Research Institute for Advanced Computer Science
NASA Ames Research Center

P-63

Gaussian Windows: A Tool for Exploring Mulitvariate Data

Louis A. Jaeckel

Research Institute for Advanced Computer Science

NASA Ames Research Center

NASA Cooperative Agreement Number NCC2-408 and NCC2-387

RIACS Technical Report 90.41 September 1990

RIACS

Gaussian Windows: A Tool for Exploring Mulitvariate Data

Louis A. Jaeckel

Research Institute for Advnanced Computer Science
NASA Ames Research Center - MS: Ellis
Moffett Field, CA 94035

RIACS Technical Report 90.41

September 1990

Abstract: This paper presents a method for interactively exploring a large set of quantitative multivariate data, in order to estimate the shape of the underlying density function. It is assumed that the density function is more or less smooth, but no other specific assumptions are made concerning its structure. The local structure of the data in a given region may be examined by viewing the data through a "Gaussian window", whose location and shape are chosen by the user. A Gaussian window is defined by giving each data point a weight based on a multivariate Gaussian function. The weighted sample mean and sample covariance matrix are then computed, using the weights attached to the data points. These quantities are used to compute an estimate of the shape of the density function in the window region. The local structure of the data is described by a method similar to the method of principal components. By taking many such local views of the data, we can form an idea of the structure of the data set. The method is applicable in any number of dimensions. The method can be used to find and describe simple structural features such as peaks, valleys, and saddle points in the density function, and also extended structures in higher dimensions. With some practice, we can apply our geometrical intuition to these structural features in any number of dimensions, so that we can think about and describe the structure of the data. Since the computations involved are relatively simple, the method can easily be implemented on a small computer.

The Research Institute of Advanced Computer Science is operated by Universities Space Research Association, The American City Building, Suite 311, Columbia, MD 244, (301)730-2656

Work reported herein was supported in part by Cooperative Agreements NCC 2-408 and NCC 2-387 between the National Aeronautics and Space Administration (NASA) and the Universities Space Research Association (USRA).

GAUSSIAN WINDOWS: A TOOL FOR EXPLORING MULTIVARIATE DATA

INTRODUCTION

Suppose that we have a large set of quantitative data consisting of N points x_i in a p-dimensional space, and that we want to explore the structure of this data set, without making many assumptions in advance concerning its structure. By "structure" I mean the shape of the underlying density function, as evidenced by the locations and shapes of concentrations of data points. I will generally think of the data set as a random sample drawn from some probability distribution or from some larger population. This assumption may not always be warranted, and I will not rely heavily on it. I will, however, assume that the density function is more or less smooth, so that if we want to learn about the structure of the distribution in a given region in the space, we can draw inferences about the structure based on the nearby data points. Note that without any knowledge or assumptions about the large-scale structure of the data, we cannot learn much, if anything, about the structure in a given region by looking at data points that are far away.

The method proposed here consists of repeatedly examining the local structure of the data by viewing the data through windows, each having a location and shape as defined below. By

taking many such local views of the data, that is, by interactively exploring or searching through the space in which the data points lie, we may be able to find and describe structural features such as peaks, valleys, and saddle points in the density function. For example, a cluster of data points is evidence of a peak in the density function. We may also find extended structures such as ridges, and analogous structures in higher dimensions. We can then put together what we have found in order to build up a general description of the structure of the data set.

In multivariate statistical analysis, the simplest and best-understood quantities to compute are the sample mean vector and the sample covariance matrix. These statistics describe the overall structure of the data, while at the same time obscuring or smearing out any fine detail that may be present. If we do not make specific assumptions about the data, the only way to discover any such small-scale structure is to look for it on a local level. In two or three dimensions we can do this by looking at a scatter plot or other graphical representation of the data, but in higher dimensions we cannot do this directly. So I will use a window to look at the data in a local region, and compute quantities such as the sample mean and covariance matrix of the data as seen through the window. If the part of the data that is in the window region happens to be a cluster with approximately a Gaussian shape, then the local mean and covariance matrix will give us a good description of the cluster. In practice, however, the data that we see in a window may have

only a part of a Gaussian shape, or the data may consist of parts of more than one cluster. Since we do not know in advance where the clusters are, or even whether there are clusters in the data, any window we might try could contain parts of one or more clusters, or parts of more complex structures.

If we use a window with sharp boundaries (for example a rectangular or an ellipsoidal window), such that each data point is either inside or outside of the window, then what is seen through the window may be overly sensitive to the exact placement and shape of the window. More importantly, if we use such a window, and if we assume that the data in the window form part of a Gaussian shape, it will be very difficult computationally to estimate the parameters of such a truncated Gaussian distribution, especially if the dimension of the space is large. (The usual way to do this would be to estimate the parameters of the Gaussian shape by the method of maximum likelihood.) Moreover, the data in a local region may not look like a cluster at all; a concentration of data points may appear more like a ridge or a valley or a saddle point. A method for exploring the data should be able to deal with such structural features. If we can choose the shape of the window so that the computational effort is reduced, then a user with a small computer will be able to try many windows with different locations and shapes quickly, and will thus be able to explore the data interactively.

Instead of a window with sharp boundaries, I will use a "shaded" window, which may be thought of as a window whose transparency is greatest at the center and which becomes

progressively more opaque as we move away from the center. The window shape I will use is defined by a multivariate Gaussian function. A *Gaussian window* is defined by choosing a p-dimensional vector a to be its center point, and a non-negative definite symmetric matrix V to describe its size and shape. For any p-dimensional vector x, let $w(x)$ be the value of the Gaussian function

$$w(x) = e^{-\frac{1}{2}(x - a)'V(x - a)} .$$

This function may be thought of as the relative transparency of the window at the point x. Note that $w(a) = 1$ and $w(x) \leq 1$ for all other x, with $w(x)$ non-increasing as x moves away from a. Each data point x_i is given the weight $w_i = w(x_i)$. The weighted (or "windowed") sample means, variances, and covariances are then computed from the weighted sums, sums of squares, and sums of products of the coordinates of the data points. The quantity Σw_i is used as the windowed equivalent of the sample size.

Suppose that the data in the region of the window (that is, in the region where $w(x)$ is not very small) happen to form a cluster with approximately a multivariate Gaussian shape. Then the "windowed" data, that is, the data points with the weights w_i attached to them, will also have a Gaussian shape, but because of the weighting of the data points, the parameters of this Gaussian shape will be different from the actual parameters of the cluster of data points without the weights. Since the window parameters are known, we can compute the biasing effect of

the window, and we can easily work backwards and "degauss" the windowed data; that is, we can remove the effect of the Gaussian window on the shape of the cluster and recover estimates of the actual parameters of the cluster. Because the windowed data in this case have a multivariate Gaussian shape, we will, by analogy with classical statistical theory, estimate the parameters of the windowed Gaussian shape by the weighted sample mean vector and sample covariance matrix. We will then degauss these estimated parameters.

In general, however, data sets will not consist of relatively isolated clusters with Gaussian shapes. In addition to peaks in the density function, we may also have valleys, ridges, saddle points, and similar but more complex features in higher dimensions, and we must be able to recognize such structural features so that we can include them in an ultimate understanding or description of the structure of the data. We will see that these local structural elements can be approximated, at least locally, by a function in the form of the exponential of a polynomial of degree at most two in the p coordinates of x. This family of functions includes the multivariate Gaussian density functions. The second-degree terms in the exponent may be expressed as a quadratic form based on a symmetric matrix that is analogous to the inverse of a covariance matrix, except that it does not have to be positive definite. If in the region of a Gaussian window the density function can be approximated by such a function, then the windowed data will have approximately a proper multivariate Gaussian shape, just as it

would if the data in the window region formed a Gaussian-shaped cluster. So we can compute the windowed sample mean and covariance matrix based on the weighted data as before, and we can then degauss the results to estimate the local shape of the density function. The difference in this case is that the degaussing process may lead to a symmetric matrix to describe the estimated shape of the density that is not positive definite. Whether it is or not, we will use the eigenvectors and eigenvalues of this matrix to describe the local structure of the data, by a method analogous to the method of principal components. Thus, Gaussian windows may be applied in situations where the data in the window region have shapes other than a Gaussian cluster, and we can use this technique to discover a variety of structural features in the data. We will have to be careful, however, about how we do the computations, so that we avoid trying to compute quantities that are numerically unstable.

The computations done for a window are the following: In the first stage we compute the weights w_i and the weighted sample means, variances, and covariances. The effort involved in this stage is proportional to Np^2. We then do standard matrix operations on $p \times p$ matrices, including inverting the windowed covariance matrix, degaussing that matrix by subtracting V from it, and then extracting the eigenvectors and eigenvalues of this degaussed symmetric matrix. I wrote a simple program in BASIC so that I could perform some experiments on artificial data sets. I also compute a variety of statistical quantities that may be useful in interpreting the results. These computations can be

done quickly on a small computer if N and p are not too large. The matrix operations can be done by whatever algorithms the user prefers; much standard software is available for this purpose. The only new software required is a program to control the process, accept the data and the user's chosen window parameters, compute the weights, weighted sums, and related quantities, and display the results. If a computer capable of parallel processing is available, the first stage of the computations can be done partly in parallel, with substantial savings in time if N and p are very large.

One of the guiding principles in this work is that the method should be applicable, at least in principle, in any number of dimensions. We must therefore develop some simple ways of thinking about geometric structures in a p-dimensional space. We can do this by making analogies with shapes that we can visualize in two or three dimensions, such as ellipsoids and hyperboloids. The shape of an analogous object in p dimensions can be described by giving its principal axes (a set of mutually orthogonal vectors) and a scale factor for each axis. These quantities are related to the eigenvectors and eigenvalues of the symmetric matrix that defines the shape of the object. Thus, by viewing the data through a window, and assuming that the local structure of the data has a simple form, we have a way of thinking about and describing what the estimated local structure looks like in any number of dimensions. Of course, when we make analogies like this so that we can apply our geometrical intuition in a space with large p, we must do so carefully so

that we are not misled.

The philosophy here is different from that in the many graphical methods which involve projecting the data onto a space of two or three dimensions, so that we can use the pattern-recognition capabilities of our own visual systems. See for example Chambers et al. (1983), Cleveland and McGill (1988), and Du Toit et al. (1986). There may well be multidimensional patterns or structures in the data that would be obscured or lost, or at least very hard to find, in a lower-dimensional projection of the data. I should emphasize, however, that the method described here is not intended to replace or compete with those other methods; instead, it is meant to complement them. When we have a large, complex data set to study, the more ways we have to look at the data, the better.

