a function of $\hat{Q}_o$ for a fixed $\hat{R}_o$.

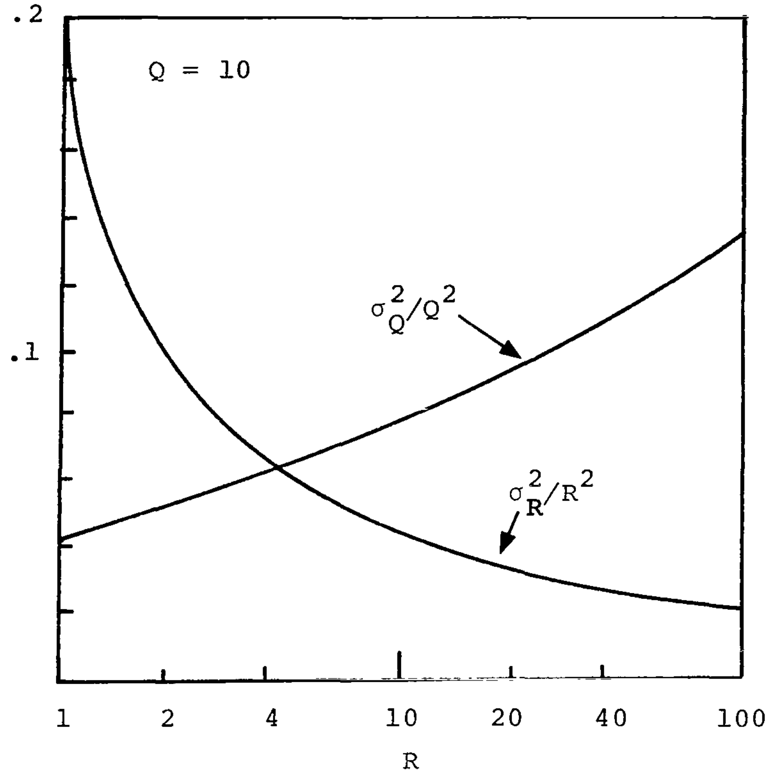

Fig. 6.17 Variance of M. L. E. vs R

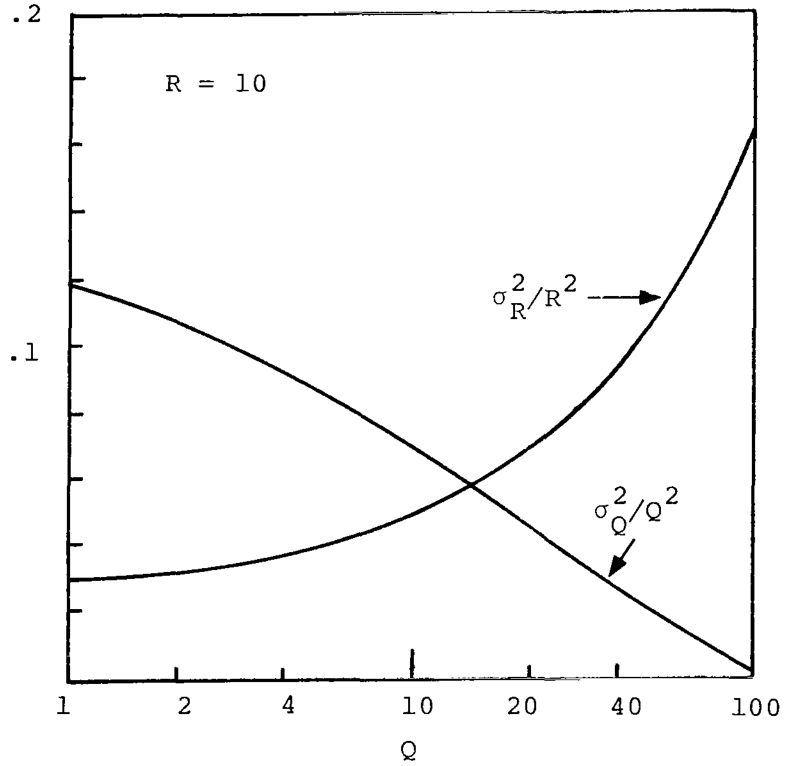

Fig. 6.18 Variance of M. L. E. vs Q

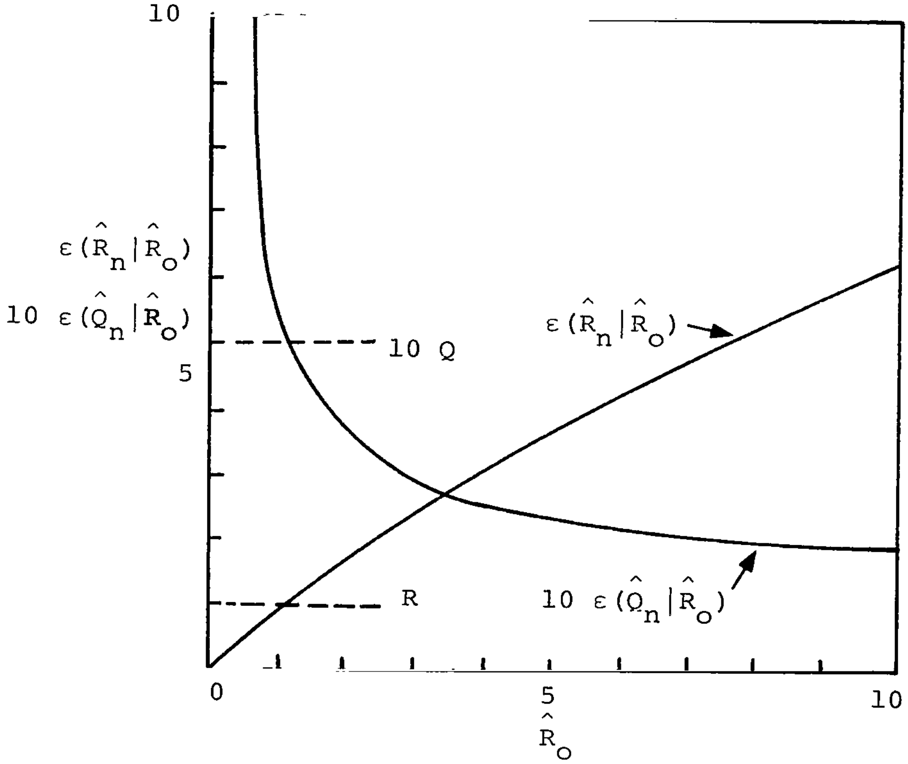

Fig. 6.19 Conditional Mean of Explicit Suboptimal Estimator vs $\hat{R}_o$

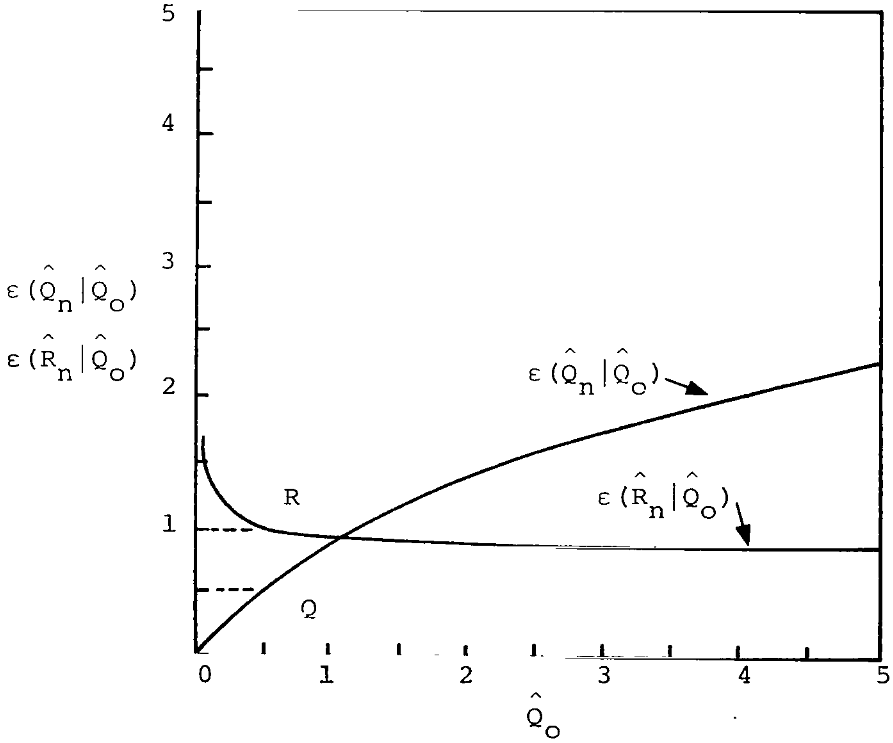

Fig. 6.20 Conditional Mean of Explicit Suboptimal Estimator vs $\hat{Q}_o$

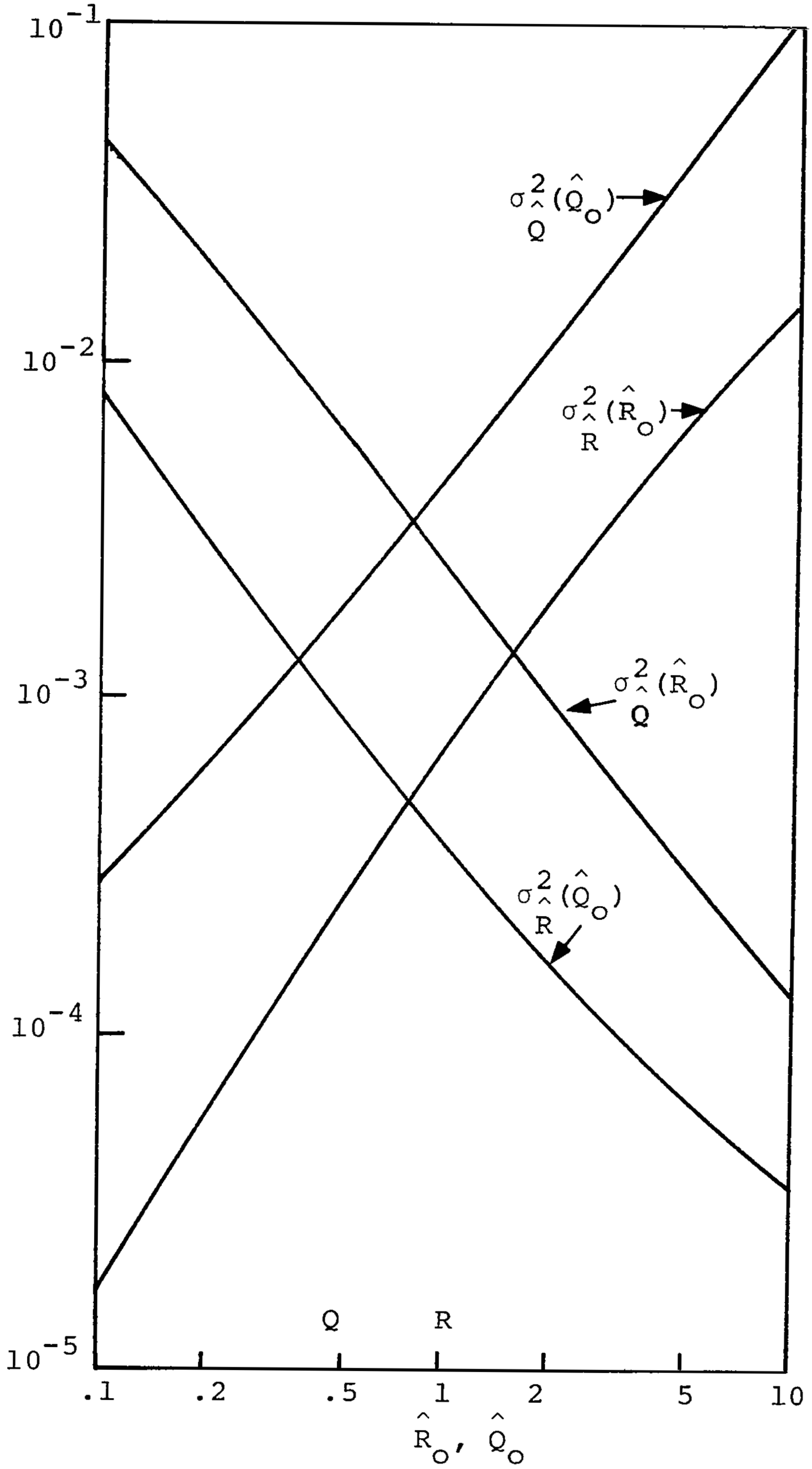

Fig. 6.21 Conditional Variance of Explicit Suboptimal Estimator about Bias Value

6.5 Testing of Statistical Hypotheses

In Chapter 5 various techniques for testing statistical hypotheses were described. In this section numerical results of a simulation of these hypothesis tests are described.

