SLAC – PUB – 4024
July 1986
M

A STATISTICIAN'S VIEW OF DATA ANALYSIS *

JEROME H. FRIEDMAN

Department of Statistics
Stanford University, Stanford, California, 94305

and

Computation Research Group
Stanford Linear Accelerator Center
Stanford University, Stanford, California, 94305

ABSTRACT

A brief overview of statistical data analysis is provided with a view towards examining its role in the analysis of astronomy data.

Invited paper presented at the International Workshop on
Data Analysis in Astronomy, Erice, Italy, April 17-30, 1986.

* Work partially supported by the Department of Energy, contract DE-A03-76SF00515.

1. INTRODUCTION

A question that is often asked in any experimental or observational science is whether statistical considerations are useful in the analysis of its data. This is a question that can only be answered by the scientists who understand the data as well as mechanisms and instruments that produce it. In order to help answer this question it is useful to know how data and data analysis is viewed by people who regard themselves as statisticians. It is the purpose of this report to give a necessarily brief overview of statistical data analysis as viewed and practiced by statisticians.

First, it is important to understand what statisticians do not regard as data analysis, but which is never-the-less an important aspect of data understanding. This is the process of data reduction. In this phase the raw data from the detectors (telescopes, counters) are reduced to move useful and understandable quantities (such as images). The software (and sometimes hardware) that perform this task are simply regarded as computing engines that transform the raw data to forms that are more convenient for further calculations. Although statistical considerations may be involved in the development of these systems, they are usually dominated by considerations specific to the scientific field and the particular instruments that produce the data.

It is the further calculations that interest statisticians. That is, how to discover from the (refined) data, the properties of the systems under study that produced the data (stars, galaxies, etc.), and deduce statistically meaningful statements about them, especially in the presence of uncertain measurements.

Statistics can be viewed as the science that studies randomness. Central to statistical data analysis is the notion of the random variable or measurement.

This is a measured quantity for which repeated observations (measurements) produce different values that cannot be (exactly) predicted in advance. Instead of a single value, repeated measurements will produce a distribution of values. The origin of the randomness can be due to random measurement errors associated with the instruments, or it could be a consequence of the fact that the measured quantity under consideration depends upon other quantities that are not (or cannot be) controlled – ie., held constant. In either case, a random variable is one for which we cannot predict exact values, only relative probabilities among all possible values the variable can assume.

The distribution of relative probabilities is quantified by the probability density function $p(X)$. Here X represents a value from the set of values that the variable can take on, and the function $p(X)$ is the relative probability that a measurement will produce that value. By convention the probability density function is required to have the properties

$$p(X) \geq 0 \text{ and } \int p(X)\, dX = 1$$

as X ranges over all of its possible values. Under the assumption that X is a random variable, the most information that we can ever hope to know about its future values is contained in its probability density function. It is the purpose of observation or experimentation to use repeated measurements of the random variable X to get at the properties of $p(X)$. It is the purpose of theory to calculate $p(X)$ from various mathematical (physical) models to compare with observation.

It is seldom the case that only one measurement is made on each object under study. Usually several simultaneous measurements of different quantities are made on each object, each of these measurements being a random variable.

In this case we can represent each observation as an n-vector of measurements

$$X_1, X_2, \ldots, X_n \tag{1}$$

where n is the number of simultaneous measurements performed on each object. We call the collection of measurements (1) a vector-valued random variable of dimension n.

Statistics as a discipline has several divisions. One such division depends upon whether one decides to study each of the random variables separately—ignoring their simultaneous measurement—or whether one uses the data (collection of simultaneous measurements) to try to access the relationships (associations) among the variables. The former approach is known as univariate statistics which reduces to studying each random variable X_i, and its corresponding probability density $P_i(X_i)$, separately and independently of the other variables.