In this paper I discuss the method of Gaussian windows as a tool to be used for exploring data interactively. It is natural to ask whether this process can be automated. If it were, we would then have an example of a process of automatic "unsupervised learning", in which a machine or an algorithm is given a set of data and is then supposed to figure out the structure of the data without the further help of a "teacher". See for example Pao (1989) and Cheeseman et al. (1988). In order to automate the process, we must have a clearly defined goal; that is, we need a clear idea of what sort of ultimate description of the structure of the data we want the process to give us. We would also have to specify the strategy it should use, both in choosing a sequence of windows through which to view

the data, and also in taking the information that it finds in those views and putting it together into an organized description of the structure of the data. Since I do not have specific ways of doing these things, I think of the method primarily as an interactive one. Perhaps with more experience we can decide on ways of automating the process, at least partially.

But there are advantages to an interactive method in its own right, besides as a stepping-stone to constructing an automatic method. When we examine the data interactively, we can proceed in a more open-ended way, feeling our way along as we go, and we can bring in any other knowledge, assumptions, or hunches about the data that we may have, without being constrained by any specific form that our results must take. With practice, we can develop skills in exploring data in this way, and in building a mental picture or description of what we find in the data. Furthermore, an interactive method can complement an automatic procedure and could be used in conjunction with it. For example, after running an automatic clustering algorithm on the data, we could use Gaussian windows to look at the results of the algorithm in more detail. We could examine the size and shape of the clusters found by the algorithm, we could look for structural features other than clusters, and we could look for fine structure that the algorithm may have been unable to find. However, even if we use an interactive method to explore the data, we cannot avoid the issues of goals and strategies raised above. We still need some idea of the kind of description of the structure of the data that we hope ultimately to find. In other

words, we need to have some ideas about objectives and strategies in mind, so that we do not wander aimlessly through the data. Some of these issues will be discussed below.

In the next section I will work out the mathematics in the one-dimensional case. I will show the effect on the density function of attaching Gaussian weights to the data points, and how we can reverse that effect — that is, degauss the windowed data — to estimate the parameters of the density function in the window region. In the section after that, I will do the same for the general case of p-dimensional data, and I will use a method analogous to the method of principal components to describe the estimated local structure of the data. We will then have the mathematical tools so that, in the final section, we can consider some possible structural features that we might find in the data, and how they would appear when viewed through a Gaussian window. Some of these features, such as peaks, valleys, and saddle points, are pivotal points that will help us to develop a description of the structure of the data. Other structural features, such as ridges, are extended features which cannot be viewed in a single window. If such an extended structure passes through a window, we will see a part of it in the window, and we will be able to tell that what we see is part of a structure that extends beyond the window. We can then try to move along the structure by using windows with different centers, so that we can map out its extent and shape. Thus we are not restricted to finding clusters, that is, sets of data points all of which are near each other. Finally, we can put together the results of our

exploration of the data set into an overall description of its structure.

THE ONE-DIMENSIONAL CASE

Many of the basic properties of the Gaussian window can be illustrated in the case $p = 1$. We will consider the general case in the next section.

Suppose that we have a sample of N data points x_i from a univariate density function $f(x)$. We will view the data through a Gaussian window. If we let α be the window center, and we let $v > 0$ be a parameter for the width of the window (if we were describing a Gaussian density function, its variance would be $\frac{1}{v}$), then the window is defined by

$$w(x) = e^{-\frac{1}{2} v(x - \alpha)^2} .$$

The data as seen through the window consist of the x_i, with each x_i given a weight $w_i = w(x_i) \leq 1$, instead of being given full weight. The "windowed" density function, that is, the effective density function of the data as viewed through this shaded window, is $w(x)f(x)$. That is, if we do computations with the weighted x_i, the results will be as if we were working with a sample from the windowed density function. Note that this function is not a proper probability density function, because its integral over x is less than 1. The integral $\int w(x)f(x)dx$ may be thought of as the expected proportion of the data that is viewed through the window. Since this integral is the expected

value of w_i for a randomly chosen data point x_i, it may be estimated by $\frac{1}{N}\Sigma w_i$.

A simple way to think of the window process is to imagine that each x_i is a small point of light with intensity 1, and that the window has transparency $w(x)$ at each x. The light from each x_i that passes through the window therefore has intensity w_i, and the total intensity of the light seen through the window is Σw_i. Another way to think about the windowed density $w(x)f(x)$ is to imagine that we take the data set and randomly throw out some of the points, using the following rule: Independently for each x_i, keep x_i with probability w_i, and throw it out with probability $1 - w_i$. The conditional density function for the remaining points would be

$$\frac{w(x)f(x)}{\int w(x)f(x)dx} .$$

The integral in the denominator is the expected proportion of the points remaining. It would of course be wasteful to throw away data in this way; we can achieve the same effect, and make better use of the data, by giving each x_i the weight w_i in our computations, where w_i is the probability that x_i would have survived the throwing-out process. The full data set, with weights attached, may be thought of as the result of averaging over all possible outcomes of the random throwing-out process.

The reason for using a Gaussian window is a simple mathematical fact: *A Gaussian times a Gaussian is a Gaussian.* Suppose that in the region near the center of a window we have chosen, the density function has (approximately) a Gaussian

shape:

$$f(x) = c \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2\sigma^2}(x - \mu)^2} ,$$

where μ, σ^2, and c are all unknown parameters. This part of $f(x)$ resembles a peak whose center is at μ and whose width is represented by σ. The multiplicative constant c represents the *probability mass* of this part of the entire probability distribution; that is, c is the probability that a point chosen at random comes from this part of the distribution. Suppose further that the rest of the probability distribution is so far away from the center of the window that the data points arising from other parts of the distribution will have only a negligible effect on the computations. Then, in the region of the window (that is, the region vaguely defined by "$w(x)$ is not too small"), we will see that the windowed density $w(x)f(x)$ has a Gaussian shape whose parameters are related to the parameters of the Gaussian function above and to the window parameters.

Before going further I will rewrite the Gaussian function above so that later we can apply the results more generally. Let $b = \frac{1}{\sigma^2}$ and let $a = c \frac{1}{\sqrt{2\pi}\sigma}$. Then, in the window region,

$$f(x) = a e^{-\frac{1}{2} b(x - \mu)^2} .$$

Since we will choose α, the window center, I will assume for simplicity that $\alpha = 0$. So the windowed density (the effective density function for the weighted data) is

$$w(x)f(x) = a\, e^{-\frac{1}{2}[b(x-\mu)^2 + vx^2]} .$$

The expression in the brackets can be rewritten by completing the square:

$$b(x-\mu)^2 + vx^2 = (b+v)x^2 - 2b\mu x + b\mu^2$$

$$= (b+v)\left(x^2 - \frac{2b\mu}{b+v}\,x + \frac{b^2\mu^2}{(b+v)^2}\right) + b\mu^2 - \frac{b^2\mu^2}{b+v}$$

$$= (b+v)\left(x - \mu\frac{b}{b+v}\right)^2 + \mu^2\frac{bv}{b+v} .$$

Therefore the windowed density is

$$w(x)f(x) = a\, e^{-\frac{1}{2}\,\mu^2\frac{bv}{b+v}}\; e^{-\frac{1}{2}(b+v)\left(x - \mu\frac{b}{b+v}\right)^2} .$$

This is a Gaussian function with "windowed mean" $\mu\frac{b}{b+v}$ and "windowed variance" $\frac{1}{b+v}$. The windowed mean has been pulled toward the window center because the data points nearer the center are given relatively greater weight. Note that the windowed variance is a function of b and v, but not of μ, and that it is less than σ^2 and also less than $\frac{1}{v}$ (the "variance" of the window).

If we write the windowed density as

$$\left[a\, e^{-\frac{1}{2}\,\mu^2\frac{bv}{b+v}}\,\frac{\sqrt{2\pi}}{\sqrt{b+v}}\right]\frac{\sqrt{b+v}}{\sqrt{2\pi}}\; e^{-\frac{1}{2}(b+v)\left(x - \mu\frac{b}{b+v}\right)^2} ,$$

then the expression to the right of the brackets is an ordinary Gaussian (normal) probability density function, whose integral over x is 1. Therefore the expression in the brackets is the integral of $w(x)f(x)$ over x. This quantity is the expected

value of the weight $w_i = w(x_i)$ to be assigned to a randomly chosen data point x_i; that is, $E(w_i) = \int w(x)f(x)dx$. A natural way to estimate the expression in the brackets, therefore, is by the average of the weights: $\frac{1}{N}\Sigma w_i$.

Since the windowed data have an approximate Gaussian shape, the simplest and most natural way to estimate the parameters of this shape is to compute the weighted sample mean and sample variance, by analogy with standard statistical theory:

$$\bar{x}_w = \frac{1}{\Sigma w_i}\Sigma w_i x_i$$

and

$$s_w^2 = \frac{1}{\Sigma w_i}\Sigma w_i(x_i - \bar{x}_w)^2 = \frac{1}{\Sigma w_i}\Sigma w_i x_i^2 - \bar{x}_w^2 \ .$$

It follows that $\bar{x}_w$ is an estimate of $\mu\frac{b}{b+v}$ and s_w^2 is an estimate of $\frac{1}{b+v}$.

We can now "degauss" the view of the data as seen through the Gaussian window; that is, we can remove the effect of the window on the shape of the density $f(x)$ in the window region. Since v is known, we have

$$s_w^2 = \frac{1}{\hat{b}+v} \ ,$$

where $\hat{b}$ will be our estimate of b. Then

$$\hat{b} = \frac{1}{s_w^2} - v \ ,$$

and our estimate of σ^2 is

$$\hat{\sigma}^2 = \frac{1}{\hat{b}} = \frac{1}{1/s_w^2 - v} \ .$$

Assume for the moment that the denominator above is positive.