As mentioned in Section 5.7, the measurement residual $z_k - H_k\hat{x}_{k|k-1}$ is a good test variable upon which hypothesis tests can be conducted. If the values of R and Q that are used to compute the state estimate $\hat{x}_{k|k-1}$ are equal to the true values, then the measurement residual is a zero mean normally distributed random variable with covariance

$$\mathrm{cov}(\Delta z_k | R,Q) = R + H_k P_{k|k-1} H_k^T$$

where $\Delta z_k = z_k - H_k \hat{x}_{k|k-1}$

However, if the measurement and driving noises are not zero mean normal variables with known covariances, then the measurement residuals may not be zero mean with covariance as given above.

Two hypothesis tests were devised to test hypotheses on the values of R and Q used to compute the measurement residual gains and to test the hypothesis concerning the unbiasedness of the measurement and driving noises. The first of these two tests will now be described.

Suppose $\hat{R}_o$ and $\hat{Q}_o$ are the a priori estimates of the measurement and driving noise covariance matrices and a maximum likelihood estimate of the state is computed as a function of the measurements and these a priori values of

$\hat{R}_o$ and $\hat{Q}_o$. In Section 2.3, the recursive equations for computing the state estimate and its "computed" covariance matrix are given. It was shown that only under the null hypothesis $\hat{R}_o = R$ and $\hat{Q}_o = Q$ does the computed covariance matrix accurately represent the covariance of the estimation error. It was also shown that the measurement residual has a zero mean even under departures from the null hypothesis, but only if $\hat{R}_o = R$ and $\hat{Q}_o = Q$ are the residuals at a time k independent of the residuals at a time j, for $k \neq j$. Therefore, with $\Delta z_k^* = z_k - H_k \hat{x}^*_{k|k-1}$,

$$\varepsilon(\Delta z_k^*) = 0$$

$$\varepsilon(\Delta z_k^* \Delta z_k^{*T}) = R + H_k P_{k|k-1} H_k^T$$

where the above conditional expected values are conditioned upon the fact that $\hat{R}_o$ and $\hat{Q}_o$ are used to compute the weighting matrices for the measurements, whereas the true values are R and Q. $P_{k|k-1}$ is the "true" state estimation error covariance and is not equal to the "computed" error covariance matrix except under the null hypothesis.

Under the null hypothesis,

$$\varepsilon(\Delta z_k^* \Delta z_j^{*T}) = 0 \qquad \text{for } k \neq j$$

Consider the variable

$$t_n = \frac{1}{\sqrt{n}} \sum_{k=1}^{n} \sqrt{B_k^{*-1}} \; (z_k - H_k \hat{x}_{k|k-1}^{*})$$

where $$B_k^{*} = \hat{R}_o + H_k P_{k|k-1}^{*} H_k^T$$

Then under the null hypothesis, t_n is a zero mean normal random variable with covariance

$$\text{cov}(t_n) = \varepsilon(t_n t_n^T) = \frac{1}{n} \sum_{j=1}^{n} \sum_{k=1}^{n} \sqrt{B_k^{*-1}} \; \varepsilon(\Delta z_k^{*} \; \Delta z_j^{*T}) \sqrt{B_j^{*-1}}$$

$$= \frac{1}{n} \sum_{k=1}^{n} \sqrt{B_k^{*-1}} \; B_k^{*} \; \sqrt{B_k^{*-1}}$$

$$= I$$

Therefore, t_n is a zero mean normal variable with covariance I. Since each component of the vector t_n is statistically independent of any other component, an independent test of each component is possible. Using the procedures of Section 5.4 concerning tests on the mean, a critical region $(-t_\alpha, t_\alpha)$ can be defined such that under the null hypothesis, the probability of the test variable t_n being in this region is $1 - \alpha$, where α is the level of significance of the test. Using the procedures outlined in Section 5.4, the test variable t_n can be used to test the hypothesis that the residual is zero mean with covariance I. A failure in this test can be caused by a bias in the measurement or driving noises or incorrect values of R and Q used to compute B_k^* which is used to normalize the residuals Δz_k^*.

Now consider the test variable

$$\chi_n^2 = \sum_{k=1}^{n} \sqrt{B_k^{*-1}}\ \Delta z_k^*\ \Delta z_k^{*T}\ \sqrt{B_k^{*-1}}$$

Under the null hypothesis

$$\varepsilon(\chi_n^2) = \sum_{k=1}^{n} \sqrt{B_k^{*-1}}\ \varepsilon(\Delta z_k^*\ \Delta z_k^{*T})\ \sqrt{B_k^{*-1}}$$

$$= n\ I$$

Since Δz_k^* are normally distributed random variables, the diagonal elements of χ_n^2 can be shown to be independent chi-square distributed varialbes under the null hypothesis, with n-1 degrees of freedom. Using the procedures outlined in Section 5.5 concerning tests on the variance, a critical region for each diagonal element of χ_n^2 can be defined such that under the null hypothesis the probability that the test variable lies within this critical region is $1 - \alpha$, where α is the level of significance of the test. The test variable χ_n^2 can then be used to test the hypothesis that the residuals are zero mean normally distributed random variables with covariance B_k^*. A failure of this test can be caused by a bias in the measurement or driving noises or incorrect values of R and Q used to compute B_k^*.

Table 6.12 shows the results of such tests of hypotheses on the values of R and Q. Shown are the true values of R and Q along with the a priori values of R and Q which were used to compute the proper test variables. The two columns, M-fail and S-fail, indicate whether or not the mean and variance test variables failed the appropriate test on the 5, 10, and 20 percent levels. A "1" indicates a test failure and a "0" indicates a passing test.

As can be seen, the mean test was not highly sensitive to departures from the null hypothesis on the values of R and Q. Only in extreme cases did the mean test fail, and then only on the 10 and 20 percent levels.

However, as would be desired, the variance test was very sensitive to moderate departures from the null hypothesis, thus indicating a powerful test of the hypothesis.

Another series of hypothesis tests was conducted to see if the above hypothesis tests could detect a bias in either the measurement or driving noises. In Chapter 2, it was shown that maximum likelihood state estimation can be adversely affected if it is assumed that the measurement and driving noises are zero mean, when in fact they are not zero mean. For this test, it was assumed that the measurement and driving noise covariance matrices were precisely known, but there was a bias in either of the two noises. In other words, hypotheses on the means of the measurement and driving noises are being tested. The results of these tests are shown in Table 6.13. B_V is the actual measurement noise bias and $\hat{B}_V$ is the hypothesized value of the measurement noise bias. B_W is the true driving noise bias and $\hat{B}_W$ is the hypothesized value of the driving noise bias. The system and measurement schedule are those given previously.

The variances of the measurement and driving noises about any possible biases were 10 and 1 respectively. It would be expected that only when the biases are comparable to the standard deviation of the noises would the tests indicate a failure. This was indeed the case. As can be seen, the mean test was somewhat more powerful in detecting departures from the hypothesis about the noise biases, but because of the non-independence of the tests, the variance test also indicated failure if the difference between the true bias and the hypothesized bias was sufficiently large.

These hypothesis test runs are not meant to be all inclusive but rather indicate that with only a moderate expenditure of computation, powerful tests on hypotheses concerning the unbiasedness and covariance of the measurement and driving noises can be implemented. The tests do not tell why the particular test failed, but they do indicate that one or more of the underlying assumptions about the system or measurements is in error. The tests might also be used to test hypotheses concerning the values of certain elements of the transition matrices, measurement matrices H_k, or any other parameter which is used to describe the system. These runs are merely meant to test the feasibility of using hypothesis tests in real time to indicate a failure of certain assumptions about the environment under which the estimation process is taking place.

Table 6.12 Hypothesis Test Run 1: R and Q Test

R	$\hat{R}$	Q	$\hat{Q}$	M-fail			S-fail		
				5	10	20	5	10	20
10	1	1	1	0	0	0	1	1	1
10	5	1	1	0	0	0	1	1	1
10	10	1	1	0	0	0	0	0	0
10	20	1	1	0	0	0	1	1	1
10	100	1	1	0	1	1	1	1	1
10	10	1	0.1	0	1	1	1	1	1
10	10	1	0.5	0	0	0	0	0	1
10	10	1	1	0	0	0	0	0	0
10	10	1	2	0	0	0	1	1	1
10	10	1	10	0	0	0	1	1	1

Table 6.13 Hypothesis Test Run 2: Noise Bias Test

B_V	$\hat{B}_V$	B_W	$\hat{B}_W$	M-fail			S-fail		
				5	10	20	5	10	20
10	0	1	1	0	0	0	0	0	0
10	5	1	1	0	0	0	0	0	0
10	10	1	1	0	0	0	0	0	0
10	20	1	1	1	1	1	0	0	0
10	100	1	1	1	1	1	1	1	1
10	10	1	0.1	1	1	1	0	0	0
10	10	1	0.5	1	1	1	1	1	1
10	10	1	1	0	0	0	0	0	0
10	10	1	2	1	1	1	1	1	1
10	10	1	10	1	1	1	1	1	1

Chapter 7

CONCLUSION

7.1 Summary of Results

The technique of maximum likelihood estimation has been shown to be effective in estimating the state and statistics of the measurement and driving noises in a linear dynamical system observed by linear noisy measurements. Theoretical and empirical results indicate that the estimator of the noise covariance parameters is asymptotically unique, unbiased, consistent, and efficient. However, the solution of the likelihood equations for the state and noise statistics generally requires considerably more computation than that normally involved in estimating the state of the system when the noise statistics are presumed known. For this reason the optimal procedures requiring an iterative solution of the likelihood equations will probably find their greatest application in data reduction rather than real time estimation problems.

In many cases, a linearized solution of the likelihood equations is quite adequate and can be used if a real time solution of the estimation problem is required. Of the suboptimal techniques studied, the linearized maximum likelihood solution is the only generally applicable technique that is effective for the real time estimation of the state and noise covariance parameters. The other techniques for estimating the noise covariance parameters are either biased with respect to the initial estimates of the parameters or result in possibly non-unique solutions.

Any technique for the estimation of the noise covariance parameters requires some additional computation. Therefore, before any estimation of these quantities is undertaken, there should be some indication that the a priori values are sufficiently in error to substantially reduce the effectiveness of the state estimation procedure. It has been shown that there are several techniques for testing certain hypotheses concerning the values of the noise statistics which allow a decision to be made concerning the correctness of the a priori estimates of these parameters.