The latter approach is known as multivariate statistics. Central to it is the motion of the joint probability density function

$$p(X_1, X_2, \ldots, X_n) \tag{2}$$

which is the relative probability that the simultaneous set of values $X_1, X_2, \ldots, X_n$ will be observed. In multivariate statistics one tries to get at the properties of the joint probability density function (2) based on repeated observation of simultaneous measurements.

Another division in the study of statistics is between parametric (model dependent) and nonparametric (model independent) analysis. We begin with a

little notation. Let

$$\underline{X} = (X_1, X_2, \ldots, X_n)$$

be an n – dimensional vector representing the simultaneous values of the n measurements made on each object. In parametric statistics the (joint) probability density function is assumed to be a member of a parameterized family of functions,

$$p(\underline{X}) \;=\; f(\underline{X}; \underline{a}), \tag{3}$$

where $\underline{a} \;=\; (a_1, a_2, \ldots, a_p)$ is a set of parameters, the values of which determine the particular member of the family. In parametric statistics the problem of determining the (joint) probability density function reduces to the determination of an appropriate set of values for the parameters. The parameterized family chosen for the analysis can come from intuition, theory, physical models, or it may just be a convenient approximation.

Nonparametric statistics, on the other hand, does not specify a particular functional form for the probability density, $p(\underline{X})$. It's properties are inferred directly from the data. As we will see, the histogram can be considered an example of a (univariate) probability density estimate.

Generally speaking, parametric statistical methods are more powerful than nonparametric methods provided that the true underlying probability density function is actually a member of the chosen parameterized family of functions. If not, parametric methods lose their power rapidly as the truth deviates from the assumptions, and the more robust nonparametric methods become the most powerful.

The final division we will discuss is between exploratory and confirmatory data analysis. With exploratory data analysis one tries to investigate the properties of the probability density function with no preconceived notions or precise questions in mind. The emphasis here is on detective work and discovering the unexpected. Standard tools for exploratory data analysis include graphical methods and descriptive statistics. Confirmatory data analysis, on the other hand, tries to use the data to either confirm or reject a specific preconceived hypothesis concerning the system under study, or to make precise probabilistic statements concerning the values of various parameters of the system.

For the most part this paper will concentrate on confirmatory aspects of data analysis with a few exploratory techniques (associated with nonparametric analysis) coming at the end.

2. Mini-Introduction to Estimation Theory

In estimation, we assume that our data, consisting of N observations, is a random sample from an infinite population governed by the probability density function $p(\underline{X})$. Our goal is to make inferences about $p(\underline{X})$. In parametric estimation we would like to infer likely values for the parameters. In nonparametric estimation we want to infer $p(\underline{X})$ directly.

Consider a parametric estimation problem. Here we have a data set $\{\underline{X}_i\}_{i=1}^N$ considered as a random sample from some (joint) probability density function $p(\underline{X})$ which is assumed to be a member of a parameterized family of functions $f(\underline{X};a)$ characterized (for this example) by a single parameter a. Our problem is to infer a likely value for (ie. estimate) a.

Let

$$Y = \phi(\underline{X}_1, \underline{X}_2, \ldots, \underline{X}_N) \tag{4}$$

be a function of N vector valued random variables. Here ϕ represents (for now) an arbitrary function. Since any function of random variables is itself a random variable, Y will be a random variable with its own probability density function $p_N(Y; a)$. This probability density function will depend on the joint probability density of the $\underline{X}_i$, $f(X_i; a)$, and through this on the (true) value of the parameter a. It will also depend on the sample size N. Suppose it were possible to choose the function ϕ in (4) so that $p_N(Y; a)$ is large only when the value of Y is close to that of a, and small everywhere else (provided the $\underline{X}_i$ follow $p(\underline{X}) = f(\underline{X}; a)$). If this were the case then we might hope that when we evaluate ϕ for our particular data set that the value for Y so obtained would be close to that of a. A function of N random variables is called a "statistic" and its value for a particular data set is called an "estimate" (for a).

