SLAC-176
UC-34d
(E/I)

DATA ANALYSIS TECHNIQUES FOR HIGH ENERGY PARTICLE PHYSICS*

JEROME H. FRIEDMAN
STANFORD LINEAR ACCELERATOR CENTER
STANFORD UNIVERSITY
Stanford, California 94305

PREPARED FOR THE U. S. ATOMIC ENERGY COMMISSION UNDER CONTRACT NO. AT(04-3)-515

September 1974

Printed in the United States of America. Available from National Technical Information Service, U. S. Department of Commerce, 5285 Port Royal Road, Springfield, Virginia 22151.
Price: Printed Copy $5.45; Microfiche $1.45.

*Lectures presented at the CERN School of Computing, Godoysund, Norway, August 11-24, 1974.

ABSTRACT

Useful techniques for the statistical analysis and presentation of high energy particle physics data are described and discussed.

DATA ANALYSIS TECHNIQUES FOR
HIGH ENERGY PHYSICS

TABLE OF CONTENTS

		Page
1.	INTRODUCTION	1
2.	COUNTED DATA AND DENSITY ESTIMATION	2
3.	A MINI-INTRODUCTION TO ESTIMATION THEORY	4
	3.1 Consistency	6
	3.2 Efficiency	6
	3.3 Bias	7
	3.4 Robustness	8
4.	ANALYSIS AND REPRESENTATION OF ONE-DIMENSIONAL DATA	9
	4.1 Non-Parametric Univariate Density Estimation	9
	4.1.1 The Histogram Approach	9
	4.1.2 The Orthogonal Function Approach	11
	4.1.3 The Rosenblatt Estimator	13
	4.1.4 Parzen Estimators	15
	4.1.5 k-th Nearest Neighbor Density Estimation	16
	4.1.6 Discussion	19
	4.2 Smoothing Counted Data	22
	4.3 Parametric Estimation	28
5.	A MINI-INTRODUCTION TO HYPOTHESIS TESTING	33
	5.1 Goodness-of-fit Testing	38
	5.2 Visual Representation of Goodness-of-fit	43
6.	MULTIVARIATE DATA ANALYSIS	47
	6.1 Mapping	54
	6.1.1 Principal Components (Linear Factor Analysis)	54
	6.1.2 Projection Pursuit	56
	6.1.3 Nonlinear Mapping Algorithms	58
	6.2 A Semi-parametric Technique for Model Fitting (Prism Plot Analysis)	60
	6.3 Generalized Nonparametric Multivariate Techniques	64
	6.3.1 A Nonparametric Procedure for Comparing Multivariate Point Sets	64
	6.3.2 Multivariate Tests for Independence	70

Page

6.3.2.1 The Mutual Information Measure for Pairwise Dependence 72

6.3.2.2 An Algorithm for the Direct Measure of Stochastic Independence 73

6.4 A Multivariate Goodness-of-fit Test 77

FOOTNOTES AND REFERENCES 83

1. INTRODUCTION

The purpose of this report is to acquaint high energy physicists with a variety of techniques for presenting and making statistical inferences from counted data. The attempt will be to introduce new techniques that are not commonly used in high energy particle physics as well as to place those methods that are familiar into the general framework of statistical data analysis. This report will not deal with the equally important problem of data reduction. That is, reducing the raw digitizations from particle detectors to more useful quantities such as particle momenta and angles. Although these calculations are often quite complex they seldom require statistical inference. (A notable exception is hypothesis discrimination in kinematic fitting.) The computer codes that perform these computations can usually be thought of as computing engines that transform the data from the raw experimental variables to those that are more convenient for further calculations.

This report is concerned with these further calculations; that is, how to discover properties of the particle interactions from the data, and deduce as well as present, statistically meaningful statements about those properties.

The methods discussed are general in the sense that they can be applied to data from any science that have similar properties to those encountered in particle physics. In fact, many of the techniques that are discussed, although new to particle physics, are commonly used in other sciences, especially pattern recognition and artificial intelligence. The emphasis, however, will be on those methods that can be most profitably applied to the types of data usually encountered in high energy particle physics experiments.

In an effort to keep this report as self-contained as possible, the first sections will deal briefly and very superficially with those concepts from probability theory and statistics that are necessary for understanding what follows.[1] The next sections will discuss various ways of analyzing and presenting univariate or one-dimensional data, that is, when only one measured aspect of the data is considered at a time. Quite often in particle physics experiments several aspects of an event are measured, and the problem is to try to understand and describe the interrelations among these quantities. This requires multivariate (or multi-dimensional) data analysis techniques where several aspects of the data can be simultaneously considered. The final sections describe some new techniques for multi-dimensional data analysis.

The emphasis throughout is on ideas and concepts rather than on specific details. When possible the procedures will be described and discussed in terms of their effect on actual or simulated data rather than with detailed analyses of their statistical properties. When needed, these properties will simply be stated and references provided where interested readers can find detailed analyses and proofs.

2. COUNTED DATA AND DENSITY ESTIMATION

Nearly all analysis on counted data centers on probability density estimation. The several measurements, $\vec{x}$, made on the events are regarded as random variables drawn from (distributed according to) a probability density function, $p(\vec{x})$. If the variables can take on only discreet (rather than a continum of) values then $p(\vec{x})$ is referred to simply as a probability distribution. There are several definitions of probability and probability density but the most intuitive is the frequency ratio definition

$$\lim_{n_i,\, N \to \infty} \frac{n_i}{N} = \int_{r_i} p(\vec{x}) d\vec{x} \quad . \tag{1}$$

Here n_i is the number of counts appearing in a sub-volume, r_i, (cell) of the measurement variables and N is the total number of counts recorded. Constructing r_i as a little sphere about some point $\vec{x}$ and letting the volume approach zero as n_i and N approach infinity, one can define the notion of the probability density, $p(\vec{x})$, at $\vec{x}$. Obvious properties of the probability density are

$$\int_R p(\vec{x}) d\vec{x} = 1 \tag{2}$$

and $p(\vec{x}) \geq 0$ for all $\vec{x} \epsilon R$

where R is the total region of measurement space.

It is clear that $p(\vec{x})$ contains all of the information of the experiment. The purpose of experimentation is to infer properties of $p(\vec{x})$ from the observed distributions of the measured counts. Conversely, it is the purpose of theory to calculate $p(\vec{x})$ from mathematical models and infer from it the results of experiments.

Data analysis is divided into two types, parametric and non-parametric. In parametric (or model dependent) analysis, $p(\vec{x})$ is assumed to be a member of a parameterized family of distributions

$$p(\vec{x}) \equiv p(\vec{a}; \vec{x}) , \tag{2a}$$

where $\vec{a}$ is the set of parameters (either discreet or continuous or both) that specify the particular distribution from the family of possible distributions. The problem of determination of the probability density function then becomes the problem of determining the appropriate values for the parameters $\vec{a}$. The particular parameterized family can come from the researchers intuition, invariance principles (such as angular momentum conservation) or specific dynamical models. For example, the Lorentz invariant amplitude squared for a reaction is the probability density in the Lorentz invariant phase space.

In non-parametric (model independent) analysis no a priori information is assumed about the probability density function. In this case one infers the probability density function directly from the counted data, with very little or no information about what form it might take. Histogramming is an example of a non-parametric (one-dimensional) density estimation.

There are relative advantages and disadvantages to both types of analysis. When it is properly applied parametric analysis is usually statistically much more powerful than non-parametric analysis. This is due to the tremendous increase of information in restricting the set of all possible probability densities to those of a particular parameterized family. The results of the analysis, however, crucially depend upon the correctness of this assumption. If the probability density function that gives rise to the data is not a member of the supposed parameterized family, then at best the statistical power is reduced compared to non-parametric techniques, and at worst (usually the case) the results are meaningless. Non-parametric techniques have the advantage of being applicable to a wide range of problems since they require few assumptions concerning

the data. It should be kept in mind, however, that even though non-parametric techniques are usually formulated independently of specific probability densities their statistical performance usually varies with the actual probability density of the data.

Statistical theory is far more developed for parametric analysis than non-parametric. This is especially true for the family of normal or Gaussian distributions

$$p(\vec{\mu}, \underset{\sim}{\Sigma}, ; \vec{x}) = \frac{1}{(2\pi)^{d/2} |\underset{\sim}{\Sigma}|^{1/2}} \exp[-1/2(\vec{x}-\vec{\mu})^T \underset{\sim}{\Sigma}^{-1}(\vec{x}-\vec{\mu})] \qquad (3)$$

where the parameters are the location vector $\vec{\mu}$ and covariance matrix $\underset{\sim}{\Sigma}$. A great many of the statistical techniques in common use were designed to be optimal for normal distributions and are referred to as normal theory techniques. These techniques can lose considerable statistical power when applied to data with non-normal density distributions.

3. A MINI-INTRODUCTION TO ESTIMATION THEORY

This section introduces the few necessary concepts in Statistics that are required to understand the sections that follow. As noted above, the set of measurements $\{\vec{x}_i\}_{i=1}^N$ comprising an experiment can be thought of as random variables drawn from a probability density function $p(\vec{x})$. The purpose of data analysis is to make inferences concerning $p(\vec{x})$. In parametric analysis one usually wishes to infer likely possible values for the parameters. In non-parametric analysis the density itself is to be inferred. This process of statistical inference is called estimation. Particle physicists quite often (incorrectly) use the terms "measurement" or "determination" for statistical estimation.

Consider a parametric example. Suppose that the set of measurements $\{\vec{x}_i\}_{i=1}^N$ are known to be distributed according to $p(a;\vec{x})$ for some (unknown) value of a. The desire is to estimate the parameter, a, from the values of the measured random variables $\vec{x}_1, \vec{x}_2, \ldots \vec{x}_N$.

Any function of a set of random variables

$$Y = \phi(\vec{x}_1, \vec{x}_2, \ldots \vec{x}_N) \qquad (4)$$

is itself a random variable with a probability density function $p_N(a;Y)$ that can (at least in principle) be calculated from $p(a;\vec{x})$. If one is sufficiently clever in choosing the function, ϕ, then $p_N(a,Y)$ might be large only for those values of Y near Y = a. That is, for any set of possible values for $\vec{x}_1, \vec{x}_2, \ldots \vec{x}_N$ drawn

from $p(a,\vec{x})$, $\varphi(\vec{x}_1,\vec{x}_2,\ldots\vec{x}_N)$ always has a value near a. Thus, for the particular set of values of $\vec{x}_1,\vec{x}_2,\ldots\vec{x}_N$ we happen to measure, the value of $Y = \varphi(\vec{x}_1,\vec{x}_2,\ldots\vec{x}_N)$ will be a good approximation to the value of a. The function $\varphi(\vec{x}_1,\vec{x}_2,\ldots\vec{x}_N)$ is called a statistic and its value for a particular set of $\vec{x}_1,\vec{x}_2,\ldots\vec{x}_N$ is called an estimate of a.

Consider the following example of how one might construct a statistic for performing an estimation. The integral

$$I(a) = \int_R f(\vec{x})\,p(a;\vec{x})\,d\vec{x} \tag{5}$$

is a function of the parameter, a. Here R is the region of all possible values for the measurements $\vec{x}$. This integral is called the expected value, E[f], (or sometimes the average or mean value) of the arbitrary function, $f(\vec{x})$, with respect to $\vec{x}$. If the integrand is integrable then the explicit functional form of I(a) can be calculated. From the central limit theorem (law of large numbers), one has the result that for sufficiently large N,

$$Y = \varphi(\vec{x}_1,\vec{x}_2,\ldots\vec{x}_N) \equiv \frac{1}{N}\sum_{i=1}^{N} f(\vec{x}_i) \tag{6}$$

has a normal probability density function

$$p_N(a;Y) = \frac{1}{\sqrt{2\pi}\,\sigma_N}\, e^{-1/2\left[(Y-a)/\sigma_N\right]^2} \tag{7}$$

where

$$\sigma_N = \left(\frac{1}{N}\int_R (f(\vec{x}) - E[f])^2\, p(a;\vec{x})\,d\vec{x}\right)^{1/2} \tag{8}$$

The integral is called the variance, V[f], of the function $f(\vec{x})$. Thus,

$$\sigma_N = \sqrt{V[f]/N}\ . \tag{9}$$

As N increases, σ_N decreases as $\sigma_N \sim 1/\sqrt{N}$. Thus, for sufficiently large N, $p_N(a;Y)$ will be a very narrow distribution centered at $Y = a$. The quantity $(1/N)\sum_{i=1}^{N} f(\vec{x}_i)$ is called the sample mean of the function $f(\vec{x})$. The central limit theorem tells us that the sample mean is a good statistic for estimating I(a) [expected value of f(x)] for sufficiently large N. One can then take as an

estimate, $\hat{a}$, for the value of the parameter, a,

$$\hat{a} = I^{-1}\left[\varphi(\vec{x}_1,\vec{x}_2,\ldots\vec{x}_N)\right] = I^{-1}\left[\frac{1}{N}\sum_{i=1}^{N} f(\vec{x}_i)\right]. \qquad (10)$$

Since the function, $f(\vec{x})$, was somewhat arbitrary it is clear that this procedure can be used to construct a variety of statistics for estimating the parameter, a. However, some will be better than others. For example σ_N, which regulates the precision with which the parameter, a, is estimated, depends on $f(\vec{x})$ (for $N < \infty$) through Eq. 8.

The field of Statistics is concerned to a great degree with finding good statistics for estimation and determining their properties. Statistics used for estimation (usually called estimators) are rated in terms of four basic properties of their probability density distributions $p_N(a;Y)$; these are consistency, efficiency, bias, and robustness.

3.1 Consistency

An estimator, $Y = \varphi(\vec{x}_1,\vec{x}_2,\ldots,\vec{x}_N)$ is consistent if the following condition holds

$$\lim_{N\to\infty} p_N(a;Y) = \delta(Y-a). \qquad (11)$$

That is as the number of samples gets arbitrarily large, p(a;Y) becomes an arbitrarily narrow function of Y about a, and the estimator provides an arbitrarily precise estimate of the parameter, a. Note that Eqs. 7 and 8 show that the estimator defined by Eq. 6 is consistent. Consistency is nearly always required for an estimator to be considered useful.