7.2 Suggestions for Future Study

In Chapters 3 and 4 techniques for the estimation of noise covariance parameters were developed under the assumptions that the measurement and driving noises were independent zero mean normally distributed random variables with diagonal, time invariant covariance matrices. These assumptions were made to simplify the estimation problem while not overly restricting the applicability of the solution. However, the techniques discussed can be extended to include cases when these assumptions are not valid. A similar structure of the problem must be retained so that definitive results can be obtained. That is, the dynamics of the state are still described by a linear differential or difference equation with normally distributed driving noise and the measurements are linear functions of the state with additive normally distributed measurement noise. In this section, the following cases will be briefly studied:

1) non-diagonal noise covariance matrices
2) time varying noise covariance matrices
3) estimation of more general parameters, such as elements of the state transition matrix

Possible extensions of the explicit suboptimal estimator will be discussed first.

The extension of the explicit suboptimal estimator to the case of non-diagonal noise covariance matrices is straightforward. In the expressions of Chapter 4 for estimators of the diagonal elements $\hat{R}_n^{jj}$ and $\hat{Q}_n^{jj}$, all that need be done is change the indices to $\hat{R}_n^{jk}$ and $\hat{Q}_n^{jk}$ with appropriate changes in the indices appearing on the right hand side of these equations. The expressions for the conditional and unconditional means of the estimators can easily be modified to include this generalization. However, extension of the expressions for the conditional and unconditional mean squared error of the estimators would be exceedingly difficult.

The case of time varying noise covariance parameters is considerably more difficult to treat. If it is assumed that R and Q vary slowly with time compared to the rate of data accumulation, then the total estimation time can be broken into segments and an independent estimate of R and Q obtained from the data gathered in each time segment. Alternatively, a different weighting of the measurement data could be proposed such that data taken in the distant past is essentially not used in the estimate of the covariance parameters. A third procedure could be used to model the noise covariance parameters in a way described by Smith and outlined in Chapter 4. In that case, the noise covariance parameters are assumed to be of the form

$$R_n = k\ R_{nom_n}$$

where R_{nom_n} is some nominal value of the measurement noise covariance at time n and k is an unknown time invariant precision factor associated with R_n. A similar equation could be used for the driving noise covariance matrices. The estimation problem is then reduced to estimating certain constants associated with each unknown noise covariance parameter. Any more general time variation of the noise covariance parameters than those outlined above cannot be adequately treated using the explicit suboptimal estimator.

There is no real possibility that the explicit estimator can be used to estimate more general parameters of the system or measurement. The estimation equations were derived with the particular goal of estimating the measurement and driving noise covariance matrices and accordingly cannot be modified to include the estimation of other system parameters.

The technique of maximum likelihood offers a procedure and formalism within which any of the extensions mentioned above can be handled. The resulting equations may be so complicated that a solution may not be practical, but at least the equations for an optimal solution of the problem can be derived.

If the measurement and driving noise covariance matrices are not assumed to be diagonal, then additional likelihood equations must be derived for estimating these off diagonal elements. This can be done quite easily. In addition to the likelihood equations for the state and diagonal elements of R and Q, one additional equation is needed for each off diagonal element of R and Q that is to be estimated. This equation has the form

$$\left(\frac{\partial L_n(R,Q,x_n,Z_n)}{\partial R^{jk}}\right)_{\substack{R\rightarrow\hat{R}_n \\ Q\rightarrow\hat{Q}_n \\ x_n\rightarrow\hat{x}_{n|n}}} = 0$$

with a similar equation for the off diagonal elements of Q. $L_n(R,Q,x_n,Z_n)$ is the logarithm of the appropriate likelihood function as derived in Chapter 3. The choice of likelihood functions is determined by whether a priori information about the noise covariance parameters is to be utilized.

As was the case of the explicit suboptimal estimator, the case of time varying noise covariance parameters is more difficult to treat. Again if it is assumed that the time variation is slow, then the total estimation time can be divided into segments and an independent estimate of the noise covariance parameters obtained assuming that R and Q are essentially constant over this time segment. Of course, the time segment over which R and Q are constant must be long enough to allow sufficient information to be gathered to obtain a reasonably good estimate of R and Q.

Alternatively, if the noise covariance parameters are assumed to change with time in a deterministic manner as proposed by Smith, the technique of maximum likelihood can easily be applied to estimate the value of the unknown precision factors k^j, with a separate precision factor associated with each unknown element of R and Q. In this case, the likelihood function can be thought to be a function of the parameters k^j rather than R_n and Q_n. For each k^j, an equation of the following form must be solved.

$$\left(\frac{\partial L_n(k^1,..,k^m,x_n,Z_n)}{\partial k^j} \right)_{\substack{x_n \to \hat{x}_{n|n} \\ k \to \hat{k}}} = 0$$

where here it has been assumed that there are m such precision factors. The solution of this equation can be obtained in a manner entirely analogous to the solution for the time invariant covariance parameters discussed in Chapter 3. There is also the likelihood equation associated with the state x_n which must be solved simultaneously with the likelihood equations for the parameters k^j.

Much more work needs to be done in the area of maximum likelihood estimation when the time variation of R and Q is more complicated than the cases given above. It is felt that there is much promise of obtaining an optimal solution to the problem. Such a solution might proceed along the following lines.

Let ξ_n represent the vector of diagonal elements of the noise covariance matrices at time n.

$$\xi_n^T = (R_n^{11},..,R_n^{\gamma\gamma},Q_n^{11},..,Q_n^{\eta\eta})$$

Let ξ_{n-1} be the vector of diagonal elements of these matrices at time n-1, and let the relationship between ξ_n and ξ_{n-1} be given by

$$\xi_n = \Psi(n,n-1)\xi_{n-1} + u_n \tag{7.2.1}$$

where $\Psi(n,n-1)$ is the "noise covariance parameter transition matrix" and u_n is the "noise covariance parameter driving noise." Since ξ_n represents a vector of noise variances, every element of ξ_n must be positive. Therefore, the distribution of the noise u_n must be chosen so that for any $\Psi(n,n-1)$ and ξ_{n-1}, the elements of ξ_n are all positive. In general this would require that the distribution of u_n be a function of $\Psi(n,n-1)$ and ξ_{n-1}. However, these problems can be avoided if it is assumed that the elements of u_n are chosen from a distribution that is independent of $\Psi(n,n-1)$ and ξ_{n-1} and only allows positive values for the elements of u_n. Such a distribution might be the Gamma distribution used in Chapter 3. Note that this choice allows ξ_n to decrease as well as increase from time n-1 to time n. If on the average ξ_n

is to be equal to ξ_{n-1}, then the parameters $\Psi(n,n-1)$ and the distribution of u_n should be chosen so that

$$(I - \Psi(n,n-1))\overline{\xi}_n = \overline{u}_n$$

where $\overline{u}_n$ is the mean of the u_n distribution and $\overline{\xi}_n$ is the average value of ξ_n. If the noise u_n on a given trial is less than its mean, then ξ_n will be less than its mean value.

It might also be reasonably assumed that $\Psi(n,n-1)$ is a diagonal matrix so that if the elements of u_n and ξ_{n-1} are mutually independent, the elements of ξ_n will also be independent.

If $\Psi(n,n-1)$ is a zero matrix, then ξ_n is completely independent of ξ_{n-1}, whereas if u_n is not present, ξ_n is a deterministic function of ξ_{n-1}. All other cases between these two extremes can be handled by appropriate choice of $\Psi(n,n-1)$ and the parameters of the distribution of u_n. It can be shown that if ξ_{n-1} and u_n have a Gamma distribution and are mutually independent, then ξ_n has a Gamma distribution.

It is desired to estimate the values of ξ_n and x_n given the measurements Z_n. The appropriate likelihood function to maximize would be

$$l(\xi_n,x_n,Z_n) = f(\xi_n,x_n|Z_n) \qquad (7.2.2)$$

where $f(\xi_n,x_n|Z_n)$ is the conditional probability density function of ξ_n and x_n given the measurements Z_n. The choice of the proper likelihood function is not as obvious in the case of time varying noise statistics as was the case when the statistics were assumed to be time invariant. Two other possibilities will be discussed subsequently.

From Bayes' rule,

$$f(\xi_n, x_n | Z_n) = \frac{f(\xi_n, x_n, z_n, Z_{n-1})}{f(z_n, Z_{n-1})} \qquad (7.2.3)$$

$$= \frac{f(z_n | Z_{n-1}, x_n, \xi_n)\; f(\xi_n, x_n | Z_{n-1})}{f(z_n | Z_{n-1})}$$

where $f(z_n | Z_{n-1}) = \iint_{-\infty}^{\infty} \frac{f(z_n, Z_{n-1}, x_n, \xi_n)}{f(Z_{n-1})}\, d\xi_n\, dx_n$

$$= \iint_{-\infty}^{\infty} f(z_n | Z_{n-1}, x_n, \xi_n)\; f(x_n, \xi_n | Z_{n-1})\; d\xi_n\, dx_n \qquad (7.2.4)$$

Then the logarithm of the likelihood function is

$$L_n(\xi_n, x_n, Z_n) = \ln f(\xi_n, x_n | Z_n)$$

$$= \ln f(\xi_n, x_n | Z_{n-1}) + \ln f(z_n | Z_{n-1}, x_n, \xi_n) \qquad (7.2.5)$$

$$- \ln f(z_n | Z_{n-1})$$

The gradient of L_n with respect to the parameters to be estimated is then

$$\frac{\partial L_n}{\partial \alpha_n} = \frac{\partial \ln f(\xi_n, x_n | Z_{n-1})}{\partial \alpha_n} + \frac{\partial \ln f(z_n | Z_{n-1}, x_n, \xi_n)}{\partial \alpha_n} \qquad (7.2.6)$$

where $\alpha_n^T = (x_n^T, \xi_n^T)$.

The density function $f(z_n|Z_{n-1},x_n,\xi_n)$ is easily found.

$$f(z_n|Z_{n-1},x_n,\xi_n) = \frac{1}{(2\pi)^{\gamma/2}|R_n|^{1/2}} e^{-\frac{1}{2}(z_n-H_nx_n)^T R_n^{-1}(z_n-H_nx_n)} \tag{7.2.7}$$

The real difficulty comes in finding the density function $f(\xi_n,x_n|Z_{n-1})$. It can be obtained, at least in theory, from previously obtained density functions. If it is assumed that initially ξ_n and x_n are independent, then

$$f(\xi_o,x_o) = f(x_o)\ f(\xi_o)$$

where $f(x_o)$ is the a priori probability density function of the initial state and $f(\xi_o)$ is the a priori probability density of the initial value of ξ_o, both of which are presumed to be known. Then before the first measurement,

$$f(\xi_1,x_1) = f(x_1|\xi_1)\ f(\xi_1)$$

Assuming $f(x_o)$ is a normal density function with mean $x_{o|o}$ and covariance $P_{o|o}$, it is easy to show that

$$f(x_1|\xi_1) = \frac{1}{(2\pi)^{\beta/2}|P_{1|o}|^{1/2}} e^{-\frac{1}{2}(x_1-\hat{x}_{1|o})^T P_{1|o}^{-1}(x_1-\hat{x}_{1|o})} \tag{7.2.8}$$

where $$\hat{x}_{1|o} = \Phi(1,0)\hat{x}_{o|o}$$

$$P_{1|o} = \Phi(1,0)\ P_{o|o}\ \Phi^T(1,0) + \Gamma_1\ Q_1\ \Gamma_1^T$$

Assuming the model for ξ_n previously given, $f(\xi_1)$ is a Gamma probability density function with known parameters. Then using (7.2.3) with (7.2.4), (7.2.7), and (7.2.8), the density function $f(\xi_1, x_1 | Z_1)$ can be found. In most cases, evaluation of this density function will be very complicated but it can be performed in theory.

Once the necessary density functions in (7.2.5) are found, then the estimates of ξ_n and x_n can be found by finding the zero points of the likelihood equations (7.2.6). Some iterative procedure will be needed for the solution of these equations.