To estimate μ we write

$$\bar{x}_w = \hat{\mu} \frac{\hat{b}}{\hat{b}+v} ,$$

so our estimate of μ will be

$$\hat{\mu} = \bar{x}_w \frac{\hat{b}+v}{\hat{b}} = \bar{x}_w \frac{1/s_w^2}{1/s_w^2 - v} = \bar{x}_w \frac{\hat{\sigma}^2}{s_w^2} .$$

We can also estimate the constants a and c. Since we have a natural estimate, $\frac{1}{N} \Sigma w_i$, for $\int w(x)f(x)dx$, the expression in brackets above, we can write

$$\hat{a}\, e^{-\frac{1}{2} \hat{\mu}^2 \frac{\hat{b}v}{\hat{b}+v}} \frac{\sqrt{2\pi}}{\sqrt{\hat{b}+v}} = \frac{1}{N} \Sigma w_i ,$$

and we have

$$\hat{a} = \frac{1}{N} \Sigma w_i \frac{\sqrt{\hat{b}+v}}{\sqrt{2\pi}} e^{\frac{1}{2} \hat{\mu}^2 \frac{\hat{b}v}{\hat{b}+v}} .$$

We can also estimate c by

$$\hat{c} = \sqrt{2\pi}\, \hat{\sigma}\, \hat{a} .$$

Thus we have a method of estimating the three parameters describing the data in the window region that is both computationally simple and easy to understand. As we will see in the next section, the same operations can be done in any number of dimensions. Bear in mind that since we compute the windowed sample mean and variance, the method estimates the overall structure of the data in the window region; any fine structure that may be present within the region is smeared out. To look for more detailed structure, we could use a smaller window. The

windows can be as small as the data will allow; if the amount of data in the window region is too small, we will not be able to obtain reliable estimates.

We could do similar computations based on a window of a different shape. For example, suppose we use a window that is simply an interval, with $w(x) = 1$ within the interval and $w(x) = 0$ outside it. Then, under the assumptions above, we would see a truncated Gaussian function in the window, and we could estimate its parameters by the method of maximum likelihood or by some other method. However, the computations would be more difficult — not so much in one dimension, but in a many-dimensional space the computational problems would be very complex.

The data present in the region of a window may not look like a single peak, as we assumed above. Since we do not know what we will find in a window before we look, a chosen window might contain two or more peaks, or none at all; we might find a valley between two peaks, or a flat area, or a gradual slope, or something more complex. Often we may be able to approximate the shape of the density function in a window region by one of the following:

$$f(x) = h \ ,$$

$$f(x) = h \, e^{rx} \ ,$$

or

$$f(x) = a \, e^{-\frac{1}{2} b(x - \mu)^2} \ .$$

In all of these cases, $f(x)$ is the exponential of a polynomial in x of degree at most two. In the third case, if $b > 0$ we have the ordinary Gaussian case which we treated above. If $b < 0$, we have what I will call a "concave Gaussian" function, which will be useful in regions where $f(x)$ is concave upward. The first two cases will be treated below. If we multiply any of these functions by $w(x)$, as we did for the Gaussian, and then complete the square in the exponent, we find that the windowed data have a Gaussian shape, as before. So we can estimate the parameters of the windowed density by computing the weighted sample mean and variance, and then degauss the data to obtain estimates of the parameters of the local approximation to $f(x)$.

In the third case above, if $b < 0$ we can find the windowed density $w(x)f(x)$ by the same algebraic steps that we used earlier. Assume for the moment that $b + v > 0$, so that $w(x)f(x)$ is a proper Gaussian shape. We can therefore estimate the parameters b, μ, and a by the same formulas as before. We will not estimate σ^2 or c in this case because they are meaningful only when $f(x)$ is an ordinary Gaussian. In this case the windowed variance $\frac{1}{b+v}$ is greater than the "variance" of the window, $\frac{1}{v}$, whereas in the ordinary Gaussian case it was smaller. We can distinguish between the ordinary and the concave Gaussian cases by looking at the sign of $\hat{b}$. However, if $\hat{b}$ is near 0, we may want to consider one of the cases discussed below.

What if $b + v \leq 0$? This would be true if $f(x)$ in the window region were very strongly concave upward, so much so that the window function $w(x)$ could not pull it down into a Gaussian

shape. In practice, however, we do not have to worry about this possibility. Since we are working with a finite data set, the data cannot continue to follow such a sharply increasing density function indefinitely as we move farther and farther from the window center. Eventually the data must taper off, so the shape of the weighted data would be something like a valley between two hills. In such a case, s_w^2 (which must be positive in any case) would be very large, so we would have a large, positive estimate of $\frac{1}{b+v}$, and hence a value of $\hat{b}$ near, but greater than, $-v$.

The second case above may be thought of as a limiting case of the third, if we let $\mu = \frac{r}{b}$ and we let b approach 0. The first case is a special case of the second. The parameter h is the density at 0, the window center. In the second case above, the windowed density is

$$w(x)f(x) = h\, e^{-\frac{1}{2} vx^2 + rx} .$$

The exponent may be written as

$$\begin{aligned} -\tfrac{1}{2} vx^2 + rx &= -\tfrac{1}{2} v\left(x^2 - 2\tfrac{r}{v} x\right) \\ &= -\tfrac{1}{2} v\left(x^2 - 2\tfrac{r}{v} x + \frac{r^2}{v^2}\right) + \frac{r^2}{2v} \\ &= -\tfrac{1}{2} v\left(x - \tfrac{r}{v}\right)^2 + \frac{r^2}{2v} . \end{aligned}$$

Therefore,

$$w(x)f(x) = \left[h\, e^{\frac{r^2}{2v}} \frac{\sqrt{2\pi}}{\sqrt{v}}\right] \frac{\sqrt{v}}{\sqrt{2\pi}} e^{-\frac{1}{2} v\left(x - \frac{r}{v}\right)^2} ,$$

where the expression in the brackets is the integral over x of this function. In this case, the mean of the Gaussian shape seen

through the window is $\frac{r}{v}$, and $\bar{x}_w$ is an estimate of $\frac{r}{v}$. Therefore r may be estimated by $\hat{r} = \bar{x}_w v$. Since the variance of this shape is $\frac{1}{v}$, s_w^2 will be near $\frac{1}{v}$, and hence $\hat{b}$ will be near 0. Thus, a value of $\hat{b}$ near 0 tells us that we should probably approximate $f(x)$ in the window region by a constant or an exponential function. Since $\frac{1}{N}\Sigma w_i$ is an estimate of the expression in the brackets above, we can estimate h by

$$\hat{h} = \frac{1}{N}\Sigma w_i \frac{\sqrt{v}}{\sqrt{2\pi}} e^{-\frac{1}{2}\bar{x}_w^2 v} .$$

If we are in the first case above, in which $f(x)$ is constant, $\bar{x}_w$ would be near 0. Since this case is like the previous case with $r = 0$, the only parameter to estimate is h, which we would estimate by

$$\hat{h} = \frac{1}{N}\Sigma w_i \frac{\sqrt{v}}{\sqrt{2\pi}} .$$

We now have a simple tool for exploring the data, based on the assumption that the shape of the density function in a window region can be approximated by the exponential of a polynomial in x of degree at most two. Since in each of the above cases the windowed density has a Gaussian shape, it is natural to estimate its parameters as we have done above. Since these estimates give us the overall shape of the density in the window region, we can look for finer detail by using smaller windows.

One of the quantities we can estimate is the value of $f(x)$ for a given x in the window region. To estimate $f(x)$ based on a given window, we simply take the estimate of the degaussed

density function, which is defined by the estimated parameters, and evaluate that function at x. In practice we might do this by trying several windows of different sizes centered at or near x, and choosing a window that seems to give a good local picture of the density. If the window is so small that few data points fall within it, the estimates will not be very accurate, and if the window is too large, the view through it might smear out some important details of the structure, and thus give us a misleading estimate of f(x). Since there is probably no generally applicable rule for choosing the best window to use, it is better to experiment with several windows to get a feeling for the data.

In many situations, however, it will be more important to study the structure of the data, rather than to estimate f(x) for particular values of x. To describe the structure of the data, we want to find and describe features such as peaks and valleys. In higher dimensions there may also be ridges, saddle points, and more complex features. The method of Gaussian windows is intended primarily for this purpose. By exploring the local structure of the data using windows with many different centers and sizes, we hope to be able to put together the information found in the windows and build up an overall understanding or description of the structure of the data.

It is natural to ask about the range of validity of the local estimate of the density function based on a window. There is not a clear-cut answer to this question. Generally the estimate should be more reliable near the center of the window, and gradually less so as we move away from the window center.

How reliable the estimate is depends partly on how many data points are involved in producing the estimate. (How involved a data point x_i is depends on w_i, so even this is a matter of degree.) The validity also depends on the true shape of $f(x)$ in the window region. It might seem that if we used a window in the form of an interval, we would have a more definite idea of where the estimated shape of $f(x)$ was valid. But this would be misleading; even in this case the estimate would probably be more reliable near the window center and less so near the endpoints of the interval, because near the center the estimate is better supported by the data. If the density function is not in the form that we expect, the estimate may not be valid anywhere, except in a very general, overall sense. On the other hand, if $f(x)$ is well-behaved, the estimated shape might continue to be accurate for some distance outside the interval. Thus the validity of the estimate would be as much a matter of degree and a function of x here as with a Gaussian window. This question of validity would be a difficult one for any shape of window, and I believe that there is no general-purpose answer to it. To give some sort of answer we would have to make additional assumptions, such as that the function $f(x)$ satisfy certain mathematical conditions and that the data comprise a random sample from some population. Questions such as these will not be addressed in this paper. In practice we can try to get a general sense of the validity of the estimates by trying several windows with different centers and sizes, so that we have some idea of the local structure of the data. For this reason it is important to

have a window shape for which the computations are simple, and also to have a way of thinking about what we find in the windows.

I should point out that it is possible to do other kinds of statistical analyses of the data as seen through a particular window. A more sophisticated analysis of the windowed data could give us a more detailed picture of the local structure of the data in the window region. I have not pursued this approach, however; instead, I search for finer structural details in the data simply by using smaller Gaussian windows and computing the basic quantities defined above. My goal has been to keep the method simple and to make minimal assumptions about the nature of the data, and then to see what could be learned by exploring a data set with this simple tool. Of course, if there is reason to believe that the data may have some particular structure, then we should use a statistical method that is specifically designed to deal with that structure.

Some of the estimates that we might want to compute may be numerically unstable; that is, a small change in the data or in the window used might make a big difference in the value of the estimate. For example, if the shape of the data in the window region is approximately Gaussian, with a variance much larger than the window variance — that is, if the window is located on a Gaussian hill so wide that only a small part of the hill appears in the window — then the estimated parameters $\hat{\mu}$, $\hat{\sigma}^2$, $\hat{a}$, and $\hat{c}$ of the degaussed density function may be so unreliable as to be almost meaningless, at least by themselves. However, the estimate of $f(x)$ by the exponential of a polynomial in x

would still be valid within the window region, provided that the number of data points in the window region is not too small. Estimates of other quantities, such as the slope or the curvature of $f(x)$, may also be valid, as long as we stay within the window region. (In the next section I will give estimates of the first and second derivatives of $\log f(x)$.) Trying to estimate a quantity that represents a feature of the data far outside the window region is risky at best and should be avoided. The quantities that we estimate should be related to the window region, since we can then hope that they can be reliably estimated based on the data appearing in the window.