3.2 Efficiency

Consistency is concerned with the precision of the estimator for infinite sample size. (In the field of Statistics, a result that holds in the limit of infinite sample sizes is called an asymptotic result.) Efficiency is concerned with the precision of the estimator for finite sample size N. An estimator is called efficient if the variance (mean squared error) of its probability density function

$$V_N = \int_R (Y-a)^2\, p_N(a;Y)\,dY \qquad (12)$$

is as small as possible[3] for a given N. The square root of the variance, $\sigma_N = \sqrt{V_N}$, is characteristic of the width of $p_N(a;Y)$ about a, and thus is directly related to the precision of the estimator. Therefore, an efficient estimator for

a given N is one that (loosely speaking) has maximal precision. The relative efficiency between two estimators is the inverse ratio of the variances of their probability densities for a given sample size, N. The efficiency of an estimator is its relative efficiency to an efficient estimator (i.e., efficient estimators are said to have 100% efficiency).

This definition of efficiency can be related to the intuitive meaning of the word in the following manner. For large sample size, N, the variance of most estimators decreases as, $V_N \sim 1/N$, for increasing N (i.e., $\sigma_N \sim 1/\sqrt{N}$). Then the efficiency of an estimator is the inverse ratio of the number of samples (events) it requires to the number an efficient estimator requires for the same precision. Clearly high efficiency is a desirable property for an estimator. However, an estimator with the highest efficiency is quite often not the most desirable. Sometimes the computational complexity of the most efficient estimator makes it more expensive for a given precision than a less efficient estimator even though the less efficient estimator requires more events.

3.3 Bias

Like efficiency, bias refers to a property of estimators for finite sample size, N. Specifically the bias of an estimator is defined as

$$b_N = \int_R Y\, p_N(a;Y)\, dY - a \qquad (13a)$$

i.e.,

$$b_N = E_N[Y] - a \quad . \qquad (13b)$$

A biased estimator is one with an expected value that is different from the true value of the parameter being estimated. The bias is just the difference between $E_N[Y]$ and the true value of the parameter.

Note that, although it might appear to be contradictory, a biased estimator can also be consistent and conversely an unbiased estimator can be inconsistent. If a biased estimator is consistent, then from Eq. 11

$$\lim_{N \to \infty} b_N = 0 \, .$$

It may at first seem that bias would be a very undesirable property for an estimator to have. This is generally not the case. It is only important that the bias be relatively small compared to the square root of the variance (Eq. 12) (standard deviation) of the probability density function. Most of the commonly used estimators in particle physics are in fact biased. There are various techniques for reducing bias in estimators but they usually do this at the expense

of precision (increasing the variance). Thus, one is forced to compromise between degree of bias and precision.

3.4 Robustness

The concept of robustness is strongly tied to the notions of non-parametric data analysis. As pointed out above parametric techniques are usually more efficient than non-parametric techniques so long as the actual probability density of the data is in fact a member of the supposed parameterized family of density distributions. If this assumption is not quite correct then the parametric estimator loses efficiency. However, some estimators lose efficiency more rapidly than others as nature deviates from the experimenter's assumptions. Estimators that maintain reasonable efficiency over a wide range of data probability densities are called robust estimators.

Robustness is an often under-rated quality of estimators. Both physicists and statisticians usually seek maximum efficiency at the expense of robustness. This effort can often be misguided since if the data deviate from the a priori assumptions then the supposed most efficient estimator can, in fact, have poor efficiency. Even when the theory from which the parameterized density function is constructed is on solid ground, measurement errors on the data points can cause the data to deviate from the parameterized family of densities.

Most highly efficient parametric estimators gain most of their information from the low density regions (tails) of the distribution. Thus, if only a few data points in these tails deviate from the parameterized probability density function the estimate will be severely effected. Physicists usually refer to this phenomenon as the "tail wagging the dog". What has actually happened is that their estimator was very non-robust. The least squares estimator, which is one of the most popular in particle physics, is a good example of an extremely non-robust estimator. Generally, order statistics such as medians, and percentiles are much more robust than arithmetic statistics such as means and standard deviations.

For example if one wished to estimate the width of a symmetric distribution he could calculate its standard deviation about the mean or he could take half the difference between its 32 and 68 percentiles. The former would be more efficient if the data had exactly a normal distribution. However, if just one of the data points near the edges of the distribution was mismeasured so that it was somewhat farther from the center than it should be, the standard deviation estimate will be severely effected. This is because the standard deviation estimate

weights each point by the square of its distance from the center. The percentile estimate on the other hand will be completely unaffected by the mismeasured point.

For exploratory data analysis especially, robustness is essential. Robust estimators generally maintain from 60% to 90% efficiencies over wide ranges of data distributions while non-robust estimators tend to have near 100% efficiency when the data distribution exactly follows the predicted probability density function, and low efficiency when it does not.

4. ANALYSIS AND REPRESENTATION OF ONE-DIMENSIONAL DATA

With the preliminaries of the preceeding section out of the way, we are ready to discuss and evaluate various techniques for analyzing and presenting data. We will start with univariate or one-dimensional data analysis. That is when only one measured quantity is considered at a time. We will discuss multivariate analysis in the following sections. Univariate analysis techniques are far more developed than corresponding multivariate techniques. This is especially true for non-parametric methods. There are many large text books devoted to statistical techniques for univariate analysis. Thus, there will be no attempt in this brief report for completeness. The purpose will be to introduce some techniques not commonly known to high energy particle physicists that could be valuable tools for analyzing particle physics data, and to relate them to the more commonly used techniques.

4.1 Non-Parametric Univariate Density Estimation

Let $\{x_i\}_{i=1}^{N}$ be a sequence of independent identically distributed random variables with some unknown probability density function $p(x)$. We wish to construct estimators $\hat{p}(x) = \varphi_N(x_1, x_2, \ldots x_N)$ for $p(x)$ that depend only on the observations, $\{x_i\}_{i=1}^{N}$.

4.1.1 The Histogram Approach

Histogramming is the most commonly used method in particle physics. In this method the real line is divided into M regions, r_i, (bins, channels) and $\hat{p}(x)$ is taken to be constant over each region r_i:

$$\hat{p}(x) = \hat{p}_i \qquad \text{if } x \in r_i, \qquad i = 1, M .$$

Let $g_i(x)$ be an indicator function for each region, i.e.,

$$g_i(x) = \begin{cases} 1 \text{ if } x \in r_i \\ 0 \text{ otherwise.} \end{cases}$$

Then we have for our estimator of p(x),

$$\hat{p}_N(x) = \frac{1}{N} \sum_{i=1}^{M} \sum_{j=1}^{N} g_i(x_j) g_i(x) \quad . \tag{14}$$

From the central limit theorem one has

$$E[\hat{p}_i] = \bar{p}_i \equiv \int_{r_i} p(x)\, dx \tag{15}$$

and

$$p_N(\hat{p}_i) = \frac{1}{\sqrt{2\pi}\,\sigma_i} e^{-\frac{(\hat{p}_i - \bar{p}_i)^2}{2\sigma_i^2}} \tag{16}$$

where $\sigma_i = \sigma_0/\sqrt{N}$,

when $\bar{n}_i \equiv N\bar{p}_i$ is large. A more careful analysis shows that for any $\bar{n}_i$, the $\hat{n}_i = N\hat{p}_i$ are distributed according to a multinomial distribution

$$p_N(\hat{n}_1, \hat{n}_2, \ldots \hat{n}_M) = N! \prod_{i=1}^{M} \frac{\bar{p}_i^{\hat{n}_i}}{\hat{n}_i!} \tag{17}$$

if the total number of events, N, is considered fixed.[4] Note from Eq. 17

$$E[\hat{n}_i] = \bar{n}_i \tag{18}$$

so that the estimator is unbiased. The variance of $\hat{n}_i$ is

$$V[\hat{n}_i] = N\bar{p}_i(1-\bar{p}_i) \tag{19a}$$

so that

$$V[\hat{p}_i] = \bar{p}_i(1-\bar{p}_i)/N \quad . \tag{19b}$$

Equation 19 shows that $\hat{p}_i$ is a consistent estimator of $\bar{p}_i$. For $\bar{p}_i \ll 1$ (large number of bins for example) Eq. 19a can be approximated by

$$V[\hat{n}_i] \simeq N\bar{p}_i \equiv \bar{n}_i \quad . \tag{20}$$

Since $\bar{n}_i$ is usually not known it seems reasonable to make the further approximation

$$\bar{n}_i \simeq \hat{n}_i \tag{21}$$

in Eq. 20, in which case

$$V[\hat{n}_i] \approx \hat{n}_i$$

or

$$\sigma[\hat{n}_i] \equiv \sqrt{V[\hat{n}_i]} = \sqrt{\hat{n}_i} \quad . \tag{22}$$

These approximations, then, yield the common "rule of thumb" result that the statistical uncertainty in the number of counts in a histogram bin is equal to the square root of the number of counts.

It can easily be shown that Eq. 17 can be approximated by Eq. 16 for large $\bar{n}_i$, with σ_i^2 given by Eq. 19b.

Although $\hat{p}_i$ is a consistent estimator of $\bar{p}_i$, $\hat{p}_N(x)$ from Eq. 14 is <u>not</u> a consistent estimator of $p(x)$ (unless by some chance $p(x)$ is exactly piece-wise constant over the chosen bins). As the total number of counts tends to infinity, the average variance (mean squared error) of the density estimation

$$V_N = E\left\{\int_R [p(x) - \hat{p}_N(x)]^2 \, dx\right\} \tag{23}$$

does <u>not</u> approach zero.

There are other shortcomings of the histogram approach. The choice of the binned intervals, r_i, and their number, M, is arbitrary. There are no general guidelines for optimum binning except looking at the result and rebinning. Also, if constant bin sizes are used many bins may have too few counts while only a few of the others may contain nearly all of the counts rendering useful density estimation impossible. Histogramming also fails to use to advantage any continuity properties of $p(x)$. Since the estimates in neighboring bins are independent of each other there will be sharp discontinuities in $\hat{p}_N(x)$, Eq. 14, at the bin boundaries.

These discontinuities (usually termed "statistical fluxations" by physicists) are normally the result of the variance of the estimation in each separate bin and do not represent actual structure in $p(x)$. Most probability densities $p(x)$, are reasonably continuous, and using this information can considerably reduce the variance, V_N, (Eq. 23) of the density estimation.

4.1.2 <u>The Orthogonal Function Approach</u>

After histogramming, the most common density estimators used by high energy physicists are orthogonal functions. Let $\{\psi_i(x)\}_{i=1}^{M}$ be a set of

orthogonal functions defined on the real line

$$\int_R \psi_i(x)\,\psi_j(x)\,dx = \delta_{ij}$$

and we wish to estimate p(x) from the data points $\{x_j\}_{j=1}^{N}$ with an estimator of the form

$$\hat{p}_N(x) = \sum_{i=1}^{M} \hat{c}_i^{(N)}\,\psi_i(x) \quad . \tag{24}$$

If the actual probability density function, p(x), were known then it is easy to show that the variance of the density estimation, V_N, (Eq. 23) is minimal for

$$\hat{c}_i^{*} = \int_R \psi_i(x)\,p(x)\,dx = E[\psi_i] \quad . \tag{25}$$

For non-parametric estimation p(x) is not known so we estimate the integral from the data sample

$$\hat{c}_i^{(N)} = \frac{1}{N} \sum_{j=1}^{N} \psi_i(x_j) \quad . \tag{26}$$

From the central limit theorem one has (for large N)

$$p_N\left[\hat{c}_i^{(N)}\right] = \frac{1}{\sqrt{2\pi}\,\sigma_N^{(i)}}\; e^{-1/2\;\frac{(\hat{c}_i^{(N)} - \hat{c}_i^{*})^2}{(\sigma_N^{(i)})^2}} \tag{27}$$

where

$$\sigma_N^2 = V(\psi)/N = E\left[(\psi - E[\psi])^2\right]/N \quad .$$

Thus, $E[\hat{c}_i^{(N)}] = \hat{c}_i^{*}$ so that the estimate is unbiased and $\lim_{N\to\infty} \sigma_N^2 = 0$ so that it is consistent.

Combining Eqs. 24 and 26 we have for our density estimate

$$\hat{p}_N(x) = \frac{1}{N} \sum_{i=1}^{M} \sum_{j=1}^{N} \psi_i(x_j)\,\psi_i(x) \quad . \tag{28}$$

The average variance of the estimate, V_N, Eq. 23, is

$$V_N = E\left\{\int_R (p - \hat{p}_N)^2\,dx\right\} = E\left\{\int_R (p - \hat{p})^2\,dx\right\} + E\left\{\int_R (\hat{p} - \hat{p}_N)^2\,dx\right\} \tag{29}$$

where

$$\hat{p}(x) = \sum_{i=1}^{M} \hat{c}_i^* \psi_i(x). \tag{30}$$

The first term on the right hand side of Eq. 29 is a constant independent of the data so that

$$V_N = \int_R p^2(x)\,dx - \int_R \hat{p}^2(x)\,dx + E\left[\int_R \left[\hat{p}(x) - \hat{p}_N(x)\right]^2 dx\right]$$

or

$$V_N = \int_R p^2(x)dx - \int_R \hat{p}^2(x)dx + \sum_{i=1}^{M} (\sigma_N^{(i)})^2 \ . \tag{31}$$

Equation 31 shows that the variance of the estimate is composed of a constant systematic part and a statistical part that approaches zero as N becomes infinite. Thus, like the histogramming approach, the orthogonal function density estimator is inconsistent (unless by some chance $p(x) = \hat{p}(x)$ for all x — i.e., either $M = \infty$, or for finite M, p(x) can exactly be expressed by Eq. 30).

It is no accident that the histogramming and orthogonal function estimators share this property of inconsistency. Inspecting Eq. 14, one sees that it is just a special case of Eq. 28 where the orthogonal functions are the indicator functions $g_i(x)$. Note that

$$\int_R g_i(x)\, g_j(x)\,dx = \delta_{ij}$$

and

$$\hat{C}_i^{(N)} = \frac{1}{N} \sum_{j=1}^{N} g_i(x_j) = \frac{\hat{n}_i}{N} \ .$$

The general orthogonal function approach suffers from generalized analogs of most of the problems discussed for histogramming. The problem of specific bin choice and number of bins becomes the problem of number and specific choice of the orthogonal functions, $\{\psi_i(x)\}_{i=1}^{M}$. Also, it may happen that $\hat{p}_N(x)$ is negative for some value of x rendering it inadmissible as a probability density function (although it still may be quite useful).