Assuming that a solution of the likelihood equations can be found, much work needs to be done to determine if such a solution is unique and if it is, what are its asymptotic properties. The situation is much more complicated than the case when the noise covariance matrices were assumed to be time invariant. If the noise covariance parameters change rapidly with time and are not sufficiently correlated with past values of the noise parameters, then there may not be sufficient information in the measurements to uniquely define the estimates. In such a situation, the maximum likelihood estimator may be required to estimate the value of the noise parameters essentially on the information contained in a single measurement. If the measurement is of small dimension compared with the number of parameters being estimated, there may be insufficient information in the measurement to estimate the noise parameters. This is not a shortcoming of this particular type of estimation, but rather a fundamental problem of trying to estimate the value of a quantity with insufficient information. A similar problem was encountered in Chapter 2 when the state of the system was being estimated without prior information. Until sufficient information was gathered, a unique state estimate could not be defined.

Assuming that a unique solution to the problem exists, finding its asymptotic properties will be difficult. However, it should not be expected that the estimator for ξ_n is a consistent estimator when there is "noise" driving the vector of noise covariance parameters. This is entirely analogous to the fact that a Kalman estimator for the state is not consistent when there is noise driving the state, or in other words, the covariance of the estimation error does not go to zero as the number of measurements goes to infinity. Therefore, it can be anticipated that the maximum likelihood estimator of the state which uses estimates of ξ_n to compute the appropriate filter gains will not converge to the estimates that would be obtained if the noise covariance parameters were known precisely. However, if the noise covariance estimator operates properly, this difference may be small.

As was mentioned previously, the likelihood function given above is not the only possibility that might be considered. Another solution to the problem might be found by simultaneously estimating the state x_n and the values of the noise covariances at all times up to and including time n. In such a situation, a likelihood function of the form

$$l(\xi_1,..,\xi_n,x_n,Z_n) = f(\xi_1,..,\xi_n,x_n|Z_n) \qquad (7.2.9)$$

might be chosen. Define

$$\Omega_n^T = (\xi_1^T,...,\xi_n^T)$$

Then by Bayes' rule

$$f(\Omega_n,x_n|Z_n) = f(x_n|\Omega_n,Z_n)\ f(\Omega_n|Z_n) \qquad (7.2.10)$$

$f(x_n|\Omega_n,Z_n)$ is the probability density function of the state given the measurements Z_n and the values of the noise covariance parameters at all times. From Chapter 2,

$$f(x_n|\Omega_n,Z_n) = \frac{1}{(2\pi)^{\beta/2}|P_{n|n}|^{1/2}} e^{-\frac{1}{2}(x_n-\hat{x}_{n|n})^T P_{n|n}^{-1}(x_n-\hat{x}_{n|n})} \qquad (7.2.11)$$

where $\hat{x}_{n|n}$ is a function of Z_n and Ω_n, and $P_{n|n}$ is a function of Ω_n.

By application of Bayes' rule,

$$f(\Omega_n|Z_n) = \frac{f(\xi_n,\Omega_{n-1},z_n,Z_{n-1})}{f(z_n,Z_{n-1})}$$

$$= \frac{f(z_n|\Omega_n, Z_{n-1})\ f(\Omega_n, Z_{n-1})}{f(z_n|Z_{n-1})\ f(Z_{n-1})}$$

$$= \frac{f(z_n|\Omega_n, Z_{n-1})\ f(\xi_n|\Omega_{n-1}, Z_{n-1})}{f(z_n|Z_{n-1})}\ f(\Omega_{n-1}|Z_{n-1})$$

Repeating the above procedure, it is easy to show that

$$f(\Omega_n|Z_n) = f(\xi_o) \prod_{i=1}^{n} \frac{f(z_i|\Omega_i, Z_{i-1})\ f(\xi_i|\Omega_{i-1}, Z_{i-1})}{f(z_i|Z_{i-1})}$$

$f(\xi_o)$ is the a priori probability density function of the initial value of the noise covariance parameters. It is easy to show that

$$f(z_i|\Omega_i, Z_{i-1}) = \frac{1}{(2\pi)^{\gamma/2}|B_i|^{1/2}}\ e^{-\frac{1}{2}(z_i - H_i\hat{x}_{i|i-1})^T B_i^{-1}(z_i - H_i\hat{x}_{i|i-1})} \qquad (7.2.12)$$

where $$B_i = R_i + H_i P_{i|i-1} H_i^T$$

$\hat{x}_{i|i-1}$ is the maximum likelihood estimate of x_i after i-1 measurements using the true values of Ω_i to compute the proper filter gains, and $P_{i|i-1}$ is the conditional covariance of x_i about $\hat{x}_{i|i-1}$.

From the model of ξ_i, it can be seen that $f(\xi_i|\Omega_{i-1}, Z_{i-1})$ is a Gamma probability density function with conditional mean

$$\varepsilon(\xi_i|\Omega_{i-1}, Z_{i-1}) = \Psi(i, i-1)\ \xi_{i-1} + \bar{u}_i$$

where $\bar{u}_i$ is the mean of the distribution of noise covariance parameter driving noise. The conditional covariance of the distribution is

$$\mathrm{cov}(\xi_i | \Omega_{i-1}, Z_{i-1}) = \mathrm{cov}(u_i)$$

In obtaining these expressions, use was made of the fact that u_i is independent of Z_{i-1} and Ω_{i-1}. $f(z_i | Z_{i-1})$ need not be evaluated since it is not a function of x_n or Ω_n.

The logarithm of the likelihood function (7.2.9) is

$$L_n(\Omega_n, x_n, Z_n) = \ln f(x_n | \Omega_n, Z_n) + \ln f(\Omega_n | Z_n) \qquad (7.2.13)$$

and the gradient of L_n with respect to the parameters to be estimated is

$$\frac{\partial L_n}{\partial \alpha_n} = \frac{\partial \ln f(x_n | \Omega_n, Z_n)}{\partial \alpha_n} + \frac{\partial \ln f(\Omega_n | Z_n)}{\partial \alpha_n} \qquad (7.2.14)$$

where now $\alpha_n^T = (x_n^T, \Omega_n^T)$. The maximum likelihood estimate of α_n is the value of α_n which makes all components of (7.2.14) zero. From an examination of (7.2.14), it can be seen that the estimate of the state x_n is just the maximum likelihood state estimate which uses the estimates of Ω_n to compute the proper filter gains. The estimates of Ω_n are found from the solution of the likelihood equations associated with the gradient of the likelihood function with respect to Ω_n.

It should be noted that finding the necessary density functions in this likelihood equation is considerably easier than finding the density functions in the previous likelihood equations (7.2.6). However, it should also be noted that the number of likelihood equations that must be simultaneously solved is much larger than in the previous case. In addition to the likelihood equation associated with the state, there is one likelihood equation associated with the value of ξ_i at every measurement time. Thus as n becomes large, the number of likelihood equations also becomes large.

A third possibility for likelihood function would be

$$l(x_n, Z_n) = f(x_n | Z_n) \qquad (7.2.15)$$

In this case, only the state is to be estimated, not the values of the noise covariance parameters. However, it will be shown that finding the above probability density function is even more difficult than in the previous two cases. From Bayes' rule,

$$f(x_n | Z_n) = \frac{f(x_n, Z_n)}{f(Z_n)}$$

$$= \int_{-\infty}^{\infty}\!\!\cdots\!\int \frac{f(x_n, Z_n, \Omega_n)}{f(Z_n)}\, d\Omega_n$$

$$= \int_{-\infty}^{\infty}\!\!\cdot\!\int f(x_n | Z_n, \Omega_n)\, f(\Omega_n | Z_n)\, d\Omega_n$$

Then the gradient of the logarithm of the likelihood function (7.2.15) is

$$\frac{\partial \ln f(x_n|Z_n)}{\partial x_n} = \frac{1}{f(x_n|Z_n)} \int_{-\infty}^{\infty}\!\!\cdot\cdot\!\!\int \frac{\partial f(x_n|Z_n,\Omega_n)}{\partial x_n} f(\Omega_n|Z_n)\, d\Omega_n$$

Evaluating this expression in a realistic situation would be very complicated and finding the zero points of the equation would be even more involved. Thus the number of equations that must be solved has been reduced over that of the two previous approaches, but the complexity of the equations is considerably increased.

In general, the state estimates obtained from the solution of these three different likelihood functions will be different. Which estimate is "better" depends upon what information is desired from the measurement information. Solution of the first problem will result in estimates of the current state and the current value of the noise covariance parameters. The solution of the second problem will result in estimates of the current state and the values of the noise covariances at all times. The solution of the third problem will result only in the estimate of the state, with no information provided about the value of the noise statistics.

The tradeoff between the number of equations to be solved and their complexity seems to be a general feature of maximum likelihood estimation. As the number of parameters to be estimated increases, there are more equations that must be simultaneously solved, but it is usually easier to find the necessary probability density functions.

Maximum likelihood estimation can be used to estimate more general parameters of the system and measurement than the statistics of the noises. These problems can be handled within the framework of the maximum likelihood estimators already developed. The likelihood functions of Chapter 3 were written as functions of the state x_n, the measurements Z_n, and the parameters R and Q. In fact, the likelihood function is a function of all parameters of the system, namely the state transition matrix, forcing function matrix, and the observation matrix. The dependency of the likelihood function on these additional parameters was not indicated because it was previously assumed that the parameters were known precisely. Now it is assumed that some of these parameters are not known precisely a priori, but rather knowledge of them is described by some a priori probability density function in a fashion similar to that used in describing the uncertainty in R and Q.

Let β represent the vector of any additional parameters of the problem that are to be estimated. For simplicity it is assumed that β is time invariant. The likelihood function appropriate for this problem is

$$l(R,Q,x_n,Z_n\beta) = f(R,Q,x_n,\beta|Z_n) \qquad (7.2.16)$$

where $f(R,Q,x_n,\beta|Z_n)$ is the joint conditional probability density function of the parameters R,Q,x_n, and β given the measurements Z_n. From Bayes' rule

$$f(R,Q,x_n,\beta|Z_n) = f(x_n|Z_n,R,Q,\beta)\ f(R,Q,\beta|Z_n) \qquad (7.2.17)$$

Then the logarithm of the likelihood function is

$$L_n(R,Q,x_n,\beta,Z_n) = \ln f(R,Q,x_n,\beta|Z_n) \qquad (7.2.18)$$

$$= \ln f(x_n|Z_n,R,Q,\beta) + \ln f(R,Q,\beta|Z_n)$$

Then

$$\frac{\partial L_n}{\partial x_n} = \frac{\partial \ln f(x_n|Z_n,R,Q,\beta)}{\partial x_n} \qquad (7.2.19)$$

But the zeros of the likelihood equations (7.2.19) can be shown to occur when

$$x_n \rightarrow \hat{x}_{n|n}(\hat{R}_n,\hat{Q}_n,\hat{\beta}_n,Z_n)$$

This says that the estimate of x_n is just the maximum likelihood estimator of the state that uses estimates of R, Q, and β to compute the proper filter gains. Estimates of R and Q are found in the same manner as in Chapter 3. Estimates of β are found from the solution of the additional likelihood equations

$$\left(\frac{\partial L_n}{\partial \beta}\right)_{\substack{x_n \rightarrow \hat{x}_{n|n} \\ R \rightarrow \hat{R}_n \\ Q \rightarrow \hat{Q}_n \\ \beta \rightarrow \hat{\beta}_n}} = 0$$

The likelihood equations for the state x_n, the noise covariance parameters R and Q, and the additional parameters β

must be solved simultaneously, this generally requiring an iterative solution.

Thus it can be seen that more general parameters can be estimated in the same way as the noise covariance parameters, except that for each additional parameter to be estimated, an additional likelihood equation must be solved.

As in the case of time varying noise statistics, much work needs to be done concerning the asymptotic properties and possible convergence problems associated with the estimation of these additional parameters.