THE MATHEMATICS OF GAUSSIAN WINDOWS

In the general case we have a large multivariate data set in p dimensions. We will examine the local structure of the data by viewing the data through Gaussian windows. As in the previous section I will consider several cases, so that we can develop the mathematics for describing what we see in a window. We saw that in one dimension the data could have peaks, valleys, and gradual slopes, and that we could distinguish these cases by looking at $\hat{b}$, which was a function of the weighted sample variance. In higher dimensions the data can also have structural features such as saddle points and ridges. A structural feature such as a peak or a valley may lie entirely within a window, or we could have a feature such as a ridge that lies partly within a window and extends beyond it in some directions. Both of these kinds of

features can be described by a method analogous to the method of principal components. In the next section we will consider some examples and some strategies for exploring the data.

Suppose that we have a sample of N data points, or vectors, $x_i = (x_{i1}, x_{i2}, \cdots, x_{ip})'$ from a multivariate density function $f(x)$ defined on a p-dimensional space (or on some region within the space). I will assume that the density function is more or less smooth, but I will not make any other specific assumptions about its structure. I will assume, however, that the data points do not lie in a linear manifold of lower dimension. (By *linear manifold* I mean the set of all vectors x satisfying $Ax = c$ for some matrix A and vector c.) If they do, we could choose a new coordinate system for that linear manifold so that the data do not lie in a linear manifold in the new coordinate system. We define a Gaussian window by choosing a center point a and a non-negative definite symmetric matrix V to describe its shape. For any p-dimensional vector x, let $w(x)$ be the value of the Gaussian function

$$w(x) = e^{-\frac{1}{2}(x - a)'V(x - a)} .$$

Note that for $p > 1$ there is a wide range of possible window shapes. Again, this function represents the relative transparency of the window at the point x, and we have $w(a) = 1$ and $w(x) \leq 1$ for all other x. If V is positive definite, its inverse may be thought of as the "covariance matrix" of the window; that is, if we were describing a multivariate Gaussian density function, V^{-1} would be its covariance matrix. In this

case we have $w(x) < 1$ for all x other than a, and the contours of $w(x)$ are ellipsoids centered at a.

Each data point x_i is given the weight $w_i = w(x_i)$. As before, the windowed density function, that is, the effective density function of the weighted data, is $w(x)f(x)$, which is not a proper density function because its integral is less than 1. The interpretation of this function is the same as in the previous section.

Suppose first that in the region of a window we have chosen, the density has (approximately) a multivariate Gaussian shape:

$$f(x) = c \frac{1}{(2\pi)^{p/2}|\Sigma|^{1/2}} e^{-\frac{1}{2}(x - \mu)'\Sigma^{-1}(x - \mu)},$$

where μ, Σ, and c are all unknown parameters. That is, we have a single peak (or cluster) in the window region. The vector μ is the center of this part of $f(x)$, and the positive definite symmetric matrix Σ is its covariance matrix, which describes the shape of the peak. (For a discussion of the properties of the multivariate Gaussian density function and the estimation of its parameters, see Morrison, 1990.) The constant c represents the probability mass of this part of the entire probability distribution. As before, suppose that the rest of the probability distribution is so far away from the window region that the data points arising from other parts of the distribution will have only a negligible effect on the computations. Then, in the window region, we will see that the windowed density $w(x)f(x)$ has a multivariate Gaussian shape.

I will rewrite the Gaussian function above: Let $B = \Sigma^{-1}$ and let $a = c \frac{1}{(2\pi)^{p/2}|\Sigma|^{1/2}}$. Then, in the window region,

$$f(x) = a\, e^{-\frac{1}{2}(x - \mu)'B(x - \mu)} .$$

As before, I will assume for simplicity that $\alpha = 0$. So the windowed density is

$$w(x)f(x) = a\, e^{-\frac{1}{2}[(x - \mu)'B(x - \mu) + x'Vx]} .$$

I will rewrite the expression in the brackets by completing the square. Let $A = B + V$. Since B is positive definite and V is non-negative definite, A is a positive definite symmetric matrix and is therefore non-singular. So we have

$$(x - \mu)'B(x - \mu) + x'Vx = x'Bx + x'Vx - 2\mu'Bx + \mu'B\mu$$

$$= x'Ax - 2\mu'BA^{-1}Ax + (\mu'BA^{-1})A(A^{-1}B\mu) - \mu'BA^{-1}B\mu + \mu'B\mu$$

$$= (x - A^{-1}B\mu)'A(x - A^{-1}B\mu) + \mu'(B - BA^{-1}B)\mu .$$

In the last term, $B - BA^{-1}B = BA^{-1}A - BA^{-1}B = BA^{-1}(A - B) = BA^{-1}V$. Therefore,

$$w(x)f(x) = a\, e^{-\frac{1}{2}\mu'BA^{-1}V\mu}\, e^{-\frac{1}{2}(x - A^{-1}B\mu)'A(x - A^{-1}B\mu)} .$$

This is a Gaussian function with "windowed mean" $A^{-1}B\mu$ and "windowed covariance matrix" A^{-1}. Note that this matrix depends on B and V but not on μ. If we write the windowed density as

$$\left[a\, e^{-\frac{1}{2}\mu'BA^{-1}V\mu} \frac{(2\pi)^{p/2}}{|A|^{1/2}}\right] \frac{|A|^{1/2}}{(2\pi)^{p/2}} e^{-\frac{1}{2}(x - A^{-1}B\mu)'A(x - A^{-1}B\mu)} ,$$

then the expression to the right of the brackets is an ordinary multivariate Gaussian (normal) density function, whose integral over the whole space is 1. Therefore the expression in the brackets is the integral of $w(x)f(x)$ over the space. As before, this quantity is the expected value of the weight $w_i = w(x_i)$ to be assigned to a randomly chosen data point x_i. We will therefore estimate the expression in the brackets by the average of the weights: $\frac{1}{N}\Sigma w_i$.

Since the windowed data have an approximate Gaussian shape, the simplest and most natural way to estimate the parameters of this shape — especially when p is large — is to compute the weighted sample mean and sample covariance matrix, by analogy with standard multivariate analysis. The sample mean vector is

$$\bar{x}_w = \frac{1}{\Sigma w_i}\Sigma w_i x_i$$

and the sample covariance matrix is

$$S_w = \frac{1}{\Sigma w_i}\Sigma w_i (x_i - \bar{x}_w)(x_i - \bar{x}_w)'$$

$$= \frac{1}{\Sigma w_i}\Sigma w_i x_i x_i' - \bar{x}_w \bar{x}_w' \ .$$

The element in the j^{th} row and the k^{th} column in this $p \times p$ matrix is the covariance of the j^{th} and k^{th} coordinates of the x_i. The matrix S_w is non-singular because I assumed that the data do not lie in a linear manifold of lower dimension; therefore it is a positive definite symmetric matrix. So $\bar{x}_w$ is an estimate of $A^{-1}B\mu$ and S_w is an estimate of A^{-1}.

We now "degauss" the view of the data as seen through the

Gaussian window; that is, we remove the effect of the window on $f(x)$ in the window region. Since V is known, and we can estimate A by $\hat{A} = S_w^{-1}$, we have

$$S_w^{-1} = \hat{A} = \hat{B} + V ,$$

so our estimate of B is

$$\hat{B} = S_w^{-1} - V .$$

We can estimate Σ by

$$\hat{\Sigma} = \hat{B}^{-1} = (S_w^{-1} - V)^{-1} ,$$

assuming that $S_w^{-1} - V$ is positive definite. To estimate μ we write

$$\bar{x}_w = \hat{A}^{-1} \hat{B} \hat{\mu} ,$$

so, assuming that $\hat{B}$ has an inverse, our estimate of μ is

$$\hat{\mu} = \hat{B}^{-1} \hat{A} \bar{x}_w = (S_w^{-1} - V)^{-1} S_w^{-1} \bar{x}_w .$$

We can also estimate the constants a and c. Since $\frac{1}{N} \Sigma w_i$ is an estimate of the expression in the brackets above, we have

$$\hat{a}\, e^{-\frac{1}{2} \hat{\mu}' \hat{B} \hat{A}^{-1} V \hat{\mu}} \frac{(2\pi)^{p/2}}{|\hat{A}|^{1/2}} = \frac{1}{N} \Sigma w_i ,$$

and, since $S_w = \hat{A}^{-1}$, our estimate of a is

$$\hat{a} = \frac{1}{N} \Sigma w_i \frac{1}{(2\pi)^{p/2} |S_w|^{1/2}} e^{\frac{1}{2} \hat{\mu}' \hat{B} S_w V \hat{\mu}} .$$

(The term $\hat{\mu}' \hat{B} S_w V \hat{\mu}$ in the exponent can also be written as $\bar{x}_w' V \hat{\mu}$.) Finally, we can estimate c by

$$\hat{c} = \frac{(2\pi)^{p/2}}{|\hat{B}|^{1/2}} \hat{a} .$$

In this case, where we have a single approximately Gaussian peak in the window region, we can describe its shape by the method of principal components (see Morrison, 1990). This method gives us a simple geometric description of the shape, which we can understand by analogy with the situation in two or three dimensions. The principal components are defined by a set of p mutually orthogonal eigenvectors of $\hat{B}^{-1} = \hat{\Sigma}$, which are imagined to emanate from $\hat{\mu}$. These eigenvectors define the principal axes of a family of concentric ellipsoids which form the contours of the estimated density function. The lengths of these axes are proportional to the square roots of the corresponding eigenvalues of $\hat{B}^{-1}$. The estimated density function is then the product of p univariate Gaussian density functions, each lying along a principal axis. The variance of each of these univariate densities is the eigenvalue corresponding to the axis for that density. We will work out the details in the general case below.

Note that the estimates of μ, Σ, a, and c may be unreliable or even meaningless by themselves unless the Gaussian shape we assumed above lies mostly within the window region. In the fully general case to be discussed below, we will estimate B as above, where B is the matrix defining the local shape of $f(x)$, and we will do a more careful analysis, based on the eigenvalues and eigenvectors of $\hat{B}$, rather than of $\hat{B}^{-1}$. ($\hat{B}$ and $\hat{B}^{-1}$ have the same eigenvectors, and the corresponding eigenvalues of $\hat{B}$ and $\hat{B}^{-1}$ are reciprocals of each other.)