4.1.3 The Rosenblatt Estimator

We will now begin to consider some consistent estimators of univariate probability density. The first is the Rosenblatt or "naive pdf" (probability density

function) estimator.[5] This estimator is defined as

$$\hat{p}_N(x) = \frac{1}{2h}\left[\hat{P}_N(x+h) - \hat{P}_N(x-h)\right] . \tag{32}$$

Here $\hat{P}_N(x)$ is the empirical cumulative distribution of the data points defined as

$$\hat{P}_N(x) = \begin{cases} 0 & x < x_1 \\ i/N & x_i \leq x < x_{i+1} \\ 1 & x \geq x_N \end{cases} \tag{32a}$$

where the data points are labeled in increasing order of x. From the central limit theorem one has

$$\lim_{N \to \infty} \hat{P}_N(x) = P(x) \equiv \int_{-\infty}^{x} p(x)\,dx . \tag{33}$$

The quantity P(x), defined in Eq. 33, is called the cumulative distribution of p(x). From its definition (Eq. 32), one sees that this estimate, $\hat{p}_N(x)$, is just the fraction of counts that lie in a window of width 2h, centered at x, divided by the window width. If we define an indicator function

$$\psi(x;x') = \begin{cases} 1, \text{ if } x - h \leq x' < x + h \\ 0, \text{ otherwise} \end{cases} \tag{34}$$

then the probability density estimate can be written as

$$\hat{p}_N(x) = \frac{1}{2hN} \sum_{i=1}^{N} \psi(x;x_i) . \tag{35}$$

Rosenblatt shows that for all N (not just for $N \to \infty$)

$$E\left[\hat{p}_N(x)\right] = \frac{1}{2h}\left[P(x+h) - P(x-h)\right] . \tag{36}$$

Expanding this in a Taylor series

$$E\left[\hat{p}_N(x)\right] = p(x) + \frac{h^2}{6}\, p''(x) + 0(h^4) . \tag{37}$$

This result shows that the estimate is biased with the bias approaching zero, quadratically, as the window size h approaches zero.

Rosenblatt also calculates the variance of the estimate as

$$E\left[\left(\hat{p}_N(x) - p(x)\right)^2\right] = \frac{p(x)}{2hN} + \frac{h^4}{36}\,|p''(x)|^2 + 0\left(\frac{1}{hn} + h^4\right) . \tag{38}$$

This estimator can be made consistent so long as h tends to zero, while the product (hN) approaches infinity. Parameterizing the window size as

$$h = CN^{\alpha} \tag{39}$$

and choosing α so as to minimize the dominant terms in Eq. 38, one obtains $\alpha = -1/5$ as the value that causes the variance to decrease most rapidly with increasing N. A careful analysis shows that the constant should be

$$C = \left[9\, p(x)/2\, |p''(x)|^2\right]^{1/5}. \tag{40}$$

The bias of this estimator (Eq. 37) is easy to understand. For finite window size the estimator $\hat{p}_N(x)$ (Eq. 32) is an unbiased estimator of the average of the probability density within the window

$$\bar{p}(x) = \frac{1}{2h} \int_{x-h}^{x+h} p(x')dx' \ . \tag{41}$$

If $p(x')$ is nonlinear within the window region, then this average will be different than the value of the probability density at the center of the window, $p(x)$. As the window size approaches zero, or as the probability density approaches linearity, this effect will disappear, as reflected by Eq. 37.

The expression for the variance (Eq. 38) shows that like histogramming, the variance of this estimate is proportional to the value of the probability density (standard deviation proportional to the square root of the probability density). Unlike histogramming, however, this probability density estimate is not piecewise constant over fixed intervals (bins) and does not suffer from the sharp discontinuities that histogramming produces at the boundaries of these intervals ("statistical fluxuations"). This estimator does, of course, suffer from statistical uncertainty as reflected by its variance (Eq. 38). However, the Rosenblatt estimator produces a relatively smooth probability density estimate which (at least in the limit of large sample size) can be shown to be more accurate than histogramming (see below for finite sample comparisons).

4.1.4 Parzen Estimators

The Rosenblatt estimator is a special case of a general class of density estimators known as Parzen estimators or Parzen windows.[6] Let K(y) be a bounded absolutely integrable function such that

$$\int_R K(y)\, dy = 1 \qquad \text{and} \qquad \lim_{|y| \to \infty} |yK(y)| = 0 . \tag{42}$$

Then the Parzen window estimators are defined as

$$\hat{p}_N(x) = \frac{1}{h(N)} \sum_{i=1}^{N} K\left(\frac{x-x_i}{h(N)}\right) . \tag{43}$$

The function K(y) is called the kernel or window function. The notation h(N) is used to explicitly indicate that the scale parameter for the kernel function depends upon the sample size, N. For the Rosenblatt estimator one has

$$K\left[\frac{x-x_i}{h(N)}\right] = \frac{\psi(x;x_i)}{2N} \tag{44}$$

where $\psi(x;x_i)$ is defined in Eq. 34. Other possible kernels are: a) the double exponential function $e^{-|y|}$; b) the standard normal (Gaussian) function; c) the Cauchy function $1/(1+y^2)$; and d) $\sin^2 y/y^2$. Using procedures analogous to those for the Rosenblatt estimator, one can show that these estimators are biased, with the bias tending to zero quadratically as the scale parameter h(N) approaches zero. Also, the variance of the estimate tends to zero as $1/Nh(N)$ for increasing sample size, N. Thus, these estimators are consistent provided that $h(N) \to 0$ while $Nh(N) \to \infty$.

4.1.5 k-th Nearest Neighbor Density Estimation

A disadvantage of the density estimators so far discussed is that there are few general guidelines for choosing the scale parameter (bin width for histogramming, window size, h(N), for Rosenblatt and Parzen estimators). For small variance (high statistical precision) the scale parameter should be as large as possible. For maximum sensitivity to p(x), rapid convergence as well as minimal bias (high systematic precision), the scale parameter should be as small as possible. The choice for a scale parameter is usually then a compromise between these two competing effects. Ideally, the scale parameter should depend upon the data. That is, on the basis of a density estimate the scale parameter can be changed and the density re-estimated. Although quite reasonable, this procedure invalidates the analyses that give rise to the statistical results stated above concerning the bias, consistency and variance of these estimators, since the analyses all assume that h(N) is a deterministic function independent of the data. Thus, the statistical properties of such a procedure are largely unknown. Even further, the scale parameter should probably change for different values of the variable, x. In denser regions, one can take advantage of the large number of counts to increase systematic precision by using smaller

values for the scale parameter. In the sparser regions the statistical precision is relatively low so that larger values of scale parameter are in order.

The k-th nearest neighbor estimator[7] allows the scale parameter h(N) to adapt to the data. Density is measured as counts per distance interval (univariate volume). For the estimators discussed so far, the interval was predetermined (by the scale parameter) and the probability content was estimated by the fraction of counts that fall in the interval. With the k-th nearest neighbor estimator, the probability content is predetermined and the interval size required to contain the probability is estimated. The estimation statistic is distance instead of number of counts. Specifically, let k(N) be a predetermined integer ($\leq$ N) and let h(N) be the distance from x to the k-th closest data point to x. Thus, h(N) is a random variable depending on the data. The number of counts within an interval of width, 2h(N) centered at x, is k(N) by definition so that the probability density function estimate at x is

$$\hat{p}_N(x) = k(N)/2Nh(N). \tag{45}$$

It is clear that this estimator overcomes many of the disadvantages of the fixed interval estimators discussed above. The interval width, h(N), becomes narrow in regions of high counting density and wider in sparser regions, tending to stabilize the variance of the estimates.

The k-th nearest neighbor estimator is biased by the same mechanism as the Rosenblatt estimator. To second order

$$E[\hat{p}_N(x)] = p(x) + \frac{1}{24}\left[\frac{k(N)}{N}\right]^2 \frac{p''(x)}{p^2(x)} \tag{46}$$

so that (like the Rosenblatt estimator) the bias is proportional to the nonlinearity of the probability density function and approaches zero quadratically as k(N)/N approaches zero. Fukunaga and Hostetler[8] show that the variance for this estimator is

$$V_N(x) = \frac{p^2(x)}{k(N)} + \left[\frac{p''(x)}{24\,p^2(x)}\left(\frac{k(N)}{N}\right)^2\right]^2 \tag{47}$$

(again to second order). From these equations we see that this estimator is consistent provided k(N) is chosen such that

$$\begin{aligned} &\lim_{N\to\infty} k(N) = \infty \\ &\lim_{N\to\infty} \frac{k(N)}{N} = 0 \ . \end{aligned} \tag{48}$$

These conditions were shown by Loftsgaarden and Quesenberry[9] to be required for consistency very early and, in fact, the k-th nearest neighbor estimator is sometimes referred to as the method of Loftsgaarden and Quesenberry.

Minimizing Eq. 47 with respect to k(N), one obtains

$$k_0(N) = \left\{ \frac{144 p^6(x)}{[p''(x)]^2 N} \right\}^{1/5} N^{4/5} \tag{49a}$$

so that

$$\frac{k_0(N)}{N} = \left\{ \frac{144 p^6(x)}{[p''(x)]^2 N} \right\}^{1/5} . \tag{49b}$$

We see that the optimum number of nearest neighbors depends upon the underlying distribution. The smaller $|p''(x)|$, the smoother the density function, and the number of nearest neighbors can be increased for greater statistical precision. For very nonlinear functions, where $|p''(x)|$ is large, the bias dominates the precision of the estimate and a smaller number of neighbors should be chosen to reduce it.

The first term in Eq. 47 is the variance of the estimate about its mean, i.e.,

$$E[\hat{p}_N^2(x)] - E^2[\hat{p}_N(x)] = \frac{p^2(x)}{k(n)} \tag{49}$$

or

$$\sigma[\hat{p}_N(x)] \simeq \hat{p}_N(x)/\sqrt{k(N)}$$

where $p(x)$ is approximated by $\hat{p}_N(x)$. Thus, the statistical uncertainty of this density estimator is proportional to the density rather than the square root of the density, as is the case for the fixed interval estimators discussed above. The coefficient of variation of the statistical uncertainty

$$C = \frac{\sigma[\hat{p}_N(x)]}{\hat{p}_N(x)} = \frac{1}{\sqrt{k(N)}} \tag{50}$$

is constant for the k-th nearest neighbor estimator. Thus, the fractional statistical precision in the estimate of $p(x)$ is uniform for all x, which overcomes one of the difficulties mentioned above for the fixed interval estimators.

4.1.6 Discussion

Several techniques for nonparametric one-dimensional density estimation have been presented in the previous sections and their properties discussed. It was shown that in the limit of very large data samples the Rosenblatt, Parzen, and near neighbor estimators are more accurate than the histogramming and orthogonal function approach. The near neighbor technique was shown to have the additional advantage of adaptability to the data.

In order to gain insight into the relative performance of these estimators for sample sizes and distributions commonly encountered with particle physics data, several Monte Carlo experiments were performed. A random sample of 710 data points was drawn from the probability density function

$$p(x) = \left(\frac{0.5}{710}\right) \frac{1}{\pi} \left[\frac{2}{1+2(x-0.5)^2} + \frac{20}{1+20(x-0.2)^2}\right] \qquad (51)$$

in the interval $0 \leq x \leq 1$. From these data points, an estimate of the probability density function, $\hat{p}_{710}(x)$, was obtained using the histogramming approach, the Rosenblatt estimator, the Parzen estimator with the standard normal density function as a kernel, and the near neighbor technique. These estimates were then compared to the true probability density, $p(x)$, of Eq. 51 by

$$V_{710} = \int_0^1 [p(x) - \hat{p}_{710}(x)]^2 \, dx \; . \qquad (52)$$

This process was repeated nine times with different random sample points drawn from $p(x)$ (Eq. 51), and the expected value $E[V_{710}]$ was estimated as the average V_{710} from these nine trials. Figure 1 (a-d) shows the results. Here the $\hat{p}_{710}(x)$, as estimated by each method, is plotted for the most "typical" data sample of the nine trials. This most "typical" trial was taken to be the one with the four values of V_{710} (from the four methods) that were closest to the averages over the nine trials. These averages were:

Table 1

Method	Parameter	Average V_{710}	V_{710} for "typical" Trial
Histogram	h = .012	.064	.064
Rosenblatt	h = .045	.038	.037
Parzen (K=Normal)	h = .030	.045	.045
Nearest Neighbor	k = 120	.023	.023

For comparison, the true density, $p(x)$, Eq. 51, is superimposed in each figure as a continuous line over the density estimate, $\hat{p}_{710}(x)$.

The parameter values used in each method were those that gave the best results (minimum V_{710}) for the particular method. The histogramming method was most sensitive to particular parameter value while the k-th nearest neighbor technique was least sensitive.

The results shown in Table 1 indicate that, for this example, the consistent estimators do provide more accurate density estimation than histogramming. The variance of the most accurate estimator, the k-th nearest neighbor, is nearly three times less than that for the histogram.

Although only a single example, Figure 1 illustrates the various different properties of these estimators. The bias of the Rosenblatt and Parzen estimators is especially apparent in the center and at the right shoulder of the peak. It is in these two regions where $p(x)$ is most nonlinear. As predicted by Eq. 37, those regions where the second derivative is large and negative (center of peak), $\hat{p}_{710}$ underestimates $p(x)$, whereas when it is large and positive (right shoulder), $\hat{p}_{710}$ overestimates $p(x)$. The k-th nearest neighbor estimator is also biased from this same mechanism. However, as predicted by Eq. 46, the bias will be small in the peaked region since the bias term is proportional to $p''(x)/p^2(x)$ and $p(x)$ is very large in this region. The bias is larger in the right shoulder region where $p(x)$ is not large enough to overcome the affect of $p''(x)$.