As an example of how it is possible to construct statistics with the properties described above, consider the method of moments. Define

$$G(a) = \int g(\underline{X}) p(\underline{X}) d\underline{X} = \int g(\underline{X}) f(\underline{X}; a) d\underline{X} \tag{5}$$

where $g(\underline{X})$ is an arbitrary function of a single (vector valued) random variable. The quantity $G(a)$ is just the average of the function of $g(\underline{X})$ with respect to the probability density $p(\underline{X})$. Its dependence on the value of a is a consequence of the fact that $p(\underline{X}) = f(\underline{X}; a)$ depends upon the value of a. Now, the law of large numbers (central limit theorem) tell us that

$$Z = \Theta(\underline{X}_1, \underline{X}_2, \ldots \underline{X}_N) = \frac{1}{N} \sum_{i=1}^{N} g(\underline{X}_i) \tag{6a}$$

has a normal (Gaussian) distribution

$$p_N(Z;a) = \frac{1}{\sqrt{2\pi}\sigma_N} \exp\left\{-\tfrac{1}{2}[Z - G(a)]^2/\sigma_N^2\right\} \tag{6b}$$

centered at $G(a)$, with standard deviation

$$\sigma_N = \frac{1}{\sqrt{N}}\{\int [g(\underline{X}) - G(a)]^2 f(\underline{X};a)d\underline{X}\}^{\frac{1}{2}} \tag{6c}$$

as the sample size becomes large. That is, the sample mean (of $g(\underline{X})$) has a Gaussian distribution centered at the true mean with a standard deviation that becomes smaller as N grows larger ($\sim \frac{1}{\sqrt{N}}$). Therefore, for large enough N, likely values of Z will always be close to $G(a)$, and Z is a good statistic for estimating $G(a)$. If $g(\underline{X})$ is chosen so that $G(a)$ is not too wild a function of a, it then follows that

$$Y = G^{-1}(Z) = G^{-1}[\frac{1}{N}\sum_{i=1}^{N} g(\underline{X}_i)]$$

will be a good statistic for estimating the value for the parameter a.

Note that in this development the moment function $g(\underline{X})$ is fairly arbitrary. Therefore, this method can be used to construct a great many statistics for estimating the (same) parameter a. Some of these estimators will be better than others. The field of statistics is concerned to a large degree with finding good estimators (statistics for estimation).

Statisticians rate the quality of estimators on the basis of four basic properties: consistency, efficiency, bias, and robustness. Consistency concerns the property of the estimator as the sample size N becomes arbitrarily large. In

particular an estimator (4) is said to be consistent if

$$\lim_{N \to \infty} p_N(Y;a) = \delta(Y-a)$$

where δ is the Dirac delta function. For a consistent estimator, the estimate becomes more and more accurate as the sample size increases. Note that (6) implies that moment estimates are consistent provided that the bracketed quantity in (6c) is finite (second central movement of $g(\underline{X})$).

Efficiency is concerned with the properties of the estimator for finite N. The efficiency of an estimator is inversely related to its expected-squared-error

$$ESE_N(Y) = \int (Y-a)^2 f_N(Y;a)dY.$$

This is the average-squared distance of the estimate from the truth. Note that if the estimator is consistent, then $\lim_{N\to\infty} ESE_N(Y) = 0$. The relative efficiency of two estimators Y and Z is defined as the inverse ratio of their corresponding expected squared errors,

$$RE_N(Y,Z) = ESE_N(Z)/ESE_N(Y).$$

Bias is concerned with whether or not the average value of a statistic is equal to the true value of the parameter it is estimating. In particular, the bias of an estimator is defined to be

$$B_N(Y) = \int Y f_N(Y;a)dY - a.$$

This is just the difference between the average value of the statistic and the truth. Note that if an estimator is consistent then $\lim_{N\to\infty} B_N(Y) = 0$. An estimator for

which $B_n(Y) = 0$ for all N is said to be unbiased. Generally speaking unbiased estimators are preferred if all other things are equal. However, all other properties are seldom equal. In particular, the efficiency of the best unbiased estimators is generally lower than that for the best biased estimators in a given problem. Unbiased estimators are almost never best in terms of expected-squared-error.