To make the above argument more general, suppose that in the window region $f(x)$ can be approximated by

$$f(x) = a\, e^{-\frac{1}{2}(x - \mu)'B(x - \mu)},$$

where B is symmetric and non-singular but not necessarily positive definite. That is, each of its eigenvalues may be positive or negative, but not 0. (The fully general case will be considered next.) If we multiply this function by $w(x)$ and then complete the square in the exponent, using the same algebraic steps as before, we find that the windowed density $w(x)f(x)$ has a Gaussian shape. I will assume here that $A = B + V$ is positive definite, so that $w(x)f(x)$ does indeed look like a Gaussian density. In practice, $S_w = \hat{A}^{-1}$ is always positive definite, since it is computed from a finite set of data, and so is its inverse, $\hat{A}$. We estimate B, μ, and a by the same formulas as above. We will not estimate Σ or c here because they are meaningful only when $f(x)$ is shaped like an ordinary Gaussian density. As in the ordinary multivariate Gaussian case above, if the point $\hat{\mu}$ and an appreciable amount of the curvature of the density function appear in the window region, the estimates should be reliable, provided that they are based on a reasonable number of data points. But if the shape cannot be discerned in the window, the estimates of μ and a may be unreliable or even meaningless.

Assuming that the shape of the density function can be discerned in the window, we can describe the shape as we did above, based on the eigenvectors and eigenvalues of $\hat{B}^{-1}$. Again, the estimated density function is the product of p functions of one variable each. But in this case these functions are ordinary

Gaussian densities in the directions of the eigenvectors corresponding to the positive eigenvalues, and are "concave Gaussian" functions in the directions corresponding to the negative eigenvalues. At $\hat{\mu}$, where the estimated first derivatives of the density function are all 0, we could have a peak, a valley, or a saddle point, depending on the signs of the eigenvalues.

The reason that the cases considered so far are not general enough is that there can be extended structural features, part of which appear in the window region, and which also extend beyond the window region in some directions. A simple example is the following: Suppose that $p = 2$, and that near the origin (say, within a radius of 4 or 5) the density function is approximately

$$f(x) = a\, e^{-\frac{1}{2} x_1^2} .$$

This function represents a long, narrow ridge, whose center line lies along the X_2-axis. The value of $f(x)$ along the center line, or the crest of the ridge, is a. The cross-section of the ridge orthogonal to the center line at any point along that line is proportional to a Gaussian density function with standard deviation 1. If a large random sample is chosen from such a probability distribution, there will be a concentration of points in the vicinity of the X_2-axis, in accordance with this part of the density function. If we then view the data through a Gaussian window centered somewhere near the origin, we will see part of this concentration of points near the X_2-axis, and we will also see that this feature extends beyond the window in both

directions. This is much like what we would see if we were looking at an ordinary bivariate Gaussian shape for which the standard deviation of x_2 was very large.

The function above cannot be treated by the method given earlier because we cannot express the exponent as a quadratic form with a non-singular matrix B. However, if we allow B to be singular we can write the function as

$$f(x) = a\, e^{-\frac{1}{2} x'Bx},$$

where $x = (x_1, x_2)'$ and $B = \begin{pmatrix} 1 & 0 \\ 0 & 0 \end{pmatrix}$, a matrix of rank 1. This matrix has an eigenvalue 1, which is related to the width of the cross-section of the ridge, and an eigenvalue 0, which can be thought of as representing an "infinite" variance along the X_2-axis. Note that there is not a unique center point μ for this function. If we estimate B by $\hat{B}$ as we did earlier, we will probably find that $\hat{B}$ has an eigenvalue near 1 and an eigenvalue near 0, which might be positive or negative. Since $\hat{B}$ will probably be non-singular, we can invert it, but we can see that the resulting $\hat{B}^{-1}$ will be very unstable; that is, a small change in the data or in the window parameters might make a big difference in $\hat{B}^{-1}$. Also, we should not try to estimate μ by $\hat{\mu}$ here, because $\hat{\mu}$ depends on $\hat{B}^{-1}$. Not only would $\hat{\mu}$ be unstable, but it would be meaningless in this case. However, we will be able to estimate the location of the center line of the ridge.

We now come to the most general case, which I will treat in a way that the quantities to be estimated will be related to the

part of the density seen in the window, so that the estimates will be relatively stable. I will do this by working with $\hat{B}$ instead of with $\hat{B}^{-1}$. We will be able to deal with structural features that appear entirely within the window region, and also with features such as ridges that extend beyond the window region. The following analysis is the central part of this paper, and it is the basis for the computations done by a computer program I wrote to test the method.

Assume that in the window region the density $f(x)$ can be approximated by the exponential of a polynomial of degree at most two in the coordinates of x. I mean by this that the approximation is relatively good near the center of the window, and that as we move away from the window center, larger deviations between the true density and the approximation become more tolerable, in inverse proportion to $w(x)$. The second-degree terms of such a polynomial can be expressed as a quadratic form in x with a symmetric matrix, and the linear terms can be expressed as $r'x$ for some vector r. Any constant term in the polynomial can be absorbed in the multiplicative constant h below. So I will approximate the density in the window region by

$$f(x) = h\, e^{-\frac{1}{2} x'Bx + r'x},$$

where the number h, the vector r, and the symmetric matrix B are unknown parameters. Note that $h = f(0)$, and that I am still assuming that the window is centered at 0. If B is singular, there is not a unique center point μ for the function, as in the ridge example above. If B is non-singular, we could

complete the square and express $f(x)$ in the form given earlier. However, as a practical matter, if some eigenvalues of B are near 0 (relative to the window size), that is, if B is close to singular, then we will not be able to estimate μ reliably, even though it is uniquely defined. So I will not assume that we can reliably invert $\hat{B}$ or estimate μ.

The windowed density is

$$w(x)f(x) = h\, e^{-\frac{1}{2}[x'Bx - 2r'x + x'Vx]} .$$

Let $A = B + V$. As before, I will make the additional assumption that A is positive definite. This amounts to assuming that $f(x)$ in the window region is not too strongly concave upward in any direction, so that multiplying it by the window function $w(x)$ will pull it down into a shape roughly like a Gaussian density function. In practice we do not have to worry about the possibility that A may not be positive definite, because we will estimate A by S_w^{-1}, a positive definite matrix. The expression in the brackets above can be rewritten as

$$\begin{aligned} x'Bx - 2r'x + x'Vx &= x'(B + V)x - 2r'x \\ &= x'Ax - 2r'A^{-1}Ax + (r'A^{-1})A(A^{-1}r) - r'A^{-1}r \\ &= (x - A^{-1}r)'A(x - A^{-1}r) - r'A^{-1}r . \end{aligned}$$

So $w(x)f(x)$ is

$$h\, e^{\frac{1}{2} r'A^{-1}r}\; e^{-\frac{1}{2}(x - A^{-1}r)'A(x - A^{-1}r)}$$

$$= \left[h\, e^{\frac{1}{2} r'A^{-1}r}\, \frac{(2\pi)^{p/2}}{|A|^{1/2}} \right] \frac{|A|^{1/2}}{(2\pi)^{p/2}}\, e^{-\frac{1}{2}(x - A^{-1}r)'A(x - A^{-1}r)} .$$

Since A is assumed to be positive definite, this function is a multivariate Gaussian shape with windowed mean $A^{-1}r$ and covariance matrix A^{-1}. The expression in the brackets is the integral of $w(x)f(x)$ over the entire space. Taking these three quantities to be the parameters of this Gaussian shape, we estimate them by $\bar{x}_w$, S_w, and $\frac{1}{N}\Sigma w_i$, respectively. These estimates give us an overall description of the shape of the weighted data, smearing out any fine structure that may be present. Since I assumed that the data do not lie in a linear maniford of lower dimension, S_w is non-singular; therefore it is positive definite, as is its inverse.

We now degauss the estimated shape of the windowed (weighted) data. As before, A is estimated by $\hat{A} = S_w^{-1}$, and B is estimated by $\hat{B} = S_w^{-1} - V$. Since $\bar{x}_w = \hat{A}^{-1}\hat{r}$, we estimate r by

$$\hat{r} = \hat{A}\,\bar{x}_w = S_w^{-1}\,\bar{x}_w \,.$$

What we have done here is that we have avoided estimating μ explicitly, for which we would have had to invert $\hat{B}$. We can also estimate h. Since we have

$$\hat{h}\, e^{\frac{1}{2}\hat{r}'\hat{A}^{-1}\hat{r}}\, \frac{(2\pi)^{p/2}}{|\hat{A}|^{1/2}} = \frac{1}{N}\Sigma w_i \,,$$

and since $\hat{r}'\hat{A}^{-1}\hat{r} = (\bar{x}_w' S_w^{-1}) S_w (S_w^{-1}\,\bar{x}_w) = \bar{x}_w' S_w^{-1}\,\bar{x}_w$, we find

$$\hat{h} = \frac{1}{N}\Sigma w_i\, \frac{1}{(2\pi)^{p/2}|S_w|^{1/2}}\, e^{-\frac{1}{2}\bar{x}_w' S_w^{-1}\,\bar{x}_w} \,.$$

This is the estimated density at the window center, rather than

the estimated density at $\hat{\mu}$, which we earlier called $\hat{a}$.

We can now analyze the estimated shape of $f(x)$ by a method analogous to the method of principal components. Let $\lambda_1, \lambda_2, \cdots, \lambda_p$ be the eigenvalues of $\hat{B}$, and let $z_1, z_2, \cdots, z_p$ be a set of eigenvectors corresponding to the λ_j, chosen so that they are mutually orthogonal and each of unit length. (The z_j are not uniquely determined by these conditions, but that does not matter.) Let Z be the matrix whose *columns* are the z_j. The matrix Z is orthogonal; that is, $Z' = Z^{-1}$. Then

$$Z'\hat{B}\,Z = \begin{pmatrix} \lambda_1 & & 0 \\ & \ddots & \\ 0 & & \lambda_p \end{pmatrix} = L ,$$

a diagonal matrix.