The k-th nearest neighbor estimate is seen from Figure 1d to become poor near the boundaries. This is a general property of this estimator. When straightforwardly applied, the near neighbor estimator will always underestimate the density whenever the interval containing the k neighbors is adjacent to a boundary edge. In this case, the actual interval size is $h(N) + B$ where B is the distance from the evaluation point to the boundary. Since this is less than $2h(N)$, which appears in Eq. 45, $\hat{p}_N(x)$ will have a strong negative bias. One could try to remedy this by using the actual interval size, $h(N) + B$, in place of $2h(N)$ whenever the interval contains a boundary edge. This, however, also causes the estimate to be biased with the bias being proportional to, and having the opposite sign of, $p'(x)$ in the interval.

A good boundary strategy (and the one used in Figure 1d) is to revert to a variable interval Rosenblatt estimate. That is, whenever the k-th nearest neighborhood contains a boundary a distance $B < h(N)$ from the evaluation point, x, then the number of points, n_B, in the smaller interval of width 2B centered

at x, is determined, and the density is estimated as

$$\hat{p}_B = \frac{n_B}{2NB} \quad . \tag{53}$$

As seen in Figure 1d, this strategy removes most of the bias. The variance of the estimate, however, becomes relatively large for those points very close to the boundary edge.

The Parzen estimate (Figure 1c) is seen to be the smoothest, while the histogram (Figure 1a) is least smooth. The relative accuracy of the estimates, as reflected from visual inspection of Figure 1, appears to correspond to the relative values of $E[V_{710}]$ given in Table 1. That is, the k-th nearest neighbor method gives the best density estimation, followed by (in order) the Rosenblatt, Parzen and Histogram methods.

Although the consistent estimators (especially the nearest neighbor) are generally more accurate than the histogramming approach, they are also computationally considerably more expensive. A histogram can be made with a single pass over the data sample so that the number of computations is simply proportional to the sample size N. Also, the whole sample need not reside in memory at one time. The most computationally efficient method for computing the Rosenblatt, Parzen and nearest neighbor estimates is to first sort the data points. This requires computation proportional to $N\log_2 N$. After sorting, these estimates can be calculated with computation simply proportional to N. Thus, these estimators require computation proportional to $N \log_2 N$. Also, all of the data points must be simultaneously in memory for the sorting.

The histogram's computational advantage is probably largely responsible for its popularity. Another reason is historical familiarity. Physicists usually learn by experience how to intuitively interpret histogram results accurately, although they seldom study the statistical foundations and approximations that lead to their techniques. Similar techniques and intuition can be learned just as well for the other density estimators. The most common objection to the consistent estimators, discussed above, is that the resulting estimation is relatively smooth and does not exhibit the sharp discontinuities ("statistical fluxuations") that are present in histograms. Statistically, this smoothness property is an asset, not a liability. It is this relative smoothness that makes these estimators more accurate than the histogram and renders them consistent

estimators. (The nearest neighbor estimator gains additional accuracy, of course, from its adaptability.)

Since histogramming is the least expensive and most popular density estimator, the following sections discuss some techniques for making the histogram a more effective tool for density estimation and data presentation.

4.2 Smoothing Counted Data

For all of the density estimators discussed in the previous sections, the variance of the estimate (statistical uncertainty) came from two sources. The first source was a systematic one. This systematic uncertainty gives rise to the inconsistency of the histogramming and orthogonal function approaches and the biases of the Rosenblatt, Parzen and k-th nearest neighbor estimators. The second source is purely statistical in nature and arises from the sampling fluxuations inherent in the random nature of the data. Associated with each of these estimators is a scale parameter (number of histogram bins, number and type of orthogonal functions, window size, h, for Rosenblatt and Parzen, and number of nearest neighbors, k). As the value of this scale parameter is varied, the amount of variance contributed from the two sources has opposite behavior. Those values that give small systematic variance usually give large statistical variance and vice versa. Thus, the choice of parameter value is a compromise between these two effects and there is usually an optimum parameter value for each specific problem where the sum from two sources is minimal. The scale can also be influenced by considerations outside the data. For example, if each data point has associated with it a measurement uncertainty, then it will make little sense to make the scale parameter much smaller than this uncertainty.

If it is possible to reduce the statistical variance by some external means, then the scale parameter can be adjusted to further reduce the systematic uncertainty, resulting in a much more accurate density estimation. This is the purpose of smoothing. Smoothing makes the assumption that the true probability density, $p(x)$, is reasonably continuous and does not change value dramatically for small changes in x. Thus, any such rapid changes in the estimate, $\hat{p}_N(x)$, must be caused by the statistical fluxuations in the estimation procedure. By taking overlapping averages of successive estimates, one hopes to dampen these fluxuations while preserving the true shape of $p(x)$. In the language of Fourier transforms, the assumption is that the Fourier transform of $\hat{p}_N(x)$, is composed of high frequency components resulting from the statistical fluxuations, and lower frequency components resulting from the true probability density $p(x)$.

The problem of smoothing is to filter out the high frequency components leaving the lower frequencies which are representative of the true probability density. [In fact, one smoothing algorithm simply Fourier transforms $\hat{p}_N(x)$, attenuates the high frequencies, and then transforms back to the original coordinate representation.]

Smoothing techniques have been extensively studied and applied in many fields to all kinds of data (not just counted data), and there are many smoothers described in the literature. Each of these smoothers has special properties and applications for which it is most effective. Only one type of smoother will be discussed here; the nonlinear, robust smoothers suggested by Tukey.[10] Although these smoothers were not specifically designed for smoothing counted data (they are probably more robust than is necessary), our experience indicates they seem to work quite well for that purpose. They are also especially easy to understand and implement.

These smoothing algorithms have three components; running medians, running means, and quadratic interpolation. Consider a sequence of observed values $\{y_i\}_{i=1}^{n}$, and it is desired to produce from them another sequence of values $\{z_i\}_{i=1}^{n}$ which will be the smoothed representation of the original set. The first ingredient of the smoothing process is running medians of three. That is

$$z_i = \text{median}\,(y_{i-1}, y_i, y_{i+1}) \,. \tag{54}$$

For the end points, we take

$$\begin{aligned} z_1 &= \text{median}\,(3z_2 - 2z_3,\ y_1,\ z_2) \\ z_n &= \text{median}\,(z_{n-1},\ y_n,\ 3z_{n-1} - 2z_{n-2}) \,. \end{aligned} \tag{55}$$

Running medians of three yield the following results:

1) monotonic sequences are unchanged.
2) points that are larger or smaller than <u>both</u> their adjacent points will be moved inward (i.e., set equal to the closest adjacent point).

Next, running medians of five are applied to the results of the running medians of three. That is,

$$z_i = \text{median}\,(z_{i-2}, z_{i-1}, z_i, z_{i+1}, z_{i+2}) \,. \tag{56}$$

For this situation, there are two special cases, the end points and next to end points. The next to end points are evaluated as medians of three:

$$z_2 = \text{median}\,(z_1, z_2, z_3)$$
$$z_{n-1} = \text{median}\,(z_{n-2}, z_{n-1}, z_n)\ . \tag{57}$$

The end points are treated as medians of one; that is, simply copied:

$$z_1 = z_1$$
$$z_n = z_n\ . \tag{58}$$

Running medians of five yield the following results:

1) monotonically rising or falling sequences are unchanged.
2) flat tops or bottoms of length three or greater are unchanged.
3) tops and bottoms (compared to two adjacent values) of length less than three are moved inward.

The final step in running medians is to apply another running medians of three to the results of the running medians of five step. In this case, however, the end points are simply copied. This part of the smoothing procedure (running medians) is called "353" for running medians of three, followed by running medians of five, followed by running medians of three. The 353 running medians procedure goes a long way toward smoothing the data. It, however, still has two shortcomings. First, rising and falling monotonic sequences are unaffected. This is not always good. A sequence can be monotonic and still not be considered smooth. As an example, consider the sequence

1 3 4 7 66 72 74 .

A second shortcoming of the 353 procedure is that it clips or flattens peaks and valleys to leave flats three values long. This gives the smoothed result an unnatural appearance of having a discontinuous derivative.

This latter deficiency can be remedied by quadratic interpolation. One looks for three adjacent equal values surrounded by values on each side that are either both lower or higher than the flat value. For each such occurrence, a quadratic fit is made through the two points adjacent to the flat, and the point in the flat next to the adjacent point with the value farthest from the three flat

value. The other two points in the flat are then given values corresponding to this quadratic.

The monotonic discontinuity problem can be dealt with by Hanning or running means (averages)

$$z_i = \frac{1}{4} z_{i-1} + \frac{1}{2} z_i + \frac{1}{4} z_{i+1} , \tag{59}$$

with simple copy

$$z_i = z_i \qquad \text{and} \qquad z_n = z_n \tag{60}$$

for the end points.

This smoother is referred to as

353QH

where 353 represents the running median block, "Q" stands for the quadratic interpolation step for the three-flats, and "H" refers to the Hanning or running averages step. This smoother (like most) follows straight lines rather well but tends to over-smooth (cut off) real peaks and valleys or any region with large second derivatives. This defect can be greatly reduced by "twicing". Consider the residuals after the smooth, defined as

$$r_i = y_i - z_i \tag{61}$$

where $\{y_i\}_{i=1}^n$ is the original data sequence and the set $\{z_i\}_{i=1}^n$ is the resulting sequence from the 353QH smoothing procedure. The sequence $\{z_i\}_{i=1}^n$ is referred to as the "smooth", whereas the sequence $\{r_i\}_{i=1}^n$ is referred to as the "rough". Twicing consists of smoothing the rough (using 353QH procedure) and adding the result to the previous smooth. That is,

$$z = \text{smooth } (y) + \text{smooth } (r)$$

or

$$z = \text{smooth } (y) + \text{smooth } [y - \text{smooth}(y)] \ . \tag{62}$$

The complete procedure, including twicing, is labeled

353QH, twice.

One can imagine many variations on this basic smoothing procedure. Beaton and Tukey[11] suggest

3G53QH, (more than twice)

as a useful alternative. Here, "G" is a conditional Hanning; if the signs of three adjacent values alternate, the middle value is replaced by the Hann of all three; otherwise, the value is unchanged. Other (simpler) algorithms that have been successfully employed are 53H, twice and 95H, twice.

As indicated above, one might consider "thricing" or even more repeated applications of twicing. In fact, one could invision a variable number of repeated applications of residual smoothing and adding until the remaining residuals meet some terminating condition.

As mentioned above, this particular type of smoother is not specifically designed for counted data. It was designed for more general situations where the fluxuations can be much more severe than those arising from purely statistical mechanisms, and where there is no information on the expected sizes of the fluxuations. In particular, it treats all similar sized fluxuations on an equal footing. For counted data, this is not desirable. For example, in histograms one knows that the statistical fluxuations are proportional to the square root of the number of counts, so that the smoother should allow larger residuals in high count regions and smaller residuals for smaller number of counts.

For histograms, this problem can be overcome by transforming the density estimate to its stable variance representation. That is

$$y_i = f[\hat{p}_N(x_i)] \tag{63}$$

where the function f(u) is chosen such that the expected statistical variance of y_i is constant. From Eq. 22, we see that

$$f(u) = \sqrt{u} \ . \tag{64}$$

Thus, if we input to the smoother

$$y_i = \sqrt{\hat{p}_N(x_i)}$$

statistical theory tells us that the statistical fluxuations are expected to be the same for all the y_i's and they should be so treated by the smoother. To obtain the smoothed density estimate, we square the smoothed output from the smoother

$$\text{smooth } [\hat{p}_N(x_i)] = z_i^2 \ . \tag{65}$$

For the other density estimators the solution is not straightforward. Since for these estimators the estimates overlap, the relative fluxuations have a more complicated dependence.

The stable variance representations for the Rosenblatt and Parzen estimators are (from Eq. 38)

$$f(u) = \sqrt{u} \tag{66}$$

and for the k-th nearest neighbor estimator (from Eq. 47)

$$f(u) = \log(u) \ . \tag{67}$$

The statistical fluxuations for these estimators are probably more constant in their stable variance representations than in their original representations. However, these estimators already provide a reasonably smooth probability density estimate so that less would be gained in applying smoothers to their results.

Figure 2a shows a histogram of the data of Figure 1, using 170 instead of 40 bins. Comparing Figures 1a and 2a, one sees that the 170 bin representation is clearly less accurate. The increase in statistical variance for the 170 bins overwhelms the decrease in systematic variance. Figure 2b superimposes over the 170 bin histogram, the result of applying the smoothing algorithm discussed above. The smoothed representation is clearly much more continuous than the unsmoothed histogram. Figure 2c plots the square root residuals between the smooth and the original data. These residuals are the differences between the square root of the original histogram and the square root of the smooth. As discussed above, these residuals are expected to have constant size. A value of ± 0.5 for a root residual corresponds to one standard deviation (values of ± 1 correspond to two standard deviations, etc.). Inspection of Figure 2c indicates that the smooth is indeed a reasonable representation of the original data. Figure 2d compares this smooth to the true probability density function, Eq. 51, from which the data were generated. The correspondence is seen to be quite good, especially in the region of the peak. To obtain a quantitative comparison, the average variance, $E[V_{710}]$, (Eq. 52) was computed for this density estimate (smooth of 170 bin histogram) using the same nine trials as for the four other density estimators. The result was

$$E[V_{710}] = .035 \ .$$

Comparing this result to those presented in Table 1, we see that this estimation procedure is more accurate than all but the k-th nearest neighbor technique.

Comparison of Figure 2d to Figures 1b-d shows why. Although the smooth has a little more statistical variance than the others, it has very little of the bias in the regions of high second derivative, $p''(x)$. This is mainly due to the twicing component of the smoother. These results also indicate that the main ingredient of the Rosenblatt and Parzen estimators that makes them more accurate than histogramming is their relative smoothness. This is also true for the k-th nearest neighbor estimate but it has the additional ingredient of adaptability to the data, yielding a further reduction in the variance.

The main advantage of the smoothed histogram estimate is its computational economy. Since it operates directly on the histogram and requires no additional information from the data, its computational requirements are very nearly the same as for the histogram. The increase in computation required for the smoothing operation is very small and is independent of the data sample size. Thus, smoothing gives the best of both worlds: the computational economy of histogramming and accuracy comparable to the consistent estimators.