Robustness concerns the sensitivity of an estimator to violations in the assumptions that went in to choosing it. In parametric statistics the assumptions center on the particular parameterized family (3) assumed to govern the probability density of the (random) variables comprising the data. For a given parametric family there is usually an optimal estimator for its parameters (in terms of efficiency). However, it is often the case that the efficiency of such an optimal estimator degrades badly if the true probability density deviates only slightly from the closest member of the assumed parameterized family. Robust estimators generally have a little less efficiency than the optimal estimator in any given situation (if the true density were known), but maintain their relatively high efficiency over a wide range of different parameterized forms for probability density functions. Robust estimators are generally preferred since it is often impossible to know for certain that the assumed parametric form for the probability density is absolutely correct.

As an example of robustness consider estimating the center of a symmetric distribution. If the probability density corresponding to the distribution were Gaussian, then the sample mean is the most efficient estimator. If, however, the distribution has higher density than the Gaussian for points far away from the center (fat tails), then the efficiency of the mean degrades badly. The sample median, on the other hand, is less efficient than the mean for Gaussian data

(relative efficiency approximately 0.64) but has much higher efficiency for fat tailed distributions.

Although the method of moments described above can be (and often is) used to construct estimators, it is not the favorite way among statisticians. Far and away the most popular method is that of maximum likelihood. By definition, the relative probability of simultaneously observing the set of values $\underline{X} = (X_1, X_2, \ldots, X_N)$ is the value of the joint probability density function $p(\underline{X})$. Let $\underline{X}_i (i = 1, N)$ be one of the observations in our data set. The relative probability of observing this observation (before we actually observed it) was $p(\underline{X}_i)$. If we believe that all of our N observations were independently drawn from a population governed by $p(\underline{X})$, then the relative probability of seeing all N of our observations (again in advance of actually seeing them) is simply the product of the probabilities for seeing the individual observations. Thus the relative probability among all possible data sets that we would have seen, the set of data that we actually saw, is

$$L_N(a) = \Pi_{i=1}^{N} p(\underline{X}_i) = \Pi_{i=1}^{N} f(\underline{X}_i; a).$$

This expression is known as the likelihood function. It is a function of the parameter a through the dependence of the probability density function on this parameter. The principal of maximum likelihood estimation is to choose as our parameter estimate that value that maximizes the probability that we would have seen the data set that we actually saw, that is the value that makes the realized data set most likely. Let $\hat{a}$ be the maximum likelihood estimate of the parameter a. Then,

$$L_N(\hat{a}) = \underset{a}{Maximum} L_N(a).$$

In practice it is usually more convenient to maximize the logarithm of the likelihood function

$$\omega_N(a) = logL_N(a) = \sum_{i=1}^{N} logf(\underline{X}_i; a)$$

since it achieves its maximum at the same value.

As an example of maximum likelihood estimation, suppose

$$f(X;a) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left\{-\tfrac{1}{2}(X-a)^2/\sigma^2\right\}$$

for a single random variable X and we wish to estimate the parameter a from a sample of size N. The logarithm of the likelihood function is

$$\omega_N(a) = \sum_{i=1}^{N} logf(X_i; a) = -\frac{1}{2\sigma^2}\sum_{i=1}^{N}(X_i - a)^2 - Nlog(\sqrt{2\pi}\sigma).$$

Taking the first derivative with respect to a and setting it equal to zero, yields the solution

$$Y_{ML}^{(a)} = \frac{1}{N}\sum_{i=1}^{N} X_i = \bar{X}$$

which is the sample mean. Thus, the sample mean is the maximum likelihood estimate for the center of a Gaussian distribution.