One final word should be said about the application of maximum likelihood estimators of the state and noise statistics in problems when there may be errors in the dynamical model of the state. Jazwinski has shown that the effects of these modeling errors can often be characterized as an additional noise driving the state, where the statistics of this noise are unknown. If a maximum likelihood estimator of the mean and covariance of the "effective driving noise" is employed, there is good reason to believe that the performance of the state estimator can be considerably improved. In such cases, the estimates of the statistics of the noise may have little physical significance, since there is actually no "modeling error noise" driving the state. However, if the effect of the modeling errors can be accurately represented as such a noise, then estimating the statistics of this noise can improve the state estimation and minimize possible divergence problems within the filter.

Appendix A

MATRIX AND VECTOR OPERATIONS

A.1 The Generalized Inverse

The generalized inverse is an important concept in matrix theory because it provides an extension of the concept of an inverse which applies to all matrices. Deutsch, Rao, and Rust discuss the theory and application of the generalized inverse in such problem areas as numerical analysis and least squares estimation. This appendix closely follows the work of Deutsch.

The generalized inverse of an m x n matrix A of rank r is a n x m matrix $A^{\#}$ of rank r such that

$$A\ A^{\#}\ A = A \qquad (A.1.1)$$

If $A = 0$, define $0^{\#} = 0^{T}$. Both $A^{\#}A$ and $A\ A^{\#}$ are idempotent because they are equal to their squares.

$$(A^{\#}A)^{2} = A^{\#}A\ A^{\#}A = A^{\#}A$$

$$(A\ A^{\#})^{2} = A\ A^{\#}A\ A^{\#} = A\ A^{\#}$$

If A is of rank $r > 0$, then it has a rank factorization of the form

$$A = B\ C$$

where B is a m x r matrix and C is an r x n matrix with the rank of both B and C equal to r.

The pseudoinverse of a matrix, often called the Moore-Penrose generalized inverse, is defined as

$$A^+ = C^T(C\ C^T)^{-1}\ (B^TB)^{-1}\ B^T \qquad (A.1.2)$$

with $0^+ = 0^T$

A pseudoinverse is a generalized inverse because (A.1.2) can be shown to satisfy (A.1.1). If A is nonsingular, then $A^+ = A^\# = A^{-1}$.

There are several advantages for employing the pseudoinverse rather than the more inclusive generalized inverse. These stem from the following properties:

1) The pseudoinverse of a pseudoinverse yields the original matrix. That is $(A^+)^+ = A$
2) $(A\ A^+)$ and (A^+A) are symmetric matrices.
3) The pseudoinverse of a matrix is unique.

Rust discusses an algorithm suitable for digital computer operation for finding the generalized inverse of a matrix. However, in certain special cases, the solution can be obtained directly.

If (A^TA) is of full rank then

$$A^+ = (A^TA)^{-1}A^T$$

If $(A\ A^T)$ is of full rank then

$$A^+ = A^T(A\ A^T)^{-1}$$

A.2 The Matrix Inversion Lemma

If A_1 is a n x n nonsingular matrix, A_2 is a n x m matrix, A_3 is a m x m nonsingular matrix, and A_4 is a m x n matrix, then

$$(A_1 + A_2A_3A_4)^{-1} = A_1^{-1} - A_1^{-1}A_2(A_4A_1^{-1}A_2 + A_3^{-1})^{-1}\ A_4A_1^{-1}$$

The proof is by direct substitution.

A.3 Matrix and Vector Derivatives

Certain matrix and vector identities are needed in the main text. The purpose of this appendix is to derive the general results applied there. The following notation is used here:

x	a n x 1 column vector
$L(x)$	a scalar function of the vector x and possibly other parameters
Y	a m x m matrix
Y^{-1}	the inverse of Y
$\|Y\|$	the determinant of Y
U	the cofactor matrix of Y such that $Y^{-1} = U^T/\|Y\|$, where U^T is the transpose of U.

$$\frac{\partial L(x)}{\partial x} = \left[\frac{\partial L}{\partial x_1}, \frac{\partial L}{\partial x_2}, \ldots, \frac{\partial L}{\partial x_n}\right]$$ a 1 x n row vector

$$\frac{\partial^2 L(x)}{\partial x \partial x} = \begin{bmatrix} \frac{\partial^2 L}{\partial x_1 \partial x_1} & \frac{\partial^2 L}{\partial x_2 \partial x_1} & \cdots & \frac{\partial^2 L}{\partial x_n \partial x_1} \\ \vdots & & & \\ \frac{\partial^2 L}{\partial x_1 \partial x_n} & \frac{\partial^2 L}{\partial x_2 \partial x_n} & \cdots & \frac{\partial^2 L}{\partial x_n \partial x_n} \end{bmatrix}$$ a n x n symmetric matrix

1. $\frac{1}{|Y|} \frac{\partial |Y|}{\partial Y} = (Y^{-1})^T$

Proof: By the cofactor expansion of the determinant of Y,

$$|Y| = \sum_{k=1}^{m} Y_{ik} U_{ik} \qquad \text{for any } i$$

Then $$\frac{\partial |Y|}{\partial Y_{j\ell}} = \sum_{k=1}^{m} \frac{\partial Y_{ik}}{\partial Y_{j\ell}} U_{ik} + Y_{ik} \frac{\partial U_{ik}}{\partial Y_{j\ell}}$$

By definition $$U_{ik} = (-1)^{i+k} M_{ik}$$

where M_{ik} is the minor of Y_{ik} which is found by evaluating the determinant of the matrix obtained by deleting the row and column containing the element Y_{ik}. Thus from this definition, U_{ik} is not a function of Y_{ik} and

$$\frac{\partial U_{ik}}{\partial Y_{ik}} = 0$$

So $$\frac{\partial|Y|}{\partial Y_{j\ell}} = \sum_{k=1}^{m} \frac{\partial Y_{ik}}{\partial Y_{j\ell}} U_{ik} = \sum_{k=1}^{m} \delta_{ij}\,\delta_{k\ell}\,U_{ik} = U_{j\ell}$$

where δ_{ij} is the Kronecker delta defined by

$$\begin{aligned} \delta_{ij} &= 0 \qquad i \neq j \\ &= 1 \qquad i = j \end{aligned}$$

Then $$\frac{1}{|Y|}\frac{\partial|Y|}{\partial Y_{j\ell}} = \frac{1}{|Y|} U_{j\ell} = (Y^{-1})_{\ell j}$$

Or $$\frac{1}{|Y|}\frac{\partial|Y|}{\partial Y} = (Y^{-1})^T$$

2. By entirely analogous procedures it can be shown that

$$\frac{1}{|Y+B|}\frac{\partial|Y+B|}{\partial Y} = [(Y+B)^{-1}]^T$$

where B is any matrix that is not a function of Y.

3. Let Y be a function of a matrix Z. Then

$$\begin{aligned} \frac{1}{|Y|}\frac{\partial|Y|}{\partial Z_{j\ell}} &= \sum_i \sum_k \frac{1}{|Y|}\frac{\partial|Y|}{\partial Y_{ik}}\frac{\partial Y_{ik}}{\partial Z_{j\ell}} \\ &= \sum_i \sum_k (Y^{-1})_{ki}\frac{\partial Y_{ik}}{\partial Z_{j\ell}} \\ &= \mathrm{Tr}(Y^{-1}\frac{\partial Y}{\partial Z_{j\ell}}) \end{aligned}$$

where Tr( ) is the trace of the enclosed matrix.

4. By analogous procedures it can be shown that

$$\frac{1}{|Y+B|}\,\frac{\partial |Y+B|}{\partial Z_{j\ell}} = \mathrm{Tr}\left[(Y+B)^{-1}\,\frac{\partial (Y+B)}{\partial Z_{j\ell}}\right]$$

where both Y and B may be functions of Z.

5. Let a and b be any constant m x 1 column vectors. Then

$$\frac{\partial (a^T Y\ b)}{\partial Y_{j\ell}} = \sum_i \sum_k \ a_i \frac{\partial Y_{ik}}{\partial Y_j} b_k = \sum_i \sum_k \ a_i\ b_k\ \delta_{k\ell}\ \delta_{ij}$$

$$= a_j b_\ell$$

Or $\dfrac{\partial (a^T Y\ b)}{\partial Y} = a\ b^T$

6. From the fact that $Y\ Y^{-1} = I$, it can be shown that

$$\frac{\partial Y}{\partial Z_{j\ell}}\ Y^{-1} + Y\ \frac{\partial Y^{-1}}{\partial Z_{j\ell}} = 0$$

Therefore $\dfrac{\partial Y^{-1}}{\partial Z_{j\ell}} = -\ Y^{-1}\ \dfrac{\partial Y}{\partial Z_{j\ell}}\ Y^{-1}$

7. $$\frac{\partial (a^T Y^{-1} b)}{\partial Z_{j\ell}} = -\ a^T\ Y^{-1}\ \frac{\partial Y}{\partial Z_{j\ell}}\ Y^{-1} b$$

If Y = A Z B + C, where A, B, and C are constant matrices,

$$\frac{\partial (a^T Y^{-1} b)}{\partial Z_{j\ell}} = -\ a^T\ Y^{-1} A\ \frac{\partial Z}{\partial Z_{j\ell}}\ B\ Y^{-1} b$$

$$= - \sum_i \sum_k (A^T Y^{-1T} a)_i \frac{\partial Z_{ik}}{\partial Z_{j\ell}} (B\ Y^{-1} b)_k$$

$$= - \sum_i \sum_k (A^T Y^{-1T} a)_i\ \delta_{ij}\ \delta_{k\ell}\ (B\ Y^{-1} b)_k$$

$$= - (A^T Y^{-1T} a\ b^T\ Y^{-1T} B^T)_{j\ell}$$

Or $$\frac{\partial (a^T\ Y^{-1} b)}{\partial Z} = - A^T Y^{-1T} a\ b^T Y^{-1T} B^T$$

Appendix B

EVALUATION OF EXPLICIT ESTIMATOR MEAN SQUARED ERROR

In Section 4.4 expressions for explicit estimators of the diagonal elements of the measurement and driving noise covariance matrices were developed. Also evaluated were the conditional and unconditional means of the estimates and the conditional mean of the squared estimation error. In this appendix the unconditional mean of the squared estimation error is obtained.

From (4.4.16) the conditional mean of the squared R estimation error is

$$\varepsilon[(\hat{R}^{jj}_n - R^{jj})^2] = G^{jj}_n + (F^{jj}_n)^2 \qquad (B.1)$$

where $$G^{jj}_n = \left(\frac{n-1}{n}\right)^2 G^{jj}_{n-1} + \frac{2}{n^2}((R + \Delta F_n - H_n P^*_{n|n} H^T_n)^{jj})^2 \qquad (B.2)$$

$$F_n = \frac{1}{n} \sum_{k=1}^{n} \Delta F_k \qquad (B.3)$$

$$\Delta F_n = H_k (P_{k|k} + P^*_{k|k}) H^T_k - H_k A^*_k R - R A^{*T}_k H^T_k \qquad (B.4)$$

The unconditional mean of the squared estimation error is then

$$E[(\hat{R}^{jj}_n - R^{jj})^2] = E(G^{jj}_n) + E[(F^{jj}_n)^2] \qquad (B.5)$$

But $$E(G_n^{jj}) = \left(\frac{n-1}{n}\right)^2 E(G_{n-1}^{jj}) + \frac{2}{n^2} E[((R + \Delta F_n - H_n P^*_{n|n} H_n^T)^{jj})^2]$$

Define $$a^{*j}_{n|n} = (H_n P^*_{n|n} H_n^T)^{jj} = h_n^{jT} P^*_{n|n} h_n^j$$

where h_n^j is the j^{th} column of H_n^T.