There is one disadvantage to this approach. Namely, there are no formulas for the variance of the estimate analogous to Eqs. 22, 38 and 47 for the other estimators. Thus, one has to essentially guéss at the statistical uncertainty of this estimate. A very crude upper limit is, of course, provided by the variance of the histogram estimate before the smoothing. Comparing Figures 2a and 2d, we see that typically the statistical uncertainty in the smooth is a small fraction of that for the unsmoothed histogram.

An important aspect of smoothing is inspection of the residuals, as in Figure 2c. If these residuals tend to be large or have a systematic trend, then one has less confidence in the smoothed result. If this is the case, one could again smooth the residuals and add it to the data smooth (i.e., thricing). This process can be repeated until there is no change in the resulting smooth (i.e., the smooth of the remaining residuals has a constant value of zero).

4.3 Parametric Estimation

As described earlier, in parametric (model dependent) analysis the data probability density function, $p(x)$, is assumed to be a member of a parameterized family of distributions

$$p(x) = p(\vec{a}; x)$$

where $\vec{a}$ is the set of parameters that specify the particular member of the parameterized family. The problem of density estimation becomes that of

estimating the parameter values from the data distribution, i.e.,

$$\hat{p}_N(x) = p(\hat{\vec{a}}_N, x) \tag{68}$$

where $\hat{\vec{a}}_N$ are the estimated values of the parameters $\vec{a}$. As described in an earlier section, one constructs a set of statistics

$$\vec{Y} = \vec{\varphi}\,(x_1, x_2, \ldots x_N)$$

which are random variables with joint probability density function $p_N(\vec{a}, \vec{Y}_N)$ that (for a good estimator) is sharply peaked near $\vec{Y}_N = \vec{a}$.

The particular set of values $\vec{Y} = \hat{\vec{a}}$, resulting from a given set of x's (experiment), is called an estimate of $\vec{a}$. Statistics used for estimation are called estimators.

There are a wide variety of estimators useful for one-dimensional data. The technique described in Eqs. 5-10 is called the method of moments. The most highly promoted estimator is called maximum likelihood. For this estimator one forms the likelihood function

$$L_N(\vec{a}, x_1 \ldots x_N) = \prod_{i=1}^{N} p(\vec{a}, x_i) \tag{69}$$

and chooses as an estimate for $\vec{a}$, the set of values $\hat{\vec{a}}$, that maximize $L_N(\vec{a}, x_1 \ldots x_N)$ with respect to $\vec{a}$. This can be expressed as

$$\hat{\vec{a}} = \max_{\vec{a}}{}^{-1} \left[L_N(\vec{a}, x_1 \ldots x_N) \right], \tag{70}$$

or as the solution to the set of simultaneous equations

$$\frac{\partial L_N}{\partial \vec{a}} (\vec{a}, x_1 \ldots x_N) = 0 \quad . \tag{71}$$

Usually, in practice, one uses as the estimator

$$\begin{aligned} w_N(\vec{a}; x_1 \ldots x_N) &= \log L_N(\vec{a}, x_1 \ldots x_N) \\ &= \sum_{i=1}^{N} \log p(\vec{a}, x_i) \\ &= N\, E_x \left[\log p(\vec{a}, x) \right] \end{aligned} \tag{72}$$

Since the logarithm is a monotonic function of its argument, the estimates will be the same. There are numerous plausible arguments for why the maximum likelihood estimator should be good but the essential results are:

1) the likelihood estimator is consistent.
2) the likelihood estimator is asymptotically efficient. That is, as $N \to \infty$ the likelihood estimate has minimum possible variance.[3)]
3) Asymptotically (again as $N \to \infty$), one has

$$p_N(\vec{a}, \hat{\vec{a}}) = L(\vec{a}; \hat{\vec{a}}) = \frac{1}{(2\pi)^{m/2} |\underset{\sim}{\Sigma}|} e^{-1/2 (\hat{\vec{a}} - \vec{a}) \underset{\sim}{\Sigma}^{-1} (\hat{\vec{a}} - \vec{a})} \quad (73)$$

i.e., $\hat{\vec{a}}$ has a normal density distribution centered at $\vec{a}$, where

$$\lim_{N \to \infty} [\text{trace}(\underset{\sim}{\Sigma}^{-1})] = 0 .$$

It must be emphasized that the maximal efficiency of the likelihood estimator is purely an asymptotic property, and for finite N, there may well be other estimators that are more efficient for particular problems. It should also be kept in mind that the likelihood estimator is biased for most problems. (Since the estimator is consistent, the bias must approach zero as the sample size increases toward infinity.)

For large sample size, N, the empirical likelihood function can be used to estimate the variance of the estimate as well as the mean. Assuming the sample size is large enough so that Eq. 73 is a good approximation to $p_N(\vec{a}; \hat{\vec{a}})$, and that $\hat{\vec{a}}$ is a good approximation to $\vec{a}$, one can estimate $\underset{\sim}{\Sigma}$ as

$$\underset{\sim}{\Sigma}^{-1} = \left[\frac{\partial^2 w}{\partial \vec{Y}^2} (\hat{\vec{a}}; \vec{Y}) \right]_{\vec{Y} = \hat{\vec{a}}} . \quad (74)$$

This equation is commonly used in particle physics to determine the variance of likelihood estimations. It should be noted that this procedure is an approximation on two counts. First, and most important that the empirical likelihood function obtained for a particular experiment, $L_N(\hat{\vec{a}}; x_1 \ldots x_N)$, is a good approximation to the true likelihood function $L_N(\vec{a}; x_1 \ldots x_N)$ (i.e., the one that would be obtained by averaging over all possible experimental results), and second, that the likelihood function has the multivariate normal shape given by Eq. 73.

Besides the method of moments and the maximum likelihood method, one can construct estimators by first performing a nonparametric density estimation, $\hat{p}_N(x)$, (as discussed in the previous sections) and then forming a dissimilarity measure between the nonparametric estimate and the parametric representation

$$d(\vec{a}) = \int_R D[\hat{p}_N(x),\ p(\vec{a}, x)]\, dx \ . \tag{75}$$

Some examples of dissimilarity functions are:

$$D[\hat{p}_N(x)\ ,\ p(\vec{a},x)] = |\hat{p}_N(x) - p(\vec{a}; x)|^{\ell} \tag{76a}$$

or

$$D[\hat{p}_N(x),\ p(\vec{a},x)] = \left\{ \frac{|\hat{p}_N(x) - p(\vec{a};x)|}{\sqrt{V_N[\hat{p}_N(x)]}} \right\}^{\ell} \tag{76b}$$

where ℓ is greater than zero. The quantity $V_N[\hat{p}_N(x)]$ is the variance of the density estimate at the point x. The estimate for the parameters, $\vec{a}$, is taken to be those values, $\hat{\vec{a}}$, that minimize the dissimilarity measure between the nonparametric density estimate and the parameterization, i.e.,

$$\hat{\vec{a}} = \min_{\vec{a}}^{-1}[d(\vec{a})] \ , \tag{77a}$$

or is the solution to the set of equations

$$\frac{\partial d(\vec{a})}{\partial \vec{a}} = 0 \ . \tag{77b}$$

As an example, if one chooses histogramming for the nonparametric density estimator, $\hat{p}_N(x)$, and Eq. 76b with $\ell = 2$ as the dissimilarity measure, then from Eqs. 75, 14, 15 and 20 one has

$$d(\vec{a}) = \sum_{i=1}^{M} \frac{[\hat{n}_i - \bar{n}_i(\vec{a})]^2}{\bar{n}_i(\vec{a})} \tag{78}$$

where M is the number of bins or channels, r_i, and

$$\hat{n}_i = N \int_{r_i} \hat{p}_N(x)\, dx$$

and (79)

$$\bar{n}_i(\vec{a}) = N \int_{r_i} p(\vec{a}; x)\, dx \ .$$

This, of course, is the familiar least squares estimator used extensively in particle physics data analysis. Another dissimilarity measure often used with the histogramming density estimator is

$$d(\vec{a}) = - \log(N!) \sum_{i=1}^{M} N \hat{p}_i \log \bar{p}_i(\vec{a}) - \log [(N\hat{p}_i)!] \ . \tag{80}$$

This dissimilarity measure is often referred to as the log binned likelihood method. However, it is important to distinguish it from the actual maximum likelihood method of Eqs. 69-72, which does not require a preliminary nonparametric density estimate.

Other nonparametric density estimators, as well as other dissimilarity measures, may be used. Since the consistent estimators (Rosenblatt, Parzen, k-th nearest neighbor) tend to be more accurate than histogramming, parametric estimation using them will also tend to be more accurate (have smaller variance). The best value for the power ℓ (or more generally the choice for a dissimilarity measure) depends on the particular problem at hand. It can be shown that if $\hat{p}_N(x)$ has a normal distribution centered at $p(\vec{a}, x)$ with variance $V_N[\hat{p}(x)]$, then Eq. 76b with $\ell=2$ (i.e., least squares) is optimum. For other distributions of $\hat{p}_N(x)$ other dissimilarity measures are best. For example, if $\hat{p}_N(x)$ has a square window function distribution centered at $p(\vec{a},x)$ with width $w(x)$, then $\ell = \infty$ would be optimum, i.e.,

$$d(\vec{a}) = \max_x \left[\frac{|\hat{p}_N(x) - p(\vec{a},x)|}{w(x)} \right] \ . \tag{81}$$

Generally, the smaller the tails of the probability distribution of $\hat{p}_N(x)$ about $p(\vec{a},x)$, the larger the optimum value of ℓ becomes. On the other hand, the higher the value for ℓ the less robust the estimate becomes. As mentioned

earlier, least squares (ℓ=2) is already a very non-robust estimator so that high values of ℓ (including ℓ=2) should be used with great care.

Estimators formed by constructing dissimilarity measures between nonparametric density estimates and the parameterized density, are usually much less efficient than the more direct methods of moments and maximum likelihood. This is because of the two-step nature of the estimation. First, the nonparametric density estimate must be made, and then this nonparametric estimation is used as input for a parametric estimation. Both stages involve statistical uncertainty. Since nonparametric procedures generally tend to have low efficiency, the first stage tends to contribute most heavily to the statistical uncertainty of the total estimate. Also, for nonparametric density estimation one must choose the value of the scale parameter and there are generally no good guidelines for this. Finally, there is the additional statistical uncertainty introduced by the second (parametric) stage of the estimation procedure, as well as further arbitrary parameters associated with the choice for a particular dissimilarity measure.

The advantage of this two-stage procedure is that it provides more information. Namely, the actual value of $d(\hat{\vec{a}})$ at the solution can be used as a measure of the goodness-of-fit of $p(\hat{\vec{a}},x)$ to the data. Goodness-of-fit testing is discussed below under hypothesis testing. The direct parametric estimators that do not involve a preliminary nonparametric density estimate cannot be used for goodness-of-fit testing. However, this is no real disadvantage since one can use them for the estimation procedure and then do a subsequent goodness-of-fit test.

Another advantage of the two-stage procedure is computational economy. Usually these statistics (especially when histogramming is used) involve considerably less computation than the direct parametric estimators.

5. A MINI-INTRODUCTION TO HYPOTHESIS TESTING

The purpose of hypothesis testing is less ambitious than density estimation. For the latter, the attempt is to infer from the random data sample the actual probability density function from which it was drawn. For hypothesis testing, one wishes to use the random sample to simply confirm or reject a preconceived notion (theory) concerning a property of $p(\vec{x})$, or to distinguish between two or several possible properties.

Like estimation, hypothesis testing divides into the two subclasses, parametric and nonparametric. In parametric, $p(\vec{x})$ is assumed to be a member of a parameterized family of density distributions $p(\vec{a}; \vec{x})$. However, instead of trying to estimate the most likely values for $\vec{a}$ as in estimation, the purpose is to accept or reject the proposition that $\vec{a}$ has a preconceived value, or to distinguish between several alternate preconceived values. In nonparametric hypothesis testing, no parameterized family is assumed for $p(\vec{x})$ and the hypotheses concern general properties of $p(\vec{x})$ that are formulated independently of its specific functional form.

One of the most common hypotheses to be tested is that $p(\vec{x})$ is a particular function of $\vec{x}$, $p(\vec{x}) = f(\vec{x})$. This preconceived notion is to be tested against the notion that $p(\vec{x}) \neq f(\vec{x})$. That is, the hypothesis $p(\vec{x}) = f(\vec{x})$ is to be tested against all possible alternate hypotheses. This type of hypothesis testing is called goodness-of-fit testing and is discussed in the next section.

This section deals with using the data points $\{\vec{x}_i\}_{i=1}^{N}$ to test a specific hypothesis, H_0, (referred to as the null hypothesis) against a specific alternate hypothesis H_1. As in estimation, one constructs a statistic

$$Y = \Phi(\vec{x}_1, \vec{x}_2, \ldots, \vec{x}_N) \tag{82}$$

from the data points. This statistic is a random variable with probability density $p_N^{(0)}(Y)$ if the null hypothesis, H_0, is true and $p_N^{(1)}(Y)$ if the alternate hypothesis, H_1, is true. The design goal is to choose a statistic such that $p_N^{(0)}(Y)$ is as different as possible from $P_N^{(1)}(Y)$ for the given two hypotheses. Specifically, the overlap

$$\int_R p_N^{(0)}(Y)\, p_N^{(1)}(Y)\, dY \tag{83}$$

should be as small as possible. Here R is the range of all possible values of Y. Statistics used for hypothesis testing are called <u>test statistics</u>. Clearly a value of Y, for which $p_N^{(0)}(Y)$ is large while $p_N^{(1)}(Y)$ is small, is evidence for the truth of H_0 and vice versa. In hypothesis testing, one divides the test statistic space, R, into two regions: a region of rejection, r, and a region of acceptance, R-r, such that H_0 will be regarded as false (H_1 true) if the value of Y is in r, and H_0 will be regarded as true (H_1 false) if the value of Y falls in the region R-r. The quantity

$$\alpha_N = \int_r p_N^{(0)}(Y)\, dY \tag{84}$$

is called the level of significance or size of the test. It is the probability that the null hypothesis, H_0, will be declared false when it is, in fact, true. Rejection of H_0 when it is true is called loss or error of the first kind. The quantity

$$\beta_N = \int_{R-r} p_N^{(1)}(Y)\, dY \tag{85}$$

is the probability that H_0 will be declared true when it is, in fact, false (H_1 is true). This is called contamination or error of the second kind. The quantity

$$1 - \beta_N = \int_r p_N^{(1)}(Y)\, dY \tag{86}$$

is the probability that H_1 will be declared true when it is, in fact, true, and is called the power of the test. Clearly the rejection region, r, should be chosen so that for a given size, α_N, the contamination, β_N, is as small as possible (power $1-\beta_N$ as large as possible). The experimenter usually decides what loss, α_N, he can tolerate and then chooses a test statistic and rejection region so as to maximize the power, $1-\beta_N$, of the test. Clearly, to be able to do hypothesis testing, the probability density functions $p_N^{(0)}(Y)$ and $p_N^{(1)}(Y)$ must be known or calculatable for the chosen test statistic.