If we want the maximum likelihood estimate for σ, the standard deviation of the Gaussian, we set the first derivative of ω_N with respect to σ equal to zero gives the solution

$$Y_{ML}^{(\sigma)} = [\frac{1}{N}\sum_{i=1}^{N}(X_i - a)^2]^{1/2}$$

which depends on the value for a. However, we know that the likelihood solution for a, $\hat{a}$, is the sample mean $\bar{X}$ independent of σ, so making this substitution we

have

$$Y_{ML}^{(\sigma)} = [\frac{1}{N}\sum_{i=1}^{N}(X_i - \bar{X})^2]^{\frac{1}{2}},$$

which is just the sample standard deviation.

In many (classic) cases it is possible to calculate in closed form the maximum likelihood estimate as was done for the simple case above. More often this is not possible and it is necessary to explicitly maximize the log-likelihood using numerical optimization techniques in order to obtain the maximum likelihood solution.

There is good reason for statisticians to like maximum likelihood estimation. First it always provides a prescription for parametric estimation. As long as one can compute the joint probability density given a set of parameter values, the likelihood function can be formed and maximized—either algebraically or numerically. The maximum likelihood estimate (MLE) can be shown to always be consistent. As the sample becomes large ($N \rightarrow \infty$), the MLE can be shown to have the highest possible efficiency. Also as the sample size becomes large, the distribution of the MLE estimate $\hat{a}$ can be shown to have a Gaussian distribution about the true value a

$$p_N(\hat{a}, a) = \frac{1}{\sqrt{2\pi}\sigma_{\hat{a}}} \exp\left\{-\tfrac{1}{2}(\hat{a} - a)^2/\sigma_{\hat{a}}^2\right\}$$

with

$$\sigma_{\hat{a}}^2 = 1/[\frac{\delta^2 \omega_N(a)}{\delta a^2}]_{a=\hat{a}}.$$

This information can be used to assign standard errors to maximum likelihood estimates.

There are a few drawbacks to maximum likelihood estimation. The estimates tend to be very non-robust. Also, if numerical optimization is used to obtain the MLE, it can be computationally expensive.

3. Nonparametric Probability Density Estimation

In nonparametric estimation we assume that the data is a random sample from some (joint) probability density, but we do not assume a particular parameterized functional form. This is usually because—for the situation at hand—the correct functional form is simply unknown. The idea is to try to directly estimate the probability density of the population directly from the data in the absence of a specific parameterization. Such estimates are generally used for exploratory data analysis purposes.

Nonparametric density estimation is well developed only for the univariate case. Here we have a set of measurements $\{X_i\}_{i=1}^N$ presumed to be a random sample from some probability density function $p(X)$. Figure 1 illustrates a possible realized configuration of data on the real line. Consider an interval centered at a point X of width ΔX. The probability that one of our data points would have a value in this interval (before we actually observed it) is just the probability content of the interval

$$Prob\{X - \Delta X/2 \leq X_i \leq X + \Delta X/2\}$$

$$= \int_{X-\Delta X/2}^{X+\Delta X/2} p(\zeta) d\zeta$$

$$\simeq p(X)\Delta x.$$

The latter approximation assumes that the width of the interval is small. A reasonable estimate for the probability content of the interval, based on the data, is simply the fraction of data that lies in the interval. (This is in fact the MLE of this quantity), ie.

$$\hat{est}[Prob\{\cdot\}] = \frac{1}{N} Num\{\cdot\}.$$

Combining these results yields an estimate for the probability density at X

$$\hat{p}_N(X) = \frac{1}{(\Delta X)N} Num\{X - \Delta/2 \leq X_i \leq X + \Delta/2\} \tag{7}$$

in terms of the number of counts in the interval of width ΔX centered at X. This result is central to two of the most popular methods of nonparametric density estimation—histograms and window estimates.