Then $$E[((R + \Delta F_n - H_n P^*_{n|n} H_n^T)^{jj})^2] = \overline{(R^{jj})^2} + E[(\Delta F_n^{jj})^2] \qquad \text{(B.6)}$$

$$+ (a^{*j}_{n|n})^2 + 2\,E(R^{jj} \Delta F_n^{jj}) - 2\,E(\Delta F_n^{jj}) a^{*j}_{n|n} - 2\,\overline{R^{jj}}\, a^{*j}_{n|n}$$

where $$\overline{R^{jj}} = E(R^{jj}) \text{ and } \overline{(R^{jj})^2} = E[(R^{jj})^2]$$

It is assumed that the a priori values of R and Q used to compute all starred quantities are equal to the means of their respective distributions. Or

$$\hat{R}_o = \overline{R} \text{ and } \hat{Q}_o = \overline{Q}$$

Then using the results of Section 2.3 it can be shown that

$$P^*_{n|n} = E(P_{n|n}) \triangleq \overline{P}_{n|n}$$

and also $$E(\Delta F_n) = 0$$

Define $$a^j_{n|n} = (H_n P_{n|n} H_n^T)^{jj} = h_n^{jT} P_{n|n} h_n^j$$

Then $$\Delta F_n^{jj} = a^j_{n|n} + a^{*j}_{n|n} - 2\, a^{*j}_{n|n} \frac{R^{jj}}{\overline{R^{jj}}}$$

and $$E[(\Delta F_n^{jj})^2] = E([a^j_{n|n})^2] - (a^{*j}_{n|n})^2$$

$$- \frac{4\, a^{*j}_{n|n}}{\overline{R^{jj}}} \left[E(a^j_{n|n} R^{jj}) - a^{*j}_{n|n} \overline{R^{jj}} - \textstyle\sum_R^{jj} \frac{a^{*j}_{n|n}}{\overline{R^{jj}}} \right]$$

where $$\textstyle\sum_R^{jk} = E[(R^{jj} - \overline{R^{jj}})^2]\delta_{jk} \quad \text{a diagonal matrix}$$

Define $$Y^j_{n|n} = E(P_{n|n} R^{jj}) - \overline{P}_{n|n} \overline{R^{jj}}$$

$$d^j_n = h^{jT}_n Y^j_{n|n} h^j_n$$

Then $$E[(\Delta F_n^{jj})^2] = E(a^j_{n|n})^2 - (a^{*j}_{n|n})^2 - \left[\frac{4\, a^{*j}_{n|n}}{\overline{R^{jj}}} \, d^j_n - \textstyle\sum_R^{jj} \frac{a^{*j}_{n|n}}{\overline{R^{jj}}} \right] \tag{B.7}$$

$E[(a^j_{n|n})^2]$ and d^j_n must now be computed. Using (2.3.43) and (2.3.44) it can be shown that

$$P_{n|n} = \lambda_{n|o} P_{o|o} \lambda^T_{n|o} + \sum_{k=1}^{n} \lambda_{n|k} (A^*_k R A^{*T}_k + D_k \Gamma_k Q \Gamma^T_k D^T_k) \lambda^T_{n|k}$$

where $$D_k \triangleq (I - A^*_k H_k)$$

$$\lambda_{n|k} \triangleq D_n \; \Phi(n,n-1) D_{n-1} \; \Phi(n-1,n-2) \ldots D_{k+1} \; \Phi(k+1,k)$$

with $$\lambda_{k|k} \triangleq I$$

Then $$h_n^{jT} P_{n|n} h_n^j = h_n^{jT} \lambda_{n|o} P_{o|o} \lambda_{n|o}^T h_n^j + \sum_{k=1}^{n} h_n^{jT} \lambda_{n|k} A_k^{*} R A_k^{*T} \lambda_{n|k}^T h_n^j$$

$$+ \sum_{k=1}^{n} h_n^{jT} \lambda_{n|k} D_k \Gamma_k Q \Gamma_k^T \lambda_{n|k}^T h_n^j$$

But $$(A_k^{*T} \lambda_{n|k}^T h_n^j)^{\ell} = (A_k^{*T} \lambda_{n|k}^T H_n^T)^{\ell j}$$

$$(\Gamma_k^T D_k^T \lambda_{n|k}^T h_n^j)^{\ell} = (\Gamma_k^T D_k^T \lambda_{n|k}^T H_n^T)^{\ell j}$$

So $$h_n^{jT} \lambda_{n|k} A_k^{*} R A_k^{*T} \lambda_{n|k}^T h_n^j = (H_n \lambda_{n|k} A_k^{*})^{js} \; (R)^{s\ell} \; (A_k^{*T} \lambda_{n|k}^T H_n^T)^{\ell j}$$

$$= ((H_n \lambda_{n|k} A_k^{*})^{j\ell})^2 \; (R)^{\ell\ell}$$

Similarly

$$h_n^{jT} \lambda_{n|k} D_k \Gamma_k Q \Gamma_k^T D_k^T \lambda_{n|k}^T h_n^j = ((\Gamma_k^T D_k^T \lambda_{n|k}^T H_n^T)^{j\ell})^2 \; (Q)^{\ell\ell}$$

Define $$(C''_{n|1})^{j\ell} = \sum_{k=1}^{n} ((H_n \lambda_{n|k} A_k^{*})^{j\ell})^2$$ a γ x γ matrix

$$(L_{n|1})^{j\ell} = \sum_{k=1}^{n} ((H_n \lambda_{n|k} D_k \Gamma_k)^{j\ell})^2$$ a γ x η matrix

$$e_{n|o}^j = h_n^{jT} \lambda_{n|o} P_{o|o} \lambda_{n|o}^T h_n^j$$ a scalar

$$r^j = (R)^{jj}$$ a γ x 1 vector

$$q^j = (Q)^{jj}$$ a η x 1 vector

Then $$h_n^{jT} P_{n|n} h_n^j = e_{n|o}^j + (C''_{n|1} \, r + L_{n|1} \, q)^j \qquad \text{(B.8)}$$

Squaring (B.8), performing the unconditional expected value, then subtracting the square of $a^{*j}_{n|n}$,

$$E[(a^{j}_{n|n})^{2}] - (a^{*j}_{n|n})^{2} = (C''_{n|1} \; \Sigma_{R} \; C''^{T}_{n|1} + L_{n|1} \; \Sigma_{Q} \; L^{T}_{n|1})^{jj}$$

where $\quad \Sigma^{jk}_{Q} = E[(Q^{jj} - \overline{Q^{jj}})^{2}]\,\delta_{jk}$ $\quad$ a diagonal matrix

In obtaining this expression it was assumed that R and Q are independent random variables and that $e^{j}_{n|o}$ is not a function of R or Q.

By a similar procedure it is easy to show that

$$d^{j}_{n} = (C''_{n|1})^{jj} \; \Sigma^{jj}_{R}$$

Define $$(C'_{n|1})^{jk} = (C''_{n|1})^{jk} - 2\,\frac{a^{*j}_{n|n}}{\overline{R^{jj}}}\,\delta_{jk}$$

Then (B.7) becomes

$$E[(\Delta F^{jj}_{n})^{2}] = (C'_{n|1} \; \Sigma_{R} \; C'^{T}_{n|1} + L_{n|1} \; \Sigma_{Q} \; L^{T}_{n|1})^{jj}$$

It can also be shown that

$$E(R^{jj} \Delta F^{jj}_{n}) = (C'_{n|1})^{jj} \; \Sigma^{jj}_{R}$$

So, after algebraic manipulation, (B.6) becomes

$$E[((R + \Delta F_n - H_n P^*_{n|n} H_n^T)^{jj})^2] = (\bar{R}^{jj} - a^{*j}_{n|n})^2 + (C_{n|1} \textstyle\sum_R C^T_{n|1} + L_{n|1} \sum_Q L^T_{n|1})^{jj}$$

where $$C_{n|1} = C'_{n|1} + I$$

So $$E(G_n^{jj}) = \left(\frac{n-1}{n}\right)^2 E(G^{jj}_{n-1}) + \frac{2}{n^2}[(\bar{R}^{jj} - a^{*j}_{n|n})^2 + (C_{n|1} \textstyle\sum_R C^T_{n|1} + L_{n|1} \sum_Q L^T_{n|1})^{jj}] \qquad \text{(B.9)}$$

Evaluation of $E[(F_n^{jj})^2]$ is considerably more difficult than evaluation of $E(G_n^{jj})$ but uses a similar procedure. It can be seen that

$$F_n^{jj} = \frac{n-1}{n} F^{jj}_{n-1} + \frac{1}{n} \Delta F_n^{jj}$$

so $$(F_n^{jj})^2 = \left(\frac{n-1}{n}\right)^2 (F^{jj}_{n-1})^2 + \frac{1}{n^2} (\Delta F_n^{jj})^2 + \frac{2}{n^2} \sum_{k=1}^{n-1} \Delta F_k^{jj} \, \Delta F_n^{jj}$$

After a slight rearranging of terms and performing the above sum to n instead of n-1, it can be seen that

$$E[(F_n^{jj})^2] = \frac{n-1}{n}^{2} E[(F^{jj}_{n-1})^2] + \frac{1}{n^2}\left[2 \sum_{k=1}^{n} E(\Delta F_k^{jj} \Delta F_n^{jj}) - E[(\Delta F_n^{jj})^2]\right] \qquad \text{(B.10)}$$

After algebraic manipulation,

$$E(\Delta F_k^{jj}\Delta F_n^{jj}) = E(a^j_{k|k}\ a^j_{n|n}) - a^{*j}_{k|k}\ a^{*j}_{n|n} - \frac{2\ a^{*j}_{n|n}}{\overline{R}^{jj}}\left[d^j_k - \sum_R^{jj} \frac{a^{*j}_{k|k}}{\overline{R}^{jj}}\right]$$

$$- \frac{2\ a^{*j}_{k|k}}{\overline{R}^{jj}}\left[d^j_n - \sum_R^{jj} \frac{a^{*j}_{n|n}}{\overline{R}^{jj}}\right]$$

Following the same procedure as in finding $E[(\Delta F_n^{jj})^2]$,

$$E(\Delta F_k^{jj}\Delta F_n^{jj}) = (C'_{k|1}\ \sum_R\ C'^T_{n|1} + L_{k|1}\ \sum_Q\ L^T_{n|1})^{jj}$$

Define
$$\overline{C}'_{n|1} = \frac{1}{n}\sum_{k=1}^{n} C'_{k|1}$$

$$\overline{L}_{n|1} = \frac{1}{n}\sum_{k=1}^{n} L_{k|1}$$

Then after algebraic manipulation, (B.10) becomes

$$E[(F_n^{jj})^2] = \left(\frac{n-1}{n}\right)^2 E[(F^{jj}_{n-1})^2] + \frac{1}{n^2}[(2\ n\ C'_{n|1}\ \sum_R \overline{C}'^T_{n|1} - C'_{n|1}\sum_R C'^T_{n|1})^{jj}$$

$$+ (2\ n\ L_{n|1}\ \sum_Q\ \overline{L}^T_{n|1} - L_{n|1}\ \sum_Q\ L^T_{n|1})^{jj}] \qquad (B.11)$$

$C''_{n|1}$ and $L_{n|1}$ can be computed through a recursive relationship. From the definition of $C''_{n|1}$,

$$(C''_{n|1})^{j\ell} = \sum_{k=1}^{n} ((H_n\lambda_{n|k}A^*_k)^{j\ell})^2 = \sum_{k=1}^{n} ((h_n^{jT}\lambda_{n|k}A^*_k)^{\ell})^2$$

$$= \sum_{k=1}^{n} (h_n^{jT}\lambda_{n|k}A^*_k)^{\ell}\ (A_k^{*T}\lambda^T_{n|k}h^j_n)^{\ell}$$

$$= (h_n^j)^m \sum_{k=1}^{n} (\lambda_{n|k} A_k^*)^{m\ell} (A_k^{*T} \lambda_{n|k}^T)^{\ell u} (h_n^j)^u$$

Define $$(\alpha_n)^{m\ell u} = \sum_{k=1}^{n} (\lambda_{n|k} A_k^*)^{m\ell} (A_k^{*T} \lambda_{n|k}^T)^{\ell u}$$

Then $$(C''_{n|1})^{j\ell} = (h_n^j)^m (\alpha_n)^{m\ell u} (h_n^j)^u \qquad (B.12)$$

But $$(\alpha_n)^{m\ell u} = \sum_{k=1}^{n-1} (\lambda_{n|k} A_k^*)^{m\ell} (A_k^{*T} \lambda_{n|k}^T)^{\ell u} + (A_n^*)^{m\ell} (A_n^{*T})^{\ell u}$$

and $$\lambda_{n|k} = D_n \Phi(n,n-1) \lambda_{n-1|k}$$

so $$(\alpha_n)^{m\ell u} = (D_n \Phi(n,n-1))^{ms} (\alpha_{n-1})^{s\ell t} (\Phi^T(n,n-1) D_n^T)^{tu}$$

$$+ (A_n^*)^{m\ell} (A_n^{*T})^{\ell u}$$

Therefore, $(\alpha_n)^{m\ell u}$ can be computed recursively and $C''_{n|1}$ found from (B.12). $L_{n|1}$ can be computed in a similar fashion.