Consider the following very simple example

H_0: p(x) is a normal distribution with mean $\mu = 0$

H_1: p(x) is a normal distribution with mean $\mu = \mu_1$ (where $\mu_1 > 0$).

A good test statistic for this problem is

$$t = \frac{\frac{1}{N}\sum_{i=1}^{N} x_i}{\left[\frac{1}{N}\sum_{i=1}^{N} x_i^2 - \left(\frac{1}{N}\sum_{i=1}^{N} x_i\right)^2\right]^{1/2}} \tag{87}$$

which is known as a t-statistic. The probability density function $p_N^{(0)}(t)$ if the null hypothesis, H_0, is true can be shown to be a Students-t distribution with (N-1) degrees of freedom

$$p_N^{(0)}(t) = \frac{\Gamma(N/2)}{\sqrt{N\pi}\,\Gamma\left(\frac{N-1}{2}\right)} \frac{1}{\left(1 + \frac{t^2}{N(N-1)}\right)^{N/2}}$$

while for the alternate hypothesis, $p_N^{(1)}(t)$ is a similar Student's-t distribution centered at μ_1. For large N ($\gtrsim$ 100), the Student's-t distribution tends toward the standard normal distribution so that

$$p_N^{(0)}(t) = \frac{1}{\sqrt{2\pi N}} \; e^{-t^2/2N} \qquad (N > 100)$$

$$p_N^{(1)}(t) = \frac{1}{\sqrt{2\pi N}} \; e^{-(t-\mu_1)^2/2N} \; .$$

If the experimenter is willing to tolerate a loss of α, then the best rejection region is defined as

$$r_N(\alpha) \leq t < \infty$$

where $r_N(\alpha)$ is the solution to

$$\alpha = \int_{r_N(\alpha)}^{\infty} \frac{1}{\sqrt{2\pi N}} \; e^{-t^2/2N} \; dt \qquad (88a)$$

or

$$r_N(\alpha) = \Phi^{-1}(1 - \alpha)/\sqrt{N} \qquad (88b)$$

where $\Phi^{-1}(x)$ is the inverse of the standard normal error function. The power of the test is

$$1 - \beta_N = \int_{r_N(\alpha)}^{\infty} \frac{1}{\sqrt{2\pi N}} \; e^{-\frac{(t-\mu_1)^2}{2N}} \; dt \qquad (89a)$$

or

$$1 - \beta_N = \Phi \left[\sqrt{N}(\mu_1 - r_N(\alpha)) \right] \qquad (89b)$$

where $\Phi(x)$ is the standard normal error function.

Like estimators, test statistics are rated by several qualities: consistency, efficiency, bias and robustness.

1. Consistency

A test is said to be consistent if the power, $1 - \beta_N$, approaches unity as N approaches infinity,

$$\lim_{N \to \infty} (1 - \beta_N) = 1 \quad . \tag{90}$$

That is, the ability to distinguish between the two hypotheses becomes better with additional data for very large samples. Note that from Eq. 89 we see that the t-statistic (Eq. 87) is consistent.

2. Bias

A test is said to be biased if the probability of accepting the null hypothesis is greater when it is false than when it is true. Conversely, a test is said to be unbiased if the probability of accepting the null hypothesis, H_0, is greatest when it is true. From Eq. 89 we see that the t-statistic (Eq. 87) is unbiased. As for estimators, a test can be both biased and consistent since bias is a property of test statistics for finite N, while consistency refers to their properties for infinite sample size.

3. Efficiency

The efficiency of a test refers to its power for given hypotheses and level of significance. A test is said to be efficient or most powerful if it has the largest power possible for a given size, α, and given hypotheses, H_0 and H_1. A test that is most powerful for all alternate hypotheses under consideration is called a uniformly most powerful test. The efficiency of a test is the ratio of its power to the most powerful test in the given situation.

4. Robustness

Robustness for test statistics has similar meaning as for estimators. Namely, the effect on the power of the test when the underlying density distribution of the data deviates from a priori assumptions. Tests that suffer great loss of power when the density distribution $p(\vec{x})$, is different than that hypothesized are non-robust, while those that maintain reasonable power over a wide range of density distributions are robust. Clearly, robust estimators are required for nonparametric applications where $p(\vec{x})$ is not known.

The t-statistic (Eq. 87) is uniformly most powerful for all alternate hypotheses $\mu = \mu_1$ (i.e., for all μ_1) provided that the distribution of the data points, $p(x)$, is normal. However, if the data points are not normally distributed then this test can become very inefficient. For example, if $p(x)$ is a Cauchy

distribution

$$p(x) = \frac{1}{1+x^2} \tag{91}$$

then this test has zero power. Thus, the t-statistic of Eq. 87 is not robust. It is possible, however, to formulate robust analogs of this t-statistic that have reasonable efficiency for all p(x).

5.1 Goodness-of-fit Testing

Goodness-of-fit testing is probably the most common type of hypothesis testing in high energy particle physics. Here the null hypothesis, H_0, is that the probability density function of the data is a specified one. That is,

$$H_0 : p(\vec{x}) = f(\vec{x})$$

where $f(\vec{x})$ is explicitly given. Here a specific alternate hypothesis is not given. Or alternatively, one can consider the alternate hypothesis to be

$$H_1 : p(\vec{x}) = \left\{g(\vec{x})\right\}$$

where $\left\{g(\vec{x})\right\}$ is the set of all possible alternatives to $f(\vec{x})$. Either way the alternate hypothesis is not explicitly specified so that it is impossible to calculate a contamination or power of the test.

As in regular hypothesis testing, goodness-of-fit testing starts with choosing a test statistic,

$$Y = \Phi(\vec{x}_1, \vec{x}_2, \ldots \vec{x}_N) .$$

The test statistic space, however, is not divided into an acceptance and rejection region since there is no specific alternate hypothesis to accept in favor of the null hypothesis. It should be noted that since the alternate hypothesis is the set of all possible alternatives to H_0, there is surely one out of the infinity of alternatives that will always fit better than H_0. A trivial example is

$$f(\vec{x}) = \frac{1}{N} \sum_{i=1}^{N} \delta(\vec{x} - \vec{x}_i) \tag{92}$$

where $\left\{\vec{x}_i\right\}_{i=1}^{N}$ are the actual data points.

In goodness-of-fit testing one calculates the probability under H_0, that the test statistic would have a value less probable than the value actually observed from the data. This is known as the level of significance for the test. For example, if the test statistic distribution under H_0, $p_N^{(0)}(Y)$, decreases for large increasing Y, then

$$\alpha_N(T) = \int_T^\infty p_N^{(0)}(Y)\, dY$$

would be this level of significance for a given value of T. This quantity is usually referred to as the "confidence level" in particle physics. In most other scientific fields it is known as the "P-value".

Clearly, the probability density of the test statistic, $p_N^{(0)}(Y)$, under the null hypothesis must be known or be calculatable from the hypothesized data probability density $p(\vec{x}) = f(\vec{x})$. For some tests the probability distribution, $p_N^{(0)}(Y)$, of the test statistic is independent of the data distribution and depends only on the number of data points, N, and the truth of the null hypothesis. These are called distribution free tests. Most of these distribution free tests are distribution free only in the limit of infinite sample size. For this case, the distribution of various types of averages will be normal from the central limit theorem independent of the underlying distribution of the data. However, the actual sample size, N, required for the asymptotic approximation to be valid does depend upon the underlying probability density distribution of the data.

Univariate goodness-of-fit tests are constructed by formulating a dissimilarity measure between a nonparametric density estimation, $\hat{p}_N(x)$, of the data and the hypothesized functional form $p(x) = f(x)$,

$$d(x_1 x_2 \ldots x_N) = \int_R D\,[\hat{p}_N(x; x_1 x_2 \ldots x_N),\, f(x)]\, dx \quad . \qquad (93)$$

Two of the most common dissimilarity measures are given by Eqs. 76a and 76b. This procedure is identical to the procedure described earlier for formulating parametric estimators from nonparametric density estimates. Here, however, the objective is to determine the goodness-of-fit from the value of the dissimilarity, d, rather than trying to find the values of parameters that minimize it. It is clear that any goodness-of-fit statistic can also be used as an estimator by simply adjusting parameters of f(x) to achieve the best fit (i.e., minimum dissimilarity, d). However, as discussed in the earlier sections, these are seldom

the best estimators for a given problem. On the other hand, there are no goodness-of-fit analogs to the method of moments and maximum likelihood estimators. That is, the value of the likelihood function at its maximum gives no information as to goodness-of-fit.

Goodness-of-fit statistics of the form given by Eq. 93 tend to be distribution free for large sample size because the distribution of $\hat{p}_N(x)$ about $p(x)$ is determined mainly by the central limit theorem (law of large numbers) since the $\hat{p}_N(x)$ are local averages.

The most common goodness-of-fit test statistic used in particle physics is the Pearson's χ^2 test, whose test statistic is given by Eq. 78, with

$$\bar{n}_i = N \int_{r_i} f(x)\, dx \,. \tag{94}$$

As discussed earlier, this statistic uses histogramming as the nonparametric estimator and a scaled Eucledian distance type measure (Eq. 76b with $\ell=2$) for a dissimilarity measure. For large $\hat{n}_i = N\hat{p}_i$, the central limit theorem requires that $\hat{n}_i$ be normally distributed about its center $\bar{n}_i$ (under H_0) with variance $\bar{n}_i$. Thus, each term in the sum is a random variable with a standard normal distribution. It can be shown that a random variable which is the sum of squares of M normally distributed random variables has the probability density distribution

$$p_M^{(0)}(\chi^2) = \frac{1}{2\Gamma(M/2)} \left(\frac{\chi^2}{2}\right)^{M/2-1} e^{-\chi^2/N} \,. \tag{95}$$

This probability density distribution is known as a chi-square distribution with M degrees-of-freedom. For a given value of χ^2, determined from an experiment, the significance level or p-value is simply given by

$$\alpha_M(\chi^2) = \int_{\chi^2}^{\infty} p_M^{(0)}(\chi^2)\, d(\chi^2) \,. \tag{96}$$

It is important to emphasize that the χ^2 test is very non-robust. It is clear from Eq. 78 that those terms in the sum, for which $\bar{n}_i$ is very small, will dominate. Thus, for these terms a very small departure from the assumptions that lead to Eq. 95 will give rise to large departure in the results. Specifically, if the expected number of counts $\bar{n}_i = N\bar{f}_i$ is small, then the central limit theorem

cannot be applied and the residuals

$$r_i = \frac{\hat{n}_i - \bar{n}_i}{\sqrt{\bar{n}_i}} \tag{97}$$

will not have the standard normal distribution. For this case, the probability distribution for the test statistic, $p_M^{(0)}(\chi^2)$, deviates considerably from the χ^2 distribution (Eq. 95).

The precise distribution for $\hat{n}_i = N\hat{p}_i$ is given by Eq. 17. This distribution is reasonably well approximated by a normal for $\bar{n}_i \gtrsim 5$. Therefore, if all of the bins have at least five expected counts, Eq. 95 can be accepted as a good approximation. If this is not the case, several remedies are possible. Tukey[13] suggests replacing the observed number of counts, $\hat{n}_i$, by

$$\hat{s}_i = 2 + 4\,(\hat{n}_i) \tag{98a}$$

and the expected number of counts, $\bar{n}_i$, by

$$\bar{s}_i = 1 + 4\,(\bar{n}_i)\ . \tag{98b}$$

These quantities are called "started counts". The motivation for using these started counts, is that if the "raw counts", $\hat{n}_i$, have the distribution given by Eq. 17 then the $\hat{s}_i$ will be much more nearly normally distributed than the raw counts. The reason for using a smaller start for the expected number of counts, $\bar{n}_i$, is due to the asymmetry of the distribution of $\hat{n}_i$ (Eq. 17) about $\bar{n}_i$. This distribution is skewed towards lower number of counts. For example, if one count is expected ($\bar{n}_i = 1$) then zero counts ($\hat{n}_i = 0$) will be observed twice as often as two counts ($\hat{n}_i = 2$). (In fact, zero counts will be observed as often as one count!) Giving the expected number of counts a smaller start helps compensate for this skewness so that $\hat{s}_i$ is more nearly symmetrically distributed about $\bar{s}_i$ as required by a normal distribution. Using started counts ($\hat{s}_i$ and $\bar{s}_i$) instead of raw counts ($\hat{n}_i$ and $\bar{n}_i$) allows Eq. 95 to remain a good approximation for smaller sample sizes.

Another method for dealing with small sample size is to use the log binned likelihood dissimilarity measure given by Eq. 80. It can be shown that -2d has (like the χ^2 test statistic) a χ^2 distribution with M-1 degrees of freedom for infinite sample size. However, it is generally felt that this property remains a good approximation for smaller sample size than with the χ^2 test statistic.

Other goodness-of-fit tests can be constructed that use analogs of the other density estimators. The two most common are the Smirnov-Cramer-Von Mises test[14] and the Kolmogorov test.[15] These both use the Rosenblatt type estimator for the nonparametric density estimation. However, instead of estimating the probability density, $\hat{p}_N(x)$, one estimates the cumulative density function

$$\hat{P}_N(x) = \int_{-\infty}^{x} \hat{p}_N(x)\, dx \tag{100}$$

using Eq. 32a. This has the advantage that no scale parameter, h(N), need be specified.