For the histogram density estimate the range of the data is divided into M bins or intervals (usually of equal width) and the density is estimated as a (different) constant within each bin using (7) (see Figure 2). The window or square kernel density estimate uses overlapping windows. At each point X for which a density estimate is required, a symmetric interval (window) centered at X of width ΔX is constructed and (7) is used to compute the density estimate (see Figure 3). The windows associated with close points will necessarily have a great deal of overlap.

For both these methods, there is an associated parameter that controls the degree of averaging that takes place. For the histogram estimate it is the number of bins, M. The larger this number, the less smooth the density estimate will

become, but the better able it will be to capture narrow effects (sharp peaks) in the density. For the window estimate this trade-off is controlled by the width ΔX chosen for the window. The smaller the value of ΔX, the rougher the estimate will be, with the corresponding increase in sensitivity to narrow structure.

For multivariate $n > 1$ data, nonparametric density estimation becomes difficult. For two dimensions ($n = 2$) the straightforward generalizations of the histogram and window estimates involving rectangular bins or windows tend to have satisfactory performance. However, for higher dimensions ($n > 2$) performance degrades severely. This is due to the so-called "curse-of-dimensionality."

Consider a histogram density estimate in ten dimensions ($n = 10$). If we choose to have ten bins on each of the ten variables then there would be a total of 10^{10} bins. Clearly for any data set of reasonable size nearly all of these bins would be empty and the few that were not empty would generally contain only one count. Even with only two bins per variable (a very coarse binning) there would be over 1000 bins.

The window estimate suffers similarly. If for a uniform distribution in a ten dimensional unit cube, we wish our window (centered at each data point) to contain ten percent of the data points on the average, the edge length of the window would have to be approximately 0.8; that is, it would have to be 80% of the extent of the data on each variable. Clearly with such a window it would be impossible to detect all but the very coarsest structure of the probability density with such an estimate. Therefore, the most we can hope for is to be able to get a general idea of the joint probability density $p(X_1, X_2, \ldots, X_n)$ in high ($n > 2$) dimensional situations.

Cluster analysis is one approach for doing this. Here the goal is to try to

determine if the joint density is very small nearly everywhere, except for a small number of isolated regions where it is large. This effect is known as clustering. Clustering algorithms attempt to determine when this condition exists and to identify the isolated regions.

Mapping the data to lower dimensional subspaces (usually one or two dimensional subspaces) and studying density estimates on the subspace is often a quite fruitful approach. Good nonparametric density estimation is possible in one and two dimensions. The trick is to perform the mapping in a way that preserves as much as possible the information contained in the full dimensional data set.

Let $\underline{X} = (X_1, X_2, \ldots, X_n)$ be a point in n-dimensions and $t = T(\underline{X})$ represent its mapping to one dimension. Here T is a single valued function of the n arguments $X_1, X_2, \ldots, X_n$. Since $\underline{X}$ is a (vector valued) random variable, t is also a random variable with a corresponding probability density function $p_T(t)$, that depends on the transformation function T. This (one-dimensional) probability density can be easily estimated and examined for different choices of transformations.

For a mapping onto two dimensions, one defines two transformation functions $t_1 = T_1(\underline{X}), t_2 = T_2(\underline{X})$ creating the random variables t_1, t_2 with joint distribution $p_{T_1,T_2}(t_1, t_2)$, depending on the choice of the transformation functions. Again, it is straightforward to estimate and examine the two-dimensional joint density of the mapped points t_1 and t_2. By performing judiciously chosen dimension reducing transformations and studying the corresponding density estimates, one can often gain considerable insight concerning the n-dimensional joint probability density $p(X_1, X_2, \ldots, X_n)$.