We will now make a change of coordinates so that the z_j form an orthonormal basis for the new coordinate system. A vector x in the original coordinate system is represented by $y = Z'x$ in the new coordinate system. That is, the j^{th} coordinate of the point x in the new coordinate system is $y_j = z_j'x$. We also have $x = Zy$. The quadratic form $x'\hat{B}x$ in the old system, which is a function of x, becomes

$$x'\hat{B}x = y'Z'\hat{B}\,Zy = y'Ly = \sum_{j=1}^{p} \lambda_j y_j^2$$

in the new coordinate system. The function $\hat{r}'x$ becomes

$$\hat{r}'x = \bar{x}_w'S_w^{-1}x = \bar{x}_w'S_w^{-1}Zy = t'y = \sum_{j=1}^{p} t_j y_j ,$$

where $t = Z'S_w^{-1}\,\bar{x}_w$ and t_j, the j^{th} coordinate of the vector

t, is defined by $t_j = z_j' S_w^{-1} \bar{x}_w$.

We can now write the estimated density function as

$$\hat{f}(x) = \hat{h}\, e^{-\frac{1}{2} x'\hat{B}x + \hat{r}'x}$$

$$= \hat{f}(Zy) = \hat{h}\, e^{-\frac{1}{2} \Sigma\lambda_j y_j^2 + \Sigma t_j y_j}$$

$$= \hat{h} \prod_{j=1}^{p} e^{-\frac{1}{2} \lambda_j y_j^2 + t_j y_j} .$$

The estimated density is now a product of p functions of one variable each, where each of these functions is either an ordinary univariate Gaussian function if $\lambda_j > 0$, or a "concave Gaussian" function if $\lambda_j < 0$. If $\lambda_j = 0$, the function is an exponential function or a constant. If $\lambda_j > 0$, then λ_j^{-1} is the variance of the Gaussian shape, and $\lambda_j^{-1/2}$ is its standard deviation. If $\lambda_j < 0$, we can interpret $(-\lambda_j)^{-1/2}$ as a scale parameter analogous to the standard deviation. In either case, λ_j is related to the curvature of the function.

For any j for which $\lambda_j \neq 0$, we can complete the square for that j, if we wish:

$$-\frac{1}{2} \lambda_j y_j^2 + t_j y_j = -\frac{1}{2} \lambda_j \left(y_j^2 - 2 \frac{t_j}{\lambda_j} y_j + \frac{t_j^2}{\lambda_j^2} \right) + \frac{t_j^2}{2\lambda_j}$$

$$= -\frac{1}{2} \lambda_j \left(y_j - \frac{t_j}{\lambda_j} \right)^2 + \frac{t_j^2}{2\lambda_j} .$$

If we let $y_j = \frac{t_j}{\lambda_j}$, that is, if we move along the axis vector z_j for a distance of $\frac{t_j}{\lambda_j}$, we come to the "center" of the function of

y_j along that direction. At this point we have either a maximum or a minimum of the j^{th} function in the product above, depending on the sign of λ_j. It follows that the point $\frac{t_j}{\lambda_j} z_j$ is the nearest point to the origin for which that function is maximized or minimized. If λ_j is near 0, then instead of completing the square along the direction of z_j, we may want to assume that we have, approximately, an exponential function or a constant in that direction. Geometrically, this amounts to concluding that, along this direction, we are looking at part of a large structure, such as a ridge or a gradual slope, that extends beyond the window region.

If none of the λ_j is 0, so that $\hat{B}^{-1}$ exists and $\hat{\mu}$ is defined, then the point $\frac{t_j}{\lambda_j} z_j$ is the projection of $\hat{\mu}$ on the line generated by z_j. (It is easy to show that $\hat{\mu} \cdot z_j = \frac{t_j}{\lambda_j}$.) Thus, even if $\hat{\mu}$ does not lie in the window region, and hence is not a stable quantity, some of its components may be reliable estimates of aspects of the data structure within the window region. (Even if $\hat{\mu}$ is not defined at all, we can compute some of what would be its components, for those λ_j not too close to 0.) For example, if $p = 2$ and a long, narrow ridge runs through the window region, $\hat{B}$ would have a positive eigenvalue, say λ_1, corresponding to an eigenvector z_1 perpendicular to the ridge, and an eigenvalue λ_2 near 0, corresponding to an eigenvector z_2 parallel to the ridge. The estimated width of the ridge would be proportional to $\lambda_1^{-1/2}$, the standard

deviation of the univariate Gaussian function of y_1, which describes the cross-section of the ridge. The point on the estimated crest, or center line, of the ridge nearest to the window center would be $\frac{t_1}{\lambda_1} z_1$, and the set of points comprising the crest would be the line parallel to z_2 through this point. The value of $\hat{f}(x)$ along the crest of the ridge might be constant or it might be gradually increasing or decreasing; its behavior would be described by the function of y_2 in the product of functions above. These considerations will be useful in describing what is seen in the window region, and also in deciding where to place the next window. For example, we might want to move the window center to the nearest point on the crest of the ridge, which would be $\frac{t_1}{\lambda_1} z_1$, and try a window there, or, if the window center is already on or very close to the crest, we might want to move along the estimated crest of the ridge and try a window centered somewhere along that line.

Above we estimated $h = f(0)$. If we want, we can also estimate $f(x)$ for any x in the window region, using the estimated parameters given above. Of course, as we move away from the window center the estimated values become less reliable. However, I believe that often it will be more important to describe the shape, or structure, of $f(x)$, rather than to estimate its value at particular points. By expressing $\hat{f}(x)$ as a product of functions of one variable, we have a way to describe and think about the local structure of the data, even if p is large. I think that computationally, the simplest way to do this

in many dimensions is the method presented here.

It may also be useful to estimate the slope and the curvature of the density function. It will be easier, however, to work with the derivatives of $\log \hat{f}(x)$, since this function is a polynomial of degree at most two. In the new coordinate system, in which $y = Z'x$, let

$$g(y) = \log \hat{f}(x) = \log \hat{h} - \frac{1}{2} \Sigma \lambda_j y_j^2 + \Sigma t_j y_j \ .$$

Then, for $j = 1, \cdots, p$:

$$\frac{\partial g}{\partial y_j} = -\lambda_j y_j + t_j$$

and

$$\frac{\partial^2 g}{\partial y_j^2} = -\lambda_j \ .$$

For $j \neq k$,

$$\frac{\partial^2 g}{\partial y_j \partial y_k} = 0 \ .$$

Thus, our assumptions as to the form of $f(x)$ amount to assuming that the second derivatives of $\log f(x)$ are constant over the window region. The above first derivatives, which indicate the relative rate of change of the density, are linear functions.

There are two special cases worth considering. We could assume that the density in the window region is approximately constant:

$$f(x) = h \ ,$$

or that it is approximately an exponential function:

$$f(x) = h \, e^{r'x} \ .$$

In other words, we let $B = 0$. As before, $h = f(0)$. When the window region does not contain much data we might want to assume that the density has one of these forms, since there are fewer unknown parameters to be estimated. Actually, in the first case we could just as well use any kind of window, and in the second case there are some other kinds of windows that could be used. I will give the estimates of the parameters using Gaussian windows, since in practice I would use a Gaussian window anyway, and then decide how to interpret the results. In these cases I will have to assume that V, the window matrix, is non-singular.

In the exponential case the windowed density is

$$w(x)f(x) = h\, e^{-\frac{1}{2}[x'Vx - 2r'x]},$$

and the expression in the brackets is

$$x'Vx - 2r'x = x'Vx - 2r'V^{-1}Vx + (r'V^{-1})V(V^{-1}r) - r'V^{-1}r$$

$$= (x - V^{-1}r)'V(x - V^{-1}r) - r'V^{-1}r .$$

So $w(x)f(x)$ is

$$h\, e^{\frac{1}{2} r'V^{-1}r}\; e^{-\frac{1}{2}(x - V^{-1}r)'V(x - V^{-1}r)}$$

$$= \left[h\, e^{\frac{1}{2} r'V^{-1}r}\, \frac{(2\pi)^{p/2}}{|V|^{1/2}} \right] \frac{|V|^{1/2}}{(2\pi)^{p/2}}\, e^{-\frac{1}{2}(x - V^{-1}r)'V(x - V^{-1}r)} .$$

The windowed mean, $V^{-1}r$, is estimated by $\bar{x}_w$. Since the covariance matrix of $w(x)f(x)$ is V^{-1}, S_w should be close to V^{-1}, and we do not have to compute S_w unless we want to use it to check our assumptions, or to decide which functional form to use. Since $\bar{x}_w = V^{-1}\hat{r}$, we can estimate r by

$$\hat{r} = V \bar{x}_w .$$

This vector is the estimated gradient of $\log f(x)$. The expression in the brackets above can be estimated as usual by $\frac{1}{N} \Sigma w_i$. Therefore we can estimate h, since

$$\hat{h}\, e^{\frac{1}{2} \hat{r}' V^{-1} \hat{r}} \frac{(2\pi)^{p/2}}{|V|^{1/2}} = \frac{1}{N} \Sigma w_i .$$

Since $\hat{r}' V^{-1} \hat{r} = (\bar{x}_w' V) V^{-1} (V \bar{x}_w) = \bar{x}_w' V \bar{x}_w$, we have

$$\hat{h} = \frac{1}{N} \Sigma w_i \frac{|V|^{1/2}}{(2\pi)^{p/2}} e^{-\frac{1}{2} \bar{x}_w' V \bar{x}_w} .$$

If $f(x)$ is assumed constant, we have the previous case with $r = 0$, so

$$w(x) f(x) = \left[h \frac{(2\pi)^{p/2}}{|V|^{1/2}} \right] \frac{|V|^{1/2}}{(2\pi)^{p/2}} e^{-\frac{1}{2} x' V x} ,$$

and we can estimate h by

$$\hat{h} = \frac{1}{N} \Sigma w_i \frac{|V|^{1/2}}{(2\pi)^{p/2}} .$$

These special cases are useful for checking to see whether an estimated shape for $f(x)$ based on a large window is valid in small subregions within that window, or whether there is some fine structure that was obscured by using the large window.

USING GAUSSIAN WINDOWS

In this section I will give some examples to illustrate the use of Gaussian windows to explore a set of data. In order to devise strategies for moving about in the space and choosing

windows to try, we must have some idea of what kinds of structural features might be present in the data, and also of how they would appear in a Gaussian window. I will interpret the results of using a window by applying the analysis given in the previous section. I said earlier that I am assuming only that the density function is more or less smooth. What I mean by this is that the density is smooth in most of the space; it may be that there are some places where the density changes abruptly. The intent of this assumption is that if we look at the number and location of the data points in a limited region, we should be able to infer something about the population from which the data were drawn, at least for that region. Since I make no assumptions about the large-scale structure of the data, what we can do is study the local structure in small regions, and then try to put that information together into a description of the structure on a larger scale. Of course, if we have any more specific ideas about the possible structure of the data, we may be able to use statistical methods designed to deal with that structure, or we may be able to use our ideas about the structure to help guide us as we move about in the space, probing the data with Gaussian windows.