The Smirnov-Cramer-Von Mises test uses a Eucledian type dissimilarity measure (Eq. 76a, $\ell=2$). That is

$$Y = d(x_1 x_2 \ldots x_N) = \int_{-\infty}^{\infty} [\hat{P}_N(x;x_1 x_2 \ldots x_N) - F(x)]^2\, dF(x) \tag{101}$$

where F(x) is the cumulative distribution function of f(x)

$$F(x) = \int_{-\infty}^{x} f(x)\, dx \ . \tag{102}$$

For the Kolmogorov test, a dissimilarity measure analogous to Eq. 81 is used (i.e., Eq. 76a, $\ell=\infty$). That is

$$Y = d(x_1 x_2 \ldots x_N) = \underset{-\infty< x <\infty}{\text{maximum}} [\,|\hat{P}_N(x;x_1 x_2 \ldots x_N) - F(x)|\,] \ . \tag{103}$$

The probability density function of the test statistic, $p_N^{(0)}(Y)$, has been calculated for these tests and their level of significance, $\alpha(Y)$, as a function of test statistic value, Y, are tabulated in standard statistical tables.

These estimators can also be used for nonparametric goodness-of-fit tests. That is, instead of comparing the experimental point set to a specific functional form, it is compared to another experimental point set. Consider two different point sets $\{x_i\}_{i=1}^{N}$ and $\{y_i\}_{i=1}^{M}$ drawn from unknown probability density functions p(x) and q(y), respectively. The null hypothesis is

$$H_0: \ p(x) = q(y) \quad \text{for all x and y.}$$

The test is nonparametric because no information is assumed to be known about either p(x) or q(y). The hypothesis is only that they are the same. Here one

constructs a test statistic by forming a dissimilarity measure between the probability density estimates for the two point sets. For example,

$$d(x_1 x_2 \ldots x_N; y_1 y_2 \ldots y_N) = \int_{-\infty}^{\infty} [\hat{P}_N(z; x_1 x_2 \ldots x_N) - \hat{P}_M(z; y_1 y_2 \ldots y_M)]^2 d\hat{P}(z) \tag{104}$$

where

$$\hat{P}(z) = [N\hat{P}_N(z) + M\hat{P}_M(z)]/(N + M)$$

for the analog of the Smirnov-Cramer-Von Mises test, and

$$d(x_1 x_2 \ldots x_N; y_1 y_2 \ldots y_M) = \underset{-\infty < z < \infty}{\text{maximum}} |\hat{P}_N(z; x_1 x_2 \ldots x_N) - \hat{P}_M(z; y_1 y_2 \ldots y_M)| \tag{105}$$

for the Kolmogorov.

As is the case for their parametric counterparts, these tests are distribution free (independent, under H_0, to whatever $p(x) = q(y)$ might be), and are tabulated in statistical tables.

5.2 Visual Representation of Goodness-of-Fit

Quite often a single number representing the significance level for a goodness-of-fit test is not enough, especially if the significance is marginal or small. The experimenter usually would like to know those values of the measured variable where the correspondence between the theory and data is relatively good and, similarly, those regions that most contribute to making the fit bad. The goodness-of-fit test statistics themselves can often be used for this purpose. For example, if a χ^2 test statistic is used, one can look for those terms in the sum that are relatively large (or small). For the Smirnov-Cramer-Von Mises and Kolmogorov tests (Eqs. 101 and 103), one can look for values of the integration variable, x, that result in large or small values of

$$\frac{d}{dx} [\hat{P}_N(x; x_1 x_2 \ldots x_N) - F(x)] \; . \tag{106}$$

Since one-dimensional density is easily represented in graphical form (plotted as density vs. coordinate value), a common procedure is to simultaneously plot on the same graph the nonparametric density estimate $\hat{p}_N(x)$ and the hypothesized functional form, $f(x)$. The experimenter can then visually evaluate the goodness-of-fit. This technique suffers from some drawbacks. First, the variance of the nonparametric estimate, $V[\hat{P}_N(x)]$ (statistical uncertainty), is

usually a nonconstant function of the coordinate, x. In order to successfully evaluate local goodness-of-fit for some value of x, one must known the value of $V[\hat{P}_N(x)]$. Also, if this variance varies widely over the range of x, then a quick recognition of those regions that represent significant departures is difficult. One way to alleviate this is to plot, in addition to $\hat{p}_N(x)$ and f(x), two more quantities

$$\hat{\sigma}_+(x) = \hat{p}_N(x) + \sqrt{V[\hat{p}_N(x)]}$$

and (107a)

$$\hat{\sigma}_-(x) = \hat{p}_N(x) - \sqrt{V[\hat{p}_N(x)]} \ ,$$

or alternatively,

$$\sigma_+(x) = f(x) + \sqrt{V[\hat{p}_N(x)]}$$

and (107b)

$$\sigma_-(x) = f(x) - \sqrt{V[\hat{p}_N(x)]} \ .$$

In the case of Eq. 107a, these are usually represented as "error bars" centered on the corresponding density estimates. This procedure overcomes the handicap at the expense of making the graph considerably more cluttered and unreadable.

When the explicit functional form of $V[\hat{p}_N(x)]$ is known, a better solution is to transform the ordinate of the graph, $\hat{p}_N(x)$, to the stable variance representation. That is, both $\hat{p}_N(x)$ and f(x) are transformed

$$\hat{p}^*_N(x) = T[\hat{p}_N(x)]$$

$$f^*(x) = T[f(x)] \qquad (108)$$

where the transformation function, T[u], is chosen such that the statistical variance is a constant. For the histogram, Rosenblatt and Parzen estimators

$$T[u] = \sqrt{u} \qquad (109a)$$

whereas for the kth nearest neighbor estimator

$$T[u] = \log u \ . \qquad (109b)$$

In particular, a histogram in which the square root of the number of counts (rather than the number of counts) is plotted is called a rootogram. Rootograms have the advantage that the variance of the estimate (expected size of statistical fluxuations) is a constant independent of x. Thus, a difference between

$\hat{p}^*_N(x) = \sqrt{\hat{p}_N(x)}$ and $f^*(x) = \sqrt{f(x)}$ is a direct indication of the lack of correspondence between them, and these regions can be identified quickly at a glance.

Another problem with this general technique is that humans are much more attuned to recognizing and evaluating departures from straight lines (especially horizontal ones) than highly curved lines. Humans easily and quickly identify and properly evaluate the significance of deviations from horizontal lines but have considerably more difficulty when the line has a steep slope or is highly curved.

This difficulty is also easily overcome by using a stable variance representation. This is because the expected departures of the residuals, $r(x) = p^*_N(x) - f^*(x)$, from zero are constant independent of $p(x)$, $f(x)$ or x. Thus, a plot of $r(x)$ vs. x can be evaluated independently of any other information. These residuals can be investigated for systematic departures from a horizontal line at $r = 0$. In the case of histograms of counts (or better "started" counts) where the fluxuations are reasonably well approximated by a normal distribution, one standard deviation corresponds to a square root residual of $r_i = \pm 0.5$, two standard deviations correspond to $r_i = \pm 1.0$ and so on, independent of the actual number of counts in the bin. Figure 2c is an example of such a plot of residuals. Using this approach of independently plotting the residuals allows one to see at a glance whether the fit is good, or to spot those regions where it is bad.

Tukey[16] recommends plotting the densities and residuals on a single plot. This procedure is illustrated in stages in Figure 3. Figure 3a shows an example of the traditional representation with a histogram plotted along with the hypothesized density function superimposed. Figure 3b shows the same plot, but where the histogram bars, instead of being aligned with the horizontal axis, are aligned with the function. The histogram residuals then appear as departures from the horizontal axis. Figure 3c shows a traditional rootogram of the same data, while Figure 3d shows the rootogram aligned with the curve. The expected constant size residuals then appear as departures from the horizontal axis. Also, the rootogram in Figure 3d is inverted so that the residuals play a more prominent role. This is called a hanging rootogram. With the hanging rootogram, the positive residuals appear as positive departures above the horizontal axis while the negative residuals appear below the axis. Finally, in Figure 3e the residuals are emphasized further by being shaded — differently for positive and negative — and the rootogram part is further suppressed. Also, horizontal lines representing ± 2 and ± 3 standard deviations are included.

The hanging rootogram allows the researcher to see at a glance both the general shape of the hypothesized function and the residuals from that function. The constant expected size residuals are contrasted against a horizontal line so that those regions of maximum departure, as well as the importance of the departure, can be recognized easily.

6. MULTIVARIATE DATA ANALYSIS

Many experiments in high energy particle physics are multivariate (sometimes referred to as multidimensional) in nature. That is, for each event several attributes or quantities are simultaneously measured. A particle reaction, resulting in n-particles in the final state, has 3n-4 independent measurables (not including spin information) associated with it. Most experiments measure several of these quantities and some can measure the complete set. Analyzing and interpreting data for these experiments is a problem in multivariate data analysis.

The concept of probability density is easily generalized for the multivariate case. Let $\vec{x}_i = \left(x_i^{(1)}, x_i^{(2)}, \ldots, x_i^{(d)}\right)$ be a set of attributes or quantities measured for each event. If the value of each quantity is plotted along a Cartesian axis, then the set of simultaneous values can be represented as a point in a Cartesian space of dimensionality, d. The entire experiment can then be regarded as a collection or swarm of such points in this d-dimensional space. Since each point contains all the information for the corresponding event, this point swarm contains all of the information of the experiment. As in the univariate case, the purpose of data analysis is to use this point swarm to make inferences concerning the joint probability density $p(\vec{x})$, defined in Eq. 1. For the multivariate case, r_i, is a small volume in the d-dimensional space. Letting this volume approach zero while n_i and N approach infinity, one defines the notion of the value of $p(\vec{x})$ at a point $\vec{x}$. Loosely speaking, $p(\vec{x}_0)$ is the probability that $x^{(1)}, x^{(2)}, \ldots, x^{(d)}$ all simultaneously have the values $x^{(1)} = x_0^{(1)}$, $x^{(2)} = x_0^{(2)}, \ldots,$ $x^{(d)} = x_0^{(d)}$. As an example, in exclusive experiments where d = 3n-4, the spin averaged Lorentz invariant amplitude squared

$$\frac{1}{\sigma} \left|M\left(x^{(1)}, x^{(2)}, \ldots, x^{(3n-4)}\right)\right|^2 = p(\vec{x}) \qquad (110)$$

is the joint probability density function in the Lorentz invariant phase space. Here, σ is the total cross section for the reaction. This joint probability density function contains all the information attainable concerning the momentum (nonspin) aspects of the experiment.

Statistical techniques for multivariate data analysis are much less well developed than univariate techniques. When they exist, multivariate techniques are usually straightforward extensions of the corresponding univariate techniques. For parametric density estimation, the method of moments and maximum likelihood are easily extended. One simply replaces the parametrized

univariate probability density, $p(\vec{a};x)$, by the parametrized joint probability density, $p(\vec{a};\vec{x})$. The general asymptotic ($N\to\infty$) statistical properties of these estimators is the same for the multivariate case as for the univariate case. However, the value of the sample size, where the asymptotic properties become good approximations, is usually much larger in the multivariate case.

Although easy to generalize conceptually, the computational complexity of both the moments and likelihood methods increases dramatically for high dimensionality. This is because of the general problems associated with the numerical evaluation of definite integrals in high dimensionality. Most joint probability density functions $p(\vec{a};\vec{x})$ that appear in high energy physics applications cannot be analytically integrated over the allowed range of the variables, R. Even for the simplest case of

$$p(\vec{a};\vec{x}) = \text{constant} \quad ,$$

the integral

$$\int_R d\vec{x}$$

cannot be explicitly evaluated when R is the region of phase space defined by momentum and energy conservation. From Eq. 5 we see that the moments method explicitly requires, in general, the evaluation of a multidimensional integral.

Although the likelihood method (Eqs. 70 or 71) does not explicitly require multidimensional integrals, it implicitly requires them through the normalization condition (Eq. 2). The likelihood function (Eq. 69) requires that $p(\vec{a};\vec{x})$ be a proper probability density function, i.e.,

$$\int_R p(\vec{a};\vec{x})\, d\vec{x} = 1 \qquad (2)$$

where R is the region of the allowed values for the variables, $\vec{x}$. Usually the Lorentz invariant amplitudes that arise in high energy physics are not so normalized, requiring the evaluation

$$p(\vec{a};\vec{x}) = \frac{|M(\vec{a};\vec{x})|^2}{\int_R |M(\vec{a};\vec{x})|^2 \, d\vec{x}} \qquad (111)$$

before the likelihood method can be applied. The value of the integral generally depends upon the values of the parameters. Thus, if an iterative scheme is used to solve Eq. 70, then the integral must be re-evaluated for each step in the iteration procedure. This can be extremely costly, computationally, especially if Monte Carlo techniques are required for the integrations.

Nonparametric multivariate density estimation is very difficult. This is due to the extreme sparseness of the multidimensional data space. Even for the very largest experiments being contemplated, the average counting density in the d-dimensional space of the measurement point swarm is very small. To make the situation even worse, even though the average density is very small there are usually one or several very small regions or surfaces of complex shape where the density becomes quite large.

For these reasons, straightforward extensions of the univariate density estimators do not usually work. For example, consider the multivariate analog of histogramming. Even if only ten bins or channels per dimension are chosen (a very course binning), there would be $(10)^d$ cells in the d-dimensional space. For d=10 and a large experiment ($N \simeq 10^6$) the average counting rate would be .0001 per cell, with a very tiny number of the cells containing all of the events.

Although avoiding cells, the Rosenblatt and Parzen estimators do not fare much better. This is because of the huge density variation in the measurement space. Thus, choosing a scale parameter, h(N), that is adequate for the sparse regions is much too big in the dense regions and vice versa. For example, consider a $p(\vec{x})$ that contains two components of equal probability content (number of events)

$$p(\vec{x}) = p_1(\vec{x}) + p_2(\vec{x})$$

where the scale (extent) of the first is 10% of the other in each of the dimensions. Then for d=10 the extent of the first component (1/2 of the events) is 10^{-10} times the extent of the other in the 10-dimensional measurement space. A scale parameter value that was useful where $p_2(\vec{x})$ dominates is 10^{10} times too large for those regions where $p_1(\vec{x})$ dominates and, conversely, if the scale parameter value was chosen to be accurate in the $p_1(\vec{x})$ region, it would not work at all in the other regions of the space (there would never be any counts within the window).