Generally the choice of mapping functions is guided by the intuition of the

researcher using his knowledge of the data and the mechanisms that give use to it. There are also techniques that attempt to use the data itself to suggest revealing mappings to lower dimensions. The useful techniques so far developed involve only linear mapping functions

$$t = \sum_{j=1}^{N} a_j X_j = \underline{a}^T \underline{X} \qquad (\text{one} - \text{dimension})$$

$$t_1 = \underline{a_1}^T \underline{X}, t_2 = \underline{a_2}^T \underline{X} \qquad (\text{two} - \text{dimensions})$$

where the projection vectors $\underline{a}, \underline{a_1}, \underline{a_2}$ depend upon the data.

The most commonly used data driven mapping technique is based on principal components analysis. Here the basic notion is that projections (linear mappings) that most spread out the data are likely to be the most interesting. This concept is illustrated in Figure 4 for the case of mapping two-dimensional data to a one-dimensional subspace. Here there are two symmetrically shaped clusters separated in one direction. This direction is the one in which the (projected) data are most spread out, and is also the direction that reveals the existence of the clustering.

Principal components mapping can be fooled, however, as illustrated in Figure 5. Here the clusters are not symmetrically shaped, but are highly elliptical. The separation of the clusters is along the minor axes in the direction for which the pooled data is least spread out. Principal components in this case would choose the direction along the major axes (direction of most data spread) which in this case does not reveal the clustering.

This shortcoming of principal components mapping has lead to the development of projection pursuit mapping. Here, instead of finding mappings (projec-

tions) that maximize the spread of the data, one tries to find those mappings that maximize the information (negative entropy) defined as

$$I(\underline{a}) = -\int p_T(t)\ log\ p_T(t)dt$$

with $t = \underline{a}^T \underline{X}$, and $p_T(t)$ the probability density function of the projected data. This approach successfully overcomes the limitations of the projection pursuit approach but at the expense of additional computation.

4. Conclusion

The purpose of this report has been to give a broad (but necessarily quite shallow) overview of statistical data analysis. The intent was to introduce astronomers to the way statisticians view data so that they can judge whether increased familiarity with statistical concepts and methods will be helpful to them.

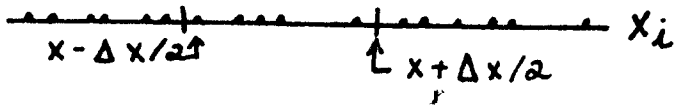

FIGURE 1.

Histogram density estimate:

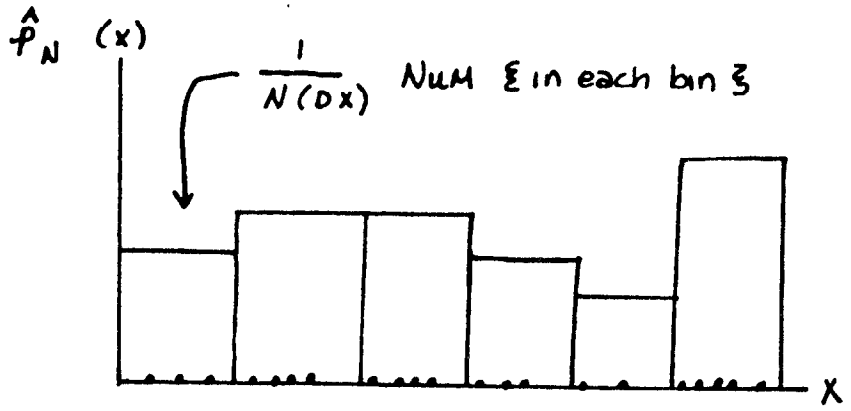

FIGURE 2.