First, a few words about the geometry of high-dimensional spaces. It seems to be more natural to define limited regions in terms of p-dimensional spheres and ellipsoids, rather than p-dimensional cubes and rectangular solids. The interior of a sphere is the set of points within a given distance of the center point; a cube, on the other hand, contains some points that are

far away from the center, but not some of the points that are much closer to the center. For example, if the length of each edge of a p-dimensional cube is 2, then the distance from the center of the cube to the nearest point on the surface of the cube is 1, while the distance from the center to a vertex is $\sqrt{p}$. Thus, if p is large, a cube has a "pointy" shape, compared to a sphere.

Consider the "spherical" multivariate Gaussian density function with mean vector $\mu = 0$ and covariance matrix $\Sigma = I$. The contours of the density function are spheres about the origin. Each component x_j of the random vector x is a random variable independent of the others and has a Gaussian distribution with mean 0 and standard deviation 1. The distance from the origin to a "typical" point drawn from this probability distribution is about $\sqrt{p}$, since $E(\Sigma x_j^2) = p$. Thus, if p is large, a "typical" point comes from a region where the density is much smaller than the density near the origin. The reason for this is that the volume of a region of a certain shape increases according to the p^{th} power of the region's size, so, if p is large, the volume of the central region where the density is large is comparatively small.

The shape of a Gaussian window is defined by the symmetric matrix V. If V is positive definite, then its inverse would be the covariance matrix of the multivariate Gaussian density function that is proportional to the window function $w(x)$. This gives us a way of thinking about the shape of the window. If we do a principal components analysis of V^{-1}, we see that the

window has essentially an ellipsoidal shape with principal axes defined by the eigenvectors of V^{-1}. The square root of an eigenvalue of V^{-1} is the standard deviation of the component of the window function in the direction of the corresponding eigenvector. I think of the "window region" as being roughly the region within about two standard deviations of the window center, at least if p is not very large. For large p, we may have to think of the window region as being broader than that. As with the multivariate Gaussian density above, the amount of data several standard deviations from a, the window center, may be so great, compared to the amount nearer to a, that the data points at the greater distance would have a predominant influence on the weighted sample mean and covariance matrix, even though $w(x)$ at that distance is small. Whether or not V has an inverse, we can decompose $w(x)$ into a product of functions of one variable each, as in the decomposition in the previous section. If V does not have an inverse, then V has some 0 eigenvalues, and the standard deviation of the window in the direction of the corresponding eigenvectors may be thought of as infinite. If we choose $V = 0$, then $w(x) = 1$ for all x, and we obtain the ordinary unweighted sample mean and covariance matrix (with denominator N). I usually use this for my first window.

Suppose that we try a Gaussian window and we find that all of the λ_j, the eigenvalues of $\hat{B}$, are positive and not too close to 0, and that $\hat{\mu}$ is in the window region. This would indicate a peak in the density, that is, a cluster of data points, centered at $\hat{\mu}$, and with a shape described by $\hat{B}$. Since the

standard deviations along the principal axes of the peak are the $\lambda_j^{-1/2}$, a large λ_j means a small standard deviation, indicating that if the data points in the window were projected onto the line generated by the corresponding eigenvector, they would be highly concentrated. A small λ_j means a large standard deviation in the corresponding direction, indicating that the data points are more spread out. If the standard deviation is more than one or two times the standard deviation of the window in that direction, we may be looking at a part of a structure that extends beyond the window region, or at least extends into its outer reaches, where $w(x)$ is small. In that case the estimates of the parameters may not be very reliable, at least in the directions corresponding to the small λ_j. (My computer program converts the λ_j to standard deviations so that it is easier to understand the results and make judgments about them.)

If we find an apparent peak in the window region, a natural next step is to try a window centered at $\hat{\mu}$, in order to obtain better estimates of the parameters of the peak. (If $\hat{\mu}$ is far from the current window center, we may want to be more cautious and move toward it in a series of steps.) We must also choose a shape for the next window. When changing the window center, I would often use the same shape for the next window that I used for the current window, so that I can think of the results of using the two windows as being comparable to each other. Whatever shape we use, the usual result of the next window is that we find that we are not exactly at the center of the peak. So we might want to try several more windows to pin down the

center and shape of the peak more precisely. Since the computational effort is not too great, unless p and N are very large, it is not hard to do this. However, in practice, a cluster of data points will generally not be exactly Gaussian in shape. Moreover, there will usually be some overlap between this peak and other parts of the data, and this mixture of points in the data will affect the estimates of the parameters. Consequently, each different window we might try would give somewhat different results. There is no single "right" window to use; therefore, there will not be a single right answer for the estimated parameters of the cluster. Trying different windows to get better estimates can be like trying to hit a moving target, and we can end up wasting time chasing after the best view of the peak and not finding it. Instead, we must content ourselves with an approximate description of the location, shape, and height of the peak. The important thing is that we have found a peak, or a cluster, and that we have an approximate description of it. By finding a peak, we have identified a structural feature which will be an element of our ultimate description of the data.

If we do have an approximate Gaussian peak in our sights, classical statistical theory suggests that the best estimates of its parameters would be obtained by using the unweighted sample mean and covariance matrix. But since there are other data points that are not part of this peak, we do not want to do that, so the best strategy would be to choose a window that gives as much weight as possible to the peak, and at the same time gives as little weight as possible to any nearby data points that are

not part of the peak. That is, we want to mask out those other data points as much as we can. For $p > 1$ we can choose among many possible window shapes. If we look at the peak with several Gaussian windows covering approximately the same region, and if the resulting estimates of the parameters are consistent with one another, we can have some confidence that the estimates are reliable. Since there is always the danger that we may be misled by results based on a window region containing few data points, the best practical safeguard against this is to view the data in the region with several windows and observe consistent results.

Suppose now that not all of the λ_j are positive, or that not all of them are safely away from 0, but that $\hat{\mu}$ is in the window region. Then, depending on the signs of the λ_j, we can think of $\hat{\mu}$ as the location of a peak, or a valley (a relative minimum in the density), or a saddle point. In such cases we are necessarily looking at a part of a structure that extends beyond the window region. In each of these cases, $\hat{\mu}$ is a point where the first derivatives of the estimated density function are 0. Such points, if we can find them, are useful because it is easy for us to think about them, and we can apply our geometrical intuition to them in higher dimensions. A simple example is a saddle point that might appear between two clusters that are near each other and have some overlap. We would expect to find a sort of ridge leading from one peak to the other. At the lowest point (the point of least density) along the ridge, we would probably find a saddle point with one negative λ_j and $p - 1$ positive λ_j. The negative λ_j would correspond to an eigenvector

parallel to the crest of the ridge at the saddle point, because the density curve is concave upward along the crest. The other λ_j would be positive because moving away from the saddle point in any direction orthogonal to the crest would mean moving to a point where the density is less than it is on the crest of the ridge. As discussed in the previous section, we can interpret $\lambda_j^{-1/2}$ as a standard deviation for positive λ_j, and $(-\lambda_j)^{-1/2}$ as an analogous scale parameter for negative λ_j. As we did above with peaks, we might want to try a window centered at $\hat{\mu}$ in order to get a better estimate of the local structure. All of the considerations above apply here. There is no "right" window to use, and therefore no single right answer. The important thing, again, is that we have found a pivotal point in the space that will be useful in thinking about and describing the structure of the data, even if we cannot estimate its parameters precisely.

Consider the example of a "ridge" in the density function. An example of such a ridge occurs in the luminosity-temperature diagram familiar to astronomers. One of the advantages of exploring the data with Gaussian windows is that we can find extended structures of this kind. That is, we do not have to assume that the data points are concentrated in a number of clusters, each of very limited extent. It might be better to think of a ridge as a kind of "bar", that is, an essentially one-dimensional structure, or concentration of data points, extending for some distance through the p-dimensional space. I do not mean by this that the data points comprising the bar lie

in a one-dimensional manifold; what I mean is that there is a line or a one-dimensional curve that acts as a "center line" for this subset of the data points, and that these points are distributed in all directions about that center line. I will assume that although these points are concentrated about a center line, they do not lie in a linear manifold of dimension less than p. The center line may be straight or it may curve gradually; it could twist in any direction as it runs through the p-dimensional space. For any point along the center line, the density function has a (p-1)-dimensional cross-section orthogonal to the center line at that point. The shape of the cross-section could vary as we move along the center line.

If we try a Gaussian window for which a bar passes through the window region, we will find one eigenvalue, say λ_p, very near 0 (it could be positive or negative), indicating a structure extending beyond the window region, with the corresponding eigenvector, z_p, parallel to the estimated center line of the bar. The other $p - 1$ eigenvalues will be positive and not too close to 0, indicating that the data points are more concentrated in the corresponding directions; they and their eigenvectors will describe the estimated (p-1)-dimensional cross-section of the bar, or at least the average cross-section in the window region. Since we find an eigenvalue near 0, indicating that we are looking at a structure extending beyond the window region, we will not try to estimate μ. However, we do want to estimate the point on the center line of the bar closest to the window center (which I assume is at the origin). To do this we change to the

coordinate system based on the z_j, the eigenvectors of $\hat{B}$, so that $\hat{f}(x)$ becomes a product of p functions of one variable each. We maximize each of the first $p - 1$ of these functions, for which λ_j is positive and not near 0, by letting $y_j = \frac{t_j}{\lambda_j}$, that is, by moving that distance from the window center in the direction of z_j. It follows that the maximum of $\hat{f}(x)$ over the (p-1)-dimensional subspace orthogonal to z_p is attained at

$$\sum_{j=1}^{p-1} \frac{t_j}{\lambda_j} z_j .$$

This is the point on the estimated center line of the bar that is closest to the window center. The estimated center line is the line parallel to z_p through this point. Note that this point may not be the maximum point for $\hat{f}(x)$ over the entire space; moving one way or the other along the estimated center line might increase the value of $\hat{f}(x)$. The p^{th} function of one variable, which we did not use above, gives us an estimate of the density function along the center line.