Define

$$(\beta_n)^{m\ell u} = \sum_{k=1}^{n} (\lambda_{n|k} D_k \Gamma_k)^{m\ell} (\Gamma_k^T D_k^T \lambda_{n|k}^T)^{\ell u}$$

Then $$(L_{n|1})^{j\ell} = (h_n^j)^m (\beta_n)^{m\ell u} (h_n^j)^u \qquad (B.13)$$

and $$(\beta_n)^{m\ell u} = (D_n\ \Phi(n,n-1))^{ms}\ (\beta_{n-1})^{s\ell t}\ (\Phi^T(n,n-1)D_n^T)^{tu}$$

$$+\ (D_n\Gamma_n)^{m\ell}\ (\Gamma_n^T D_n^T)^{\ell u}$$

From (4.4.21) the conditional mean of the squared Q estimation error is

$$\varepsilon[(\hat{Q}_n^{jj} - Q^{jj})^2] = J_n^{jj} + (M_n^{jj})^2 \qquad \text{(B.14)}$$

where $$J_n^{jj} = \left(\frac{n-1}{n}\right)^2 J_{n-1}^{jj} + \frac{2}{n^2}\ ((Q + \Delta M_n - T_n^*)^{jj})^2 \qquad \text{(B.15)}$$

$$T_n^* = \Gamma_n^{\#}(P_{n|n}^* - U_n^*)\Gamma_n^{\#T}$$

$$U_n^* = \Phi(n,n-1)P_{n-1|n-1}^*\ \Phi^T(n,n-1)$$

$\Gamma_n^{\#}$ is the generalized inverse of Γ_n, which in most cases is equal to

$$(\Gamma_n^T\Gamma_n)^{-1}\ \Gamma_n^T$$

$$\Delta M_n = \Gamma_n^{\#}(P_{n|n} + P_{n|n}^* - P_{n|n}^* P_{n|n-1}^{*-1}\ P_{n|n-1} - P_{n|n-1}\ P_{n|n-1}^{*-1}\ P_{n|n}^*$$

$$+\ U_n - U_n^*)\Gamma_n^{\#T}$$

$$U_n = \Phi(n,n-1) P_{n-1|n-1} \Phi^T(n,n-1)$$

$$M_n = \frac{1}{n} \sum_{k=1}^{n} \Delta M_k$$

The unconditional mean of the squared estimation error is then

$$E[(\hat{Q}_n^{jj} - Q^{jj})^2] = E(J_n^{jj}) + E[(M_n^{jj})^2]$$

But $E(J_n^{jj}) = \left(\frac{n-1}{n}\right)^2 E(J_{n-1}^{jj}) + \frac{2}{n^2} E[((Q + \Delta M_n - T_n^*)^{jj})^2]$ (B.16)

$$E[((Q + \Delta M_n - T_n^*)^{jj})^2] = \overline{(Q^{jj})^2} + E[(\Delta M_n^{jj})^2] + (T_n^{*jj})^2 \quad \text{(B.17)}$$

$$+ 2\, E(Q^{jj}\ \Delta M_n^{jj}) - 2\, E(\Delta M_n^{jj})\ T_n^{*jj} - 2\, \bar{Q}^{jj}\ T_n^{*jj}$$

If $\hat{R}_o = \bar{R}$ and $\hat{Q}_o = \bar{Q}$, then $E(\Delta M_n) = 0$ and

$$E[((Q + \Delta M_n - T_n^*)^{jj})^2] = \overline{(Q^{jj})^2} + E[(\Delta M_n^{jj})^2] + (T_n^{*jj})^2$$

$$+ 2\, E(Q^{jj}\ \Delta M_n^{jj}) - 2\, \bar{Q}^{jj}\ T_n^{*jj}$$

After some manipulation, ΔM_n can be expressed in the following form.

$$\Delta M_n = \bar{Q} - Q + f_n[(R - \bar{R}) + H_n(P_{n|n-1} - P^*_{n|n-1})H_n^T] f_n^T$$

where $f_n = \Gamma_n^{\#} A_n^{*}$ $\qquad$ a η x γ matrix

Define $f_n^j \triangleq$ the j^{th} column of f_n^T $\qquad$ a γ x 1 vector

$$g_n^j \triangleq H_n^T \, f_n^j \qquad \text{a } \beta \text{ x 1 vector}$$

$$b^j_{n|n-1} = g_n^{jT} P_{n|n-1} g_n^j \qquad \text{a scalar}$$

$$b^{*j}_{n|n-1} = g_n^{jT} P^*_{n|n-1} g_n^j \qquad \text{a scalar}$$

$$m_n^j = \begin{bmatrix} ((f_n^j)^1)^2 \\ ((f_n^j)^2)^2 \\ \cdot \\ \cdot \\ \cdot \\ ((f_n^j)^\gamma)^2 \end{bmatrix} \qquad \text{a } \gamma \text{ x 1 vector}$$

Then $\Delta M_n^{jj} = \bar{Q}^{jj} - Q^{jj} + m_n^{jT}(r - \bar{r}) + b^j_{n|n-1} - b^{*j}_{n|n-1}$

where as before r and $\bar{r}$ are γ x 1 vectors composed of the diagonal elements of R and $\bar{R}$ respectively. Then, after some algebraic manipulation, it can be shown that

$$E[(\Delta M_n^{jj})^2] = \sum\nolimits_Q^{jj} + m_n^{jT} \sum\nolimits_R m_n^j + E[(b^j_{n|n-1})^2] - (b^{*j}_{n|n-1})^2$$

$$-2[E(Q^{jj} b^j_{n|n-1}) - \bar{Q}^{jj} b^{*j}_{n|n-1}] + 2m_n^{jT}[E(r\, b^j_{n|n-1}) - \bar{r}\, b^{*j}_{n|n-1}]$$

Now $E[(b^j_{n|n-1})^2]$, $E(Q^{jj}b^j_{n|n-1})$ and $E(r\ b^j_{n|n-1})$ must be found.

From (2.3.43)

$$P_{n|n} = D_n P_{n|n-1} D_n^T + A_n^* R A_n^{*T}$$

so
$$P_{n|n-1} = D_n^{-1}(P_{n|n} - A_n^* R A_n^{*T}) D_n^{T-1}$$

and
$$b^j_{n|n-1} = g_n^{jT} D_n^{-1}(P_{n|n} - A_n^* R A_n^{*T}) D_n^{T-1} g_n^j$$

Or
$$b^j_{n|n-1} = g_n^j D_n^{-1} \lambda_{n|o} P_{o|o} \lambda_{n|o}^T D_n^{T-1} g_n^j$$

$$+ \sum_{k=1}^{n} g_n^{jT} D_n^{-1} \lambda_{n|k} (A_k^* R A_k^{*T} + D_k \Gamma_k Q \Gamma_k^T D_k^T) \lambda_{n|k}^T D_n^{T-1} g_n^j$$

$$- g_n^{jT} D_n^{-1} A_n^* R A_n^{*T} D_n^{T-1} g_n^j$$

Define
$$(U''_{n|1})^{j\ell} = \sum_{k=1}^{n} ((g_n^{jT} D_n^{-1} \lambda_{n|k} A_k^*)^{\ell})^2 - ((g_n^{jT} D_n^{-1} A_n^*)^{\ell})^2$$

$$(W'_{n|1})^{j\ell} = \sum_{k=1}^{n} ((g_n^{jT} D_n^{-1} \lambda_{n|k} D_k \Gamma_k)^{\ell})^2$$

$$s^j_{n|o} = g_n^{jT} D_n^{-1} \lambda_{n|o} P_{o|o} \lambda_{n|o}^T D_n^{T-1} g_n^j$$

Then
$$b^j_{n|n-1} = s^j_{n|o} + (U''_{n|1}\, r + W'_{n|1}\, q)^j$$

From this it follows that

$$E[(b^j_{n|n-1})^2] - (b^{*j}_{n|n-1})^2 = (U''_{n|1} \sum_R U''^T_{n|1} + W'_{n|1} \sum_Q W'^T_{n|1})^{jj}$$

In a similar fashion it can be shown that

$$E(Q^{jj} b^j_{n|n-1}) - \bar{Q}^{jj} b^{*j}_{n|n-1} = (W'_{n|1})^{jj} \sum_Q^{jj}$$

and $$E(r^k b^j_{n|n-1}) - \bar{r}^k b^{*j}_{n|n-1} = (\sum_R U''^T_{n|1})^{kj}$$

So $$E[(\Delta M^{jj}_n)^2] = \sum_Q^{jj} + m^{jT}_n \sum_R m^j_n + (U''_{n|1} \sum_R U''^T_{n|1} + W'_{n|1} \sum_Q W'^T_{n|1})^{jj}$$

$$- 2 W'^{jj}_{n|1} \sum_Q^{jj} + 2 (m^{jT}_n \sum_R U''^T_{n|1})^j$$

Define $$W_{n|1} = W'_{n|1} - I$$

$$(U'_{n|1})^{jk} = (U''_{n|1})^{jk} + (m^j_n)^k$$

Then $$E[(\Delta M^{jj}_n)^2] = (U'_{n|1} \sum_R U'^T_{n|1} + W_{n|1} \sum_Q W^T_{n|1})^{jj}$$

In a similar fashion it can be shown that

$$E(Q^{jj} \Delta M^{jj}_n) = W^{jj}_{n|1} \sum_Q^{jj}$$

So $E[((Q + \Delta M_n - T_n^*)^{jj})^2] = (\bar{Q}^{jj} - T_n^{*jj})^2 + (U'_{n|1} \sum_R U'^T_{n|1} + W'_{n|1} \sum_Q W'^T_{n|1})^{jj}$

and $E(J_n^{jj}) = \left(\frac{n-1}{n}\right)^2 E(J_{n-1}^{jj}) + \frac{2}{n^2}[(\bar{Q}^{jj} - T_n^{*jj})^2 + (U'_{n|1} \sum_R U'^T_{n|1} + W'_{n|1} \sum_Q W'^T_{n|1})^{jj}]$

Using a procedure similar to that of finding $E[(F_n^{jj})^2]$, it can be shown that

$$E[(M_n^{jj})^2] = \left(\frac{n-1}{n}\right)^2 E[(M_{n-1}^{jj})^2] + \frac{1}{n^2}[2 \sum_{k=1}^{n} E(\Delta M_k^{jj} \Delta M_n^{jj}) - E(\Delta M_n^{jj})^2]$$