Rosenblatt (square mernal) estimate:

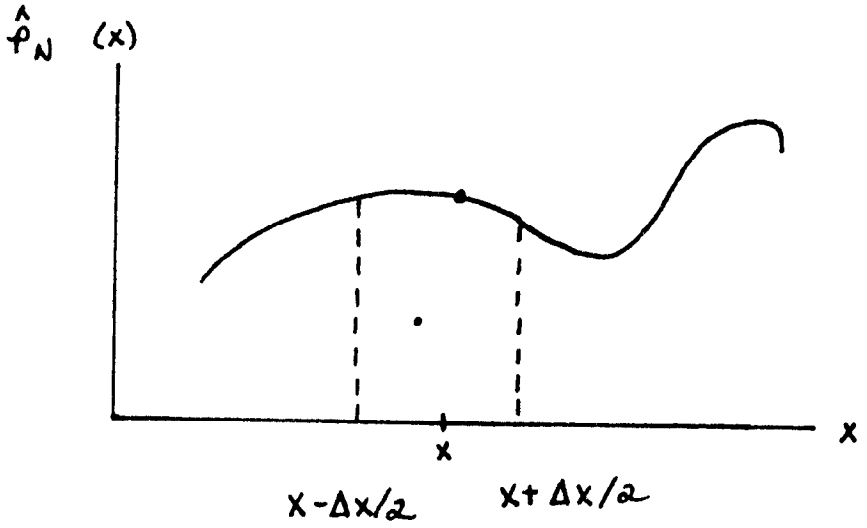

FIGURE 3. Overlapping Bins

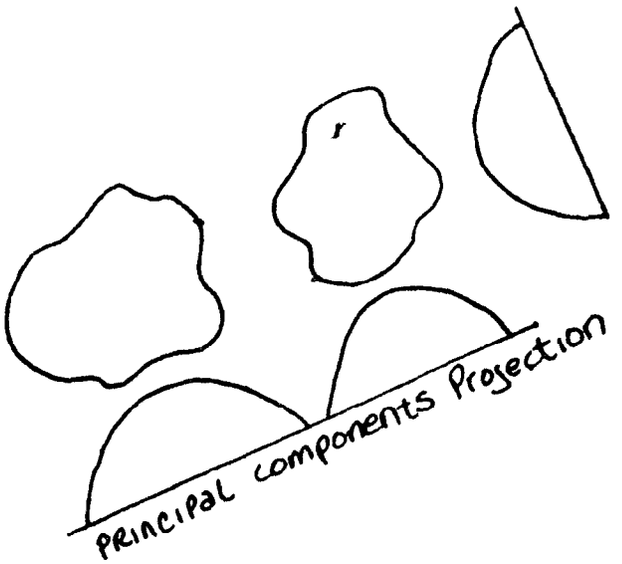

FIGURE 4.

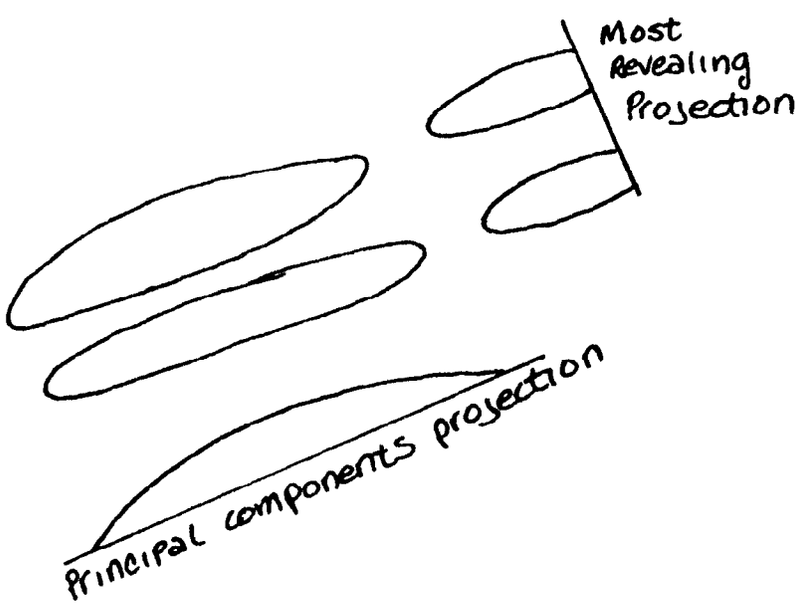

FIGURE 5.