We can now consider choosing a window to try next. If the current window center is not on or very near the center line, we may want to move the window center to the point defined above, so that we can try a window centered on the estimated center line. This window should give us a better estimate of the shape of the bar. (If the point above is far from the current window center, we may want to move toward it in steps, for example by including in the sum above only those terms for which we think the estimated coefficient of z_j is reliable.) Once we are on or

very near the center line of the bar, the natural thing to do is to move along the center line, that is, to change the window center by a multiple of z_p. We should move in short hops along this line, so that the next window has some overlap with the current window. Since the center line may curve, and since, even if it does not, our estimate of its direction is only approximate, the new estimate of the center line, based on the new window, will probably not go through the center of that window. So we could then move the window center to the point on the new estimate of the center line closest to the new window center, as we did above, and then resume moving along the center line. In this way we can follow along the center line as far as we can in both directions, and map out a description of where the center line goes and what the cross-section of the bar looks like. Some experiments with artificial data show that this can be done. Note that I am not recommending moving along the gradient. This is because I am not looking for a relative maximum; instead, I am trying to understand and describe the shape of the density function by studying the structural features found in the data.

There could also be similar structures of higher dimension in the data. For example, the data points in a region could be concentrated in an essentially two-dimensional structure like a "pancake". That is, instead of a center line, the center of the pancake would be a two-dimensional "center sheet", with the data points distributed in all directions about the sheet. The sheet could be flat, or it could curve gradually as it runs through the

space. At any point on the sheet the density function would have a (p-2)-dimensional orthogonal cross-section, whose shape could vary from point to point. If we try a Gaussian window with such a structure passing through it, we would find two eigenvalues very near 0, indicating a structure extending beyond the window region; the two corresponding eigenvectors would define a plane parallel to the estimate of the center sheet in the window region. The other $p - 2$ eigenvalues would be positive and not too close to 0, indicating that the data points are more concentrated in the corresponding directions; they and their eigenvectors would describe the estimated (p-2)-dimensional cross-section of the pancake, or at least the average cross-section. If we find such a structure in a window, we can estimate the nearest point on its center sheet by forming a linear combination of the $p - 2$ eigenvectors orthogonal to the estimated center sheet, just like the estimate above of the nearest point on the center line of a bar. As before, we can then try a window centered at that point (or we can move toward it in steps), in order to get a better estimate of the shape of the structure. Once we have a window centered on or near the center sheet, we can move along the sheet by choosing a new window center somewhere in the estimated plane of the center sheet. Here we have to search in two dimensions, instead of simply following a curve; that is, we would have to try points in the plane to the north, south, east, and west, so to speak. After trying a window at such a point, we would probably find that the new window center is off of the center sheet, for the

same reasons as with the bar above, so we would want to move over to the nearest point on the new estimate of the center sheet, and then resume moving along the plane of the center sheet. If we continue this process in all directions, we will eventually map out a description of the extent and shape of the pancake, including its center sheet and its cross-section.

Similarly, we might find an essentially k-dimensional structure, for any k less than p. Such a structure would have a k-dimensional manifold as a "center", and a (p-k)-dimensional cross-section. We would recognize such a structure in a window by observing k eigenvalues near 0, indicating a structure extending beyond the window region in k dimensions, in the directions of the corresponding eigenvectors, and p - k positive eigenvalues not too close to 0, indicating that the structure is limited in extent in the corresponding directions. We could then try to follow the structure and map out its extent and shape, as with the examples above. To do this we would need a strategy for moving in all directions in a k-dimensional manifold and keeping track of the results. Note that I have been vague about how near 0 an eigenvalue has to be to indicate a structure extending beyond the window region. As a rule of thumb I consider a standard deviation more than one or two times the standard deviation of the window in the corresponding direction to be an indication that the structure extends beyond the window region. I think that it would be unwise to try to give a definite cutoff point for the size of λ_j, since any such rule would be arbitrary. Consequently, the dimension k of an

apparent structure in the data would not be specified, at least not at first. An analogous situation occurs in principal components analysis, where it is often unclear as to how many of the principal components to regard as significant. Since we are exploring the data interactively, and since it is not costly to try several windows, I think it is better not to commit ourselves to a specific value of k until the data have been explored rather thoroughly. Since we are free to move about in the space, we can move the window center along whichever eigenvectors we want; that is, we can try different possibilities without deciding in advance which eigenvectors define the center of the structure and which define its cross-section. For example, we can begin by moving along only those eigenvectors for which we seem to have a good estimate of where the density is maximized in that direction. Then, as we move toward a region of higher density, we may obtain better estimates of the shape of the density function in other directions.

A Gaussian window focusses a spotlight on the data in a particular region. But the density function in that region may or may not satisfy the basic assumption that I have been making — that it can be approximated there by the exponential of a second-degree polynomial. In other words, there may be some "fine structure" in the data in that region. In fact, if a window contains a large amount of data, it is not unlikely that there will be some fine structure. Or, to put it the other way around, if a window contains only a small amount of data, we will not be able to tell whether any fine structure is present, and we

will have to be content with a simple, overall description of the data in that region. One way to look for fine structure would be to do a more sophisticated analysis of the data seen in a window. Instead of doing that, however, I will use subwindows; that is, I will try smaller Gaussian windows within the region of the given window, and I will compute the usual quantities for those windows. For a given window, we have an estimate, say $\hat{f}_1(x)$, of the density in the region of that window, derived from the estimated overall shape of the weighted data points. For any point x in the window region, we can try a small window centered at x, to see whether the estimated density at x based on the small window agrees with $\hat{f}_1(x)$, the estimate based on the large window. Since the true density might vary greatly from $\hat{f}_1(x)$ at any point in the region, there is no way to tell whether it does so, other than by looking at the data near that point. So what we can do is to try a series of small windows, each centered at one of a set of trial points spread out through the region, and compare the estimated density at those points with $\hat{f}_1(x)$. These points could be a systematic set of regularly spaced points, or a random set of points. If p is large, however, a set of points covering the entire window region might have to be a very large set. In that case, one way to choose a set of trial points would be to choose a number of the data points in the region at random. Since these points would tend to be where the bulk of the data points are, we would be checking the density in the places where it is probably most important to do so. In these small windows, there might be only a small

amount of data, in which case we might estimate the density at the window center using one of the simple special cases treated at the end of the previous section.

In two or three dimensions we can look for fine structure, or any other unexpected features in the data, by examining a scatter plot or other such graphical representation of the data. When we look at a scatter plot, we can move our eyes around the diagram and focus on any small part of it; that is how we discover features on a smaller scale. Trying subwindows of a Gaussian window is the analogue of this in higher dimensions. If p is large we can project the data, or a subset of the data, onto a space of lower dimension, so that we can then use a graphical technique for studying the data. See for example Chambers et al. (1983), Cleveland and McGill (1988), and Du Toit et al. (1986). But when we do this we risk obscuring the structural features we are trying to find, and we may be limiting the dimensionality of the features that we can find in this way. Since Gaussian windows can be used in any number of dimensions, I prefer using subwindows to search for fine structure when p is large. Of course, other methods for searching for fine structure could be used in conjunction with Gaussian windows.

I usually begin exploring a set of data by computing the unweighted sample mean and covariance matrix (that is, by using a window with $V = 0$). Then I use large windows to find the overall shape of the data for large regions, and then I work my way down to smaller windows. We can try windows as small as the data will allow. If a window is too small, the window region

will not contain enough data to give reliable estimates of the parameters, especially the quantities derived from $\hat{B}$. However, it is not clear how to tell whether a window region contains enough data to give reliable estimates. I think that in general the best thing to do is to try several windows with various centers and shapes covering approximately the same region, to see whether the results are consistent. If they are, we can be confident that what we think we see in the data is really there, at least for the purpose of describing the shape of the data. If we want to draw broader statistical inferences about the population from which the data were drawn, we will need to make additional assumptions about the process by which the data were generated.

It may be possible to estimate the standard deviations of the various estimates based on a window. Further work is needed to devise simple measures of accuracy for those estimates and to determine whether such measures would be useful. Any such measures of accuracy, however, would have to be taken with a grain of salt, since, after the first window used, the choice of the window parameters will be influenced by what was seen in the previous windows; that is, the succeeding windows will not be independent of the data.

After we have explored a set of data, we can put the results together into a final description of its structure. Such a description might include a list of the structural features found, with a description of each one, and of how they are related to one another. The list could include pivotal points

such as peaks, valleys, and saddle points, and also extended structures such as bars and pancakes, and similar structures in higher dimensions. The description can include as much or as little detail as is desired. Note that any structural feature might have a more detailed structure on a smaller scale; what appears to be a cluster or a bar in a window of a certain size might turn out, upon closer examination, to be composed of smaller structures that are not separately visible in a larger window. There may also be partial structures that merge into one another. For each feature, we could give its location, size, shape, and extent. We might also estimate its "mass", that is, the proportion of the data points that are part of that feature. For a cluster of Gaussian shape that can be viewed in a single window, we can estimate its mass by $\hat{c}$. For an extended structure we might be able to estimate its mass by considering its extent and its cross-section.

As with any new tool, using Gaussian windows takes some practice. Since I have kept the assumptions about the data to a minimum, the method is widely applicable. Since it is interactive, it is flexible and open-ended, and the user is free to experiment and to follow a variety of strategies. If we have some additional knowledge or beliefs about the data, we can use them to guide us in choosing windows to try and in interpreting the results. Also, the method can be used in conjunction with other methods. The method is computationally simple, compared to many other multivariate methods, and it can be implemented on a small computer. To implement the method, the user can

incorporate any standard algorithms for inverting a matrix and for finding the eigenvalues and eigenvectors of a symmetric matrix. The other computations are simple to program. Finally, the method presented here provides a way to apply our geometrical intuition, so that we can think about and describe the structure of a set of data in any number of dimensions.

I would like to thank Dr. Mike Raugh of RIACS for providing me with the opportunity and the freedom to do this work.

REFERENCES

Chambers, J., W. Cleveland, B. Kleiner, and P. Tukey (1983). *Graphical Methods for Data Analysis*. Wadsworth & Brooks/Cole.

Cheeseman, P., J. Kelly, M. Self, J. Stutz, W. Taylor, and D. Freeman (1988). AUTOCLASS: A Bayesian classification system. In *Proc. Fifth Machine Learning Workshop*, pp. 54-64. Morgan Kaufmann.

Cleveland, W., and M. McGill (eds.) (1988). *Dynamic Graphics for Statistics*. Wadsworth & Brooks/Cole.

Du Toit, S., Steyn, and Stumpf (1986). *Graphical Exploratory Data Analysis*. Springer-Verlag.

Morrison, D. (1990). *Multivariate Statistical Methods* (3rd ed.). McGraw-Hill.

Pao, Y-H. (1989). *Adaptive Pattern Recognition and Neural Networks*. Addison-Wesley.