It is easy to show that

$$E(\Delta M_k^{jj} \; \Delta M_n^{jj}) = (U'_{k|1} \; \sum_R \; U'^T_{n|1} + W_{k|1} \; \sum_Q \; W^T_{n|1})^{jj}$$

So $E[(M_n^{jj})^2] = \left(\frac{n-1}{n}\right)^2 E[(M_{n-1}^{jj})^2] + \frac{1}{n^2} [(2 \; n \; U'_{n|1} \; \sum_R \bar{U}'^T_{n|1} - U'_{n|1} \sum_R U'^T_{n|1})^{jj}$

$$+ (2 \; n \; W'_{n|1} \; \sum_Q \; \bar{W}'^T_{n|1} - W'_{n|1} \; \sum_Q \; W'^T_{n|1})^{jj}]$$

where

$$\bar{U}'_{n|1} = \frac{1}{n} \sum_{k=1}^{n} U'_{k|1}$$

$$\bar{W}'_{n|1} = \frac{1}{n} \sum_{k=1}^{n} W'_{k|1}$$

$U''_{n|1}$ and $W'_{n|1}$ can be computed as functions of $(\alpha_n)^{m\ell u}$ and $(\beta_n)^{m\ell u}$. It can be seen that

$$(U''_{n|1})^{j\ell} = (g_n^{jT} D_n^{-1})^m \, (\alpha_n)^{m\ell u} \, (D_n^{T-1} g_n^j)^u - ((g_n^{jT} D_n^{-1} A_n^*)^{\ell})^2$$

$$(W'_{n|1})^{j\ell} = (g_n^{jT} D_n^{-1})^m \, (\beta_n)^{m\ell u} \, (D_n^{T-1} g_n^j)^u$$

REFERENCES

1. Abramowitz, M. and I. A. Stegun (1966), Handbook of Mathematical Functions, National Bureau of Standards Applied Mathematics Series 55, U. S. Government Printing Office, Washington, D. C.

2. Blum, M. (1966), "Best Linear Unbiased Estimation by Recursive Methods," SIAM Journal on Applied Math., Vol. 14, No. 1, pp. 167-181, January.

3. Cramer, H. (1955), The Elements of Probability Theory, John Wiley & Sons, New York.

4. Davisson, L. D. (1966), "Adaptive Linear Filtering when Signal Distributions are Unknown," IEEE Trans. on Auto. Control, Vol. 11, No. 4, pp. 740-742, October.

5. Dennis, A. R. (1967), "Functional Updating and Adaptive Noise Variance Determination in Recursive-Type Trajectory Estimators," presented at the Special Projects Branch Astrodynamics Conference, NASA Goddard Space Flight Center, Maryland, May.

6. Deutsch, R. (1965), Estimation Theory, Prentice-Hall, Englewood Cliffs, N. J.

7. Deyst, J. J. (1964), "Optimal Continuous Estimation of Nonstationary Random Variables," (S. M. Thesis), M. I. T. Instrumentation Laboratory, Report T-369, Cambridge, Massachusetts.

8. Drenick, R. F. and L. Shaw (1964), "Optimal Control of Linear Plants with Random Parameters," IEEE Trans. on Auto. Control, Vol. 9, No. 3, pp. 236-244, October.

9. Fitzgerald, R. J. (1967), "Error Divergence in Optimal Filtering Problems," presented at the Second IFAC Symposium on Automatic Control in Space, Vienna, Austria, September.

10. Fraser, D. C. (1967), "A New Technique for the Optimal Smoothing of Data," (Sc.D. Thesis), M. I. T. Instrumentation Laboratory, Report T-474, Cambridge, Massachusetts.

11. Friedman, A. L., A. Dushman, and A. Gelb (1964), "Optimization of Sampling Long-Term Inertial Navigation Systems," IEEE Trans.on Aerospace and Navigational Electronics, Vol. 11, No. 3, pp. 142-150, September.

12. Gelb, A., A. Dushman, and H. J. Sandberg (1963), "A Means for Optimum Signal Identification," NEREM Records, IEEE Northeast Electronics Research and Engineering Meeting, Vol. 5, pp. 80-81.

13. Hays, W. L. (1965), Statistics for Psychologists, Holt, Rinehart, & Winston, New York.

14. Hildebrand, F. B. (1961), Methods of Applied Mathematics, Prentice Hall, Englewood Cliffs, N. J.

15. Jazwinski, A. H. (1967), "Adaptive Filtering," Interim Report No. 67-6, Analytical Mechanics Associates, Inc., Lanham, Maryland.

16. Kalman, R. E. (1960), "A New Approach to Linear Filtering and Prediction Problems," Trans. ASME, Series D, Jour. of Basic Eng., Vol. 82, pp. 35-45, March.

17. Kalman, R. E., and R. S. Bucy (1961), "New Results in Linear Filtering and Prediction Theory," Trans. ASME, Series D, Jour. of Basic Eng., Vol. 83, pp. 95-107, March.

18. Kozin, C. H., and G. R. Cooper (1966), "The Use of Prior Estimates in Successive Point Estimation of a Random Signal," SIAM Journal on Applied Math., Vol. 14, No. 1, pp. 112-130, January.

19. Le Cam, L. (1956), "On the Asymptotic Theory of Estimation and Testing Hypotheses," Proc. of Third Berkeley Symp. on Math. Stat. and Prob., Vol. 1, pp. 129-156, University of California Press, Berkeley.

20. Lee, R. C. K. (1964), Optimal Estimation, Identification, and Control, Research Monograph No. 28, M. I. T. Press, Cambridge, Massachusetts.

21. Magill, D. T. (1965), "Optimal Adaptive Estimation of Sampled Stochastic Processes," IEEE Trans. on Auto. Cont., Vol. 10, No. 4, pp. 434-439, October.

22. Neyman, J. (1949), "Contribution to the Theory of the Chi Squared Test," Proc. of First Berkeley Symp. on Math. Stat. and Prob., Vol. 1, pp. 239-273, University of California Press, Berkeley.

23. Potter, J. E., and R. G. Stern (1963), "Statistical Filtering of Space Navigation Measurements," presented as Preprint 63-333 at the AIAA Guidance and Control Conf., Cambridge, Mass., August.

24. Potter, J. E. (1965), "A Matrix Equation Arising in Statistical Filter Theory," NASA Contractor Report NASA CR-270, NASA, Washington, D. C.

25. Rao, C. R. (1965), Linear Statistical Inference and Its Applications, John Wiley & Sons, New York.

26. Rauch, H. E., F. Tung, and C. T. Striebel (1965), "Maximum Likelihood Estimates of Linear Dynamic Systems," AIAA Journal, Vol. 3, No. 8, pp. 1445-1450, August.

27. Richards, F. S. G. (1961), "A Method of Maximum Likelihood Estimation," Jour. Roy. Statis. Soc., Series B, Vol. 23, pp. 469-475.

28. Robbins, H. (1956), Proc. of Third Symp. on Math. Stat. and Prob., Vol. 1, pp. 157-163, University of California Press, Berkeley.

29. Rust, B. (1966), "A Simple Algorithm for Computing the Generalized Inverse of a Matrix," Communications of the ACM, Vol. 9, No. 5, pp. 381-387, May.

30. Silvey, S. D. (1961), "A Note on the Maximum-Likelihood in the Case of Dependent Random Variables," Jour. Roy. Statis. Soc., Series B, Vol. 23, pp. 444-452.

31. Shellenbarger, J. C. (1966), "Estimation of Covariance Parameters for an Adaptive Kalman Filter," (Ph. D. Thesis), Iowa State University Department of E. E., Ames, Iowa.

32. Sklansky, J. (1966), "Learning Systems for Automatic Control," IEEE Trans. on Auto. Cont., Vol. 11, No. 1, pp. 6-19, January.

33. Smith, G. L. (1967), "Sequential Estimation of Observation Error Variances in a Trajectory Estimation Problem," AIAA Journal, Vol. 5, No. 11, pp. 1964-1969, November.

34. Spang, H. A. (1965), "Optimum Control of an Unknown Linear Plant Using Bayesian Estimation of the Error," IEEE Trans. on Auto. Cont., Vol. 10, No. 1, pp. 80-83, January.

35. Wald, A. (1949), "Note on the Consistency of the Maximum Likelihood Estimate," Annals of Math. Statistics. Vol. 20.

36. Wiener, N. (1949), Extrapolation, Interpolation, and Smoothing of Stationary Time Series, M. I. T. Press, Cambridge, Massachusetts.

37. Wilks, S. S. (1963), Mathematical Statistics, John Wiley & Sons, New York.

38. Zaborszky, J. (1966), "An Information Theory Viewpoint for the General Identification Problem," IEEE Trans. on Auto. Cont., Vol. 11, No. 1, pp. 130-131, January.

BIOGRAPHY

Paul D. Abramson was born January 14, 1942, in Shreveport, Louisiana. He attended public schools in Shreveport and graduated from C. E. Byrd High School.

Mr. Abramson entered M. I. T. in September 1959 and received the degrees of Bachelor of Science and Master of Science in Aeronautics and Astronautics from M. I. T. in June 1964. As an undergraduate he held a Civil Air Patrol scholarship and was a member of the Honors Program of the Department of Aeronautics and Astronautics.

While a graduate student, Mr. Abramson held the A. C. Sparkplug and Civil Air Patrol Fellowships. During the summer of 1961 he was employed as a technical assistant at the M. I. T. Instrumentation Laboratory and in the summer of 1963 he was employed by Mithras, Inc. as a research engineer. In September 1963 he joined the M. I. T. Instrumentation Laboratory as a research assistant under the supervision of Mr. John Hatfield. In this capacity he worked on the analysis of "strapped down" inertial guidance systems and various techniques for optimal navigation in space. In October 1967 he jointed the M. I. T. Experimental Astronomy Laboratory under the supervision of Kenneth R. Britting where the remainder of this thesis research was conducted.

Mr. Abramson is married to the former Colene Piercy of Shreveport, Louisiana. He is a member of Tau Beta Pi, Sigma Xi, and Sigma Gamma Tau honorary fraternities.

NASA-Langley, 1970 —— 19

"The aeronautical and space activities of the United States shall be conducted so as to contribute . . . to the expansion of human knowledge of phenomena in the atmosphere and space. The Administration shall provide for the widest practicable and appropriate dissemination of information concerning its activities and the results thereof."

—NATIONAL AERONAUTICS AND SPACE ACT OF 1958

NASA SCIENTIFIC AND TECHNICAL PUBLICATIONS

TECHNICAL REPORTS: Scientific and technical information considered important, complete, and a lasting contribution to existing knowledge.

TECHNICAL NOTES: Information less broad in scope but nevertheless of importance as a contribution to existing knowledge.

TECHNICAL MEMORANDUMS: Information receiving limited distribution because of preliminary data, security classification, or other reasons.

CONTRACTOR REPORTS: Scientific and technical information generated under a NASA contract or grant and considered an important contribution to existing knowledge.

TECHNICAL TRANSLATIONS: Information published in a foreign language considered to merit NASA distribution in English.

SPECIAL PUBLICATIONS: Information derived from or of value to NASA activities. Publications include conference proceedings, monographs, data compilations, handbooks, sourcebooks, and special bibliographies.

TECHNOLOGY UTILIZATION PUBLICATIONS: Information on technology used by NASA that may be of particular interest in commercial and other non-aerospace applications. Publications include Tech Briefs, Technology Utilization Reports and Technology Surveys.

Details on the availability of these publications may be obtained from:

SCIENTIFIC AND TECHNICAL INFORMATION DIVISION
NATIONAL AERONAUTICS AND SPACE ADMINISTRATION
Washington, D.C. 20546