The only density estimator that has a chance of being useful for multivariate applications is the kth nearest neighbor. Because of its property of adapting to the data, it does not suffer from the aforementioned difficulties. However, there are additional problems that severely limit its usefulness as a density estimator. These problems can be understood by inspecting Eq. 46. The multivariate analog

of this equation[8] for the bias is (to second order)

$$b_N(x) \equiv E[\hat{p}_N(x)] - p(x) = \frac{\Gamma^{2/d}\left(\frac{d+2}{2}\right)}{2\pi(d+2)}\left(\frac{k}{N}\right)^{2/d}\frac{\mathrm{tr}\left[p''_{ij}(\vec{x})\right]}{p^{2/d}(\vec{x})} \qquad (112)$$

where $\mathrm{tr}[p''_{ij}(x)]$ is the trace of the Hessian (second derivative $\partial^2 p(x)/\partial x_i \partial x_j$) matrix evaluated at x. As discussed above, typical joint probability density functions in high dimensions are characterized by very small average density ($p(\vec{x})$ very small) and large variability ($\mathrm{tr}[p''_{ij}(x)]$ big). Thus, the bias of the estimate can be expected to be very large.

This large bias can be understood intuitively as follows. Consider a point located in a sparse region of the measurement space. As one expands a spherical volume centered at the point, data points will be encountered very slowly. Large increases in volume result in small accumulation of data points until the sphere borders a dense region where vast numbers of data points will be included for small changes in volume. Thus, the kth nearest neighbors of the sparse points will tend to include considerable contamination from dense regions, resulting in a large overestimate of the density at the point. As the sample size becomes infinite ($N \to \infty$) while $k(N)/N \to 0$, the kth nearest neighbors to a point will all be very close, eliminating this effect and sending the bias to zero, rendering the estimate consistent (as seen from Eq. 112). However, the sample sizes required for these asymptotic results to be useful are truly astronomical.

Because of its large bias, the kth nearest neighbor technique is not useful for direct density estimation. However, it can be successfully used for multivariate techniques where absolute density estimation is not required. Some of these techniques are discussed below.

Unlike univariate data, multivariate data is difficult to present (except for the special case of d=2). This is due to the inability of humans to perceive in more than three dimensions. Methods using interactive computer graphics to aid humans to perceive and manipulate multidimensional data is described in Refs. 17 and 18.

Because of the great difficulties in directly dealing with multivariate data, nonparametric techniques have, in the past, sought to reduce d-dimensional data to one, two or three dimensions where human perception can be employed and where nonparametric density estimation is practical. The most common tool for this dimensionality reduction is projection. With projection, one integrates over all but one or a few of the variables leaving a density function of lower

dimensionality. For example, with one dimensional projection

$$p(x_1) = \int_R p(x_1 x_2 \ldots x_d)\, dx_2 dx_3 \ldots dx_d$$

or more generally (113)

$$p(y) = \frac{d}{dy}\left[\int_R p(\vec{x})\, d\vec{x}\right]$$

where $y = T(\vec{x})$. Here $T(\vec{x})$ is some arbitrary function of the measurement variables. For two dimensions

$$p(y_1, y_2) = \frac{d^2}{dy_1 dy_2}\left[\int_R p(\vec{x})\, d\vec{x}\right] \tag{114}$$

where $y_1 = T_1(\vec{x})$ and $y_2 = T_2(\vec{x})$.

Operationally, p(y) (one-dimensional projection) can be estimated by evaluating y_i for each event, $y_i = T(\vec{x}_i)$, and performing a univariate density estimation on the resulting set of values, $\{y_i\}_{i=1}^N$. That is,

$$\hat{p}_N(y; y_1, y_2, \ldots, y_N) = \hat{p}_N\left[y; T(\vec{x}_1), T(\vec{x}_2), \ldots, T(\vec{x}_N)\right] . \tag{115}$$

By judicious choices for various transformation functions, $T(\vec{x})$, one hopes to infer some of the salient features of the multivariate density, $p(\vec{x})$. Similarly, for two-dimensional projections, one estimates

$$\hat{p}_N(y_1, y_2) = \hat{p}_N\left[y_1, y_2; T_1(\vec{x}_1) \ldots T_1(\vec{x}_N), T_2(\vec{x}_1) \ldots T_2(\vec{x}_N)\right], \tag{116}$$

or simply maps the points $\{(y_1, y_2)_i\}_{i=1}^N$ onto the two-dimensional plane (scatter plot). All of the one-dimensional density estimators discussed earlier can usually be extended to two dimensions without encountering the difficulties of high dimensionality (d > 3).

As an aid to this process, masking (making cuts) is sometimes used in conjunction with projection. With masking, only a preselected subregion of the total measurement space is chosen for projection

$$p_r(y) = \frac{d}{dy}\left[\int_{r \in R} p(\vec{x})\, d\vec{x}\right]$$

or equivalently

$$p_r(y) = \frac{d}{dy}\left[\int_R p(\vec{x})\, B(\vec{x})\, d\vec{x}\right]$$

where

$$B(\vec{x}) = \begin{cases} 1 \text{ if } \vec{x} \in r \\ 0 \text{ otherwise} \end{cases} \tag{117}$$

Operationally, $B(\vec{x}_i)$ is evaluated for each data point, x_i, and those points for which $B(\vec{x}_i) = 0$ are excluded from the density estimate, $\hat{p}_N(y)$. By clever choices for both transformation functions, $T(\vec{x})$, and masking functions, $B(\vec{x})$, one can often learn a great deal about the underlying multivariate joint probability density function, $p(\vec{x})$, describing the data.

There are several fundamental drawbacks to the projection approach. Most important is the tremendous loss of information inherent in the projection process. By integrating over all of the measurement variables but one (or two), a great deal of the information contained in the data is lost. This makes complex interrelationships between the variables very difficult to discover using projection and masking only.

Another problem is reflections due to the non-rectangular shape of the boundaries of the measurement space. Momentum and energy conservation impose complicated boundaries on the measurement variables. These boundaries cause two problems. First, their shape appears in the projection along with any structure coming from the actual density function (dynamics) of the data. Even for constant multivariate data density, projected densities have nonconstant distributions due to the shape of these boundaries. From the projections alone, it is not possible to tell whether an effect comes from the data density (dynamics) or from the boundaries (kinematics).

Another more serious problem caused by the non-rectangular nature of the boundaries is the reflection of actual dynamical effects in the multivariate density. This is illustrated in Fig. 4 with the classical example of projecting a two-dimensional Dalitz plot onto two one-dimensional projections. Here the density is made up of a constant background plus an enhancement in the vertical coordinate. Projecting the data onto the vertical coordinate, the enhancement appears and one makes the correct inference concerning its nature. Projecting the data onto the horizontal axis also reveals an enhancement in this coordinate. Inspecting the data in its full dimensionality ($d = 2$ in this case) reveals that these two enhancements result from the same cause and are not independent. However, with only the projected data this inference could not be made and the

experimenter could incorrectly interpret the horizontal coordinate enhancement as an additional independent effect in the data.

More generally, the non-rectangular nature of the boundaries causes spurious interrelationships to appear between measurement variables when they are projected onto lower dimensional manifolds that are not contained in the data. The preceding example shows that the problem can be serious when two-dimensional data are projected onto one dimension. The problem is even more serious when 7- or 13-dimensional data are projected onto one- or two-dimensional manifolds.

Another limitation of projections that is unique to high energy physics has to do with the symmetry properties of the probability density function required by identical particles. Probability density functions that describe particle physics interactions must be invariant to interchange of identical particles. The nature of this problem can be illustrated by again referring to an example of two-dimensional data projected onto one-dimensional manifolds. Consider a two-dimensional density, $p(x,y)$, that is required to be invariant to the interchange of x and y. If during the measurement process x and y do not happen to be treated equally (which is usually the case), then the measured density will not have the proper symmetry. There are two ways to remedy this situation. One is to symmetrize the data by using each data point twice. That is, the data point is entered as measured, and then entered again with its values of x and y permuted. Another technique is to completely asymmetrize the data. That is, the measured coordinates are always entered in order, say x always larger than y. This is equivalent to folding the density about the line x=y into the lower diagonal part of the plane.

The information content of the two procedures is equivalent (even though the symmetrization procedure introduces twice as many points). That is, the data density $p(x,y)$ is the same for both cases; only the boundaries have been changed. For those techniques that deal directly with the density, the asymmetrization procedure is preferred since it requires half the computation. However, the procedures are <u>not</u> equivalent if projection is used. This is due to the change in the boundaries. This is illustrated in Fig. 5 with a uniform density. In Fig. 5a, where symmetrization is used, the projections have uniform densities and the correct inference is made about the bivariate density. If asymmetrization is used, however, the projected distributions are no longer uniform but have a linear shape. Without knowing the true nature of the data in the full

dimensionality, one might incorrectly infer that this nonuniformity was a property of the data density.

In high multiplicity final states, where there tend to be several identical particles, this effect can be especially severe. Projection techniques require that symmetrization be used causing the computation to increase by 2^m where m is the multiplicity of identical particles. Fully multidimensional techniques that deal directly with the data density can use asymmetrization, avoiding this increased cost.

6.1 Mapping

With projection, the choice of transformation functions, $y = T(\vec{x})$, is usually dictated by the intuition of the researcher or suggested by some theoretical model. In addition, some techniques have been developed that use the data points themselves to suggest useful transformations. These techniques use the data point swarm in the full dimensionality to suggest those transformations or mappings to lower dimensionality that best reveal the salient features of the full dimensional data.

These mapping techniques are divided into linear and nonlinear dimension reducers. The linear methods consider only linear mapping functions

$$y = \vec{a} \cdot \vec{x} \tag{118}$$

where the vector $\vec{a} = \vec{a}(\vec{x}_1 \vec{x}_2 \ldots \vec{x}_N)$ is determined from the full dimensional data set by some criteria.

The nonlinear methods do not define a specific transformation at all. Instead, they directly associate with each point in the full dimensional space, a point in the lower dimensional space. The points in the lower dimensional space are moved around with respect to each other until their relative positions have some relationship to the relative positions of the actual data points in the full dimensional space. This relationship is usually based on some criterian involving the mutual interpoint distances from each point to all of the others.

There are a wide variety of both linear and nonlinear mapping algorithms discussed in the literature. This section will discuss two linear methods and one nonlinear method. Most of the other methods are modifications and generalizations of those discussed here.

6.1.1 Principal components (linear factor analysis)

The most widely used mapping algorithm is principal components. The assumption with this method is that those coordinate projections that exhibit the

largest data spread are most likely to reveal interesting structure. One considers the class of transformations

$$y = \hat{a} \cdot \vec{x} \tag{119}$$

where $|\hat{a}| = 1$, and varies the direction, $\hat{a}$, seeking to maximize

$$V_{\hat{a}}[y] = \frac{1}{N} \sum_{i=1}^{N} \left(\hat{a} \cdot \vec{x}_i\right)^2 - \left(\frac{1}{N} \sum_{i=1}^{N} \hat{a} \cdot \vec{x}_i\right)^2 \tag{120}$$

with respect to $\hat{a}$,

$$\hat{a}^* = \max_{\hat{a}}^{-1} \left[V_{\hat{a}}(y)\right] \quad . \tag{121}$$

One can then map the data onto the solution projection via

$$y_i = \hat{a}^* \cdot \vec{x}_i \quad . \tag{122}$$

If the mapping subspace is to be two-dimensional, then one can consider $\hat{a}^*$ as the transformation for the first dimension, i.e.,

$$y_i^{(1)} = \hat{a}^* \cdot \vec{x}_i \tag{123}$$

and consider another similar transformation for the second coordinate; that is,

$$y_i^{(2)} = \hat{b}^* \cdot \vec{x}_i$$

where $\hat{b}^*$ is the solution (124)

$$\hat{b}^* = \max_{\hat{b}}^{-1} \left[V_{\hat{b}}(y)\right]$$

subject to constraint

$$\hat{a}^* \cdot \hat{b}^* = 0 \ .$$

That is, $\hat{b}^*$ is the direction in which the data has the largest spread orthogonal to $\hat{a}^*$. In a similar manner, one could consider a third direction, $\hat{c}^*$, such that

$$\hat{a}^* \cdot \hat{c}^* = \hat{b}^* \cdot \hat{c}^* = 0$$

and (125)

$$\hat{c}^* = \max_{\hat{c}}^{-1} \left[V_{\hat{c}}(y)\right] \ ,$$

for a three-dimensional map.

Since both the criteria and transformations are linear, the directions $\hat{a}^*$, $\hat{b}^*$, $\hat{c}^*$, etc., can be explicitly solved by linear methods. First, one forms

the sample covariance matrix

$$V_{ij} = \frac{1}{N} \sum_{k=1}^{N} x_k^{(i)} x_k^{(j)} - \left[\frac{1}{N} \sum_{k=1}^{N} x_k^{(i)}\right]\left[\frac{1}{N} \sum_{k=1}^{N} x_k^{(j)}\right] \tag{126}$$

from the data. This matrix is real, symmetric, and non-negative. Thus, its eigenvectors are mutually orthogonal and its eigenvalues are all non-negative. The eigenvector associated with the largest eigenvalue corresponds to the direction $\hat{a}^*$; the eigenvector associated with the second largest eigenvalue corresponds to $\hat{b}^*$ and the third largest to $\hat{c}^*$, and so on. The eigenvalues themselves are the sample variances of the data as projected onto the eigenvectors.

If the data sample is drawn from a normal distribution (Eq. 3), then the sample covariance matrix is an estimate of the true covariance matrix and rotation to the principal components as axes will cause the probability density function to completely factor

$$p(\vec{y}) = \prod_{i=1}^{d} p_i(y_i) \quad . \tag{127}$$

That is, each of the y_i are totally independent of all the others.

The principal components method is not particularly useful as a mapping technique for exploratory data analysis in particle physics since densities are seldom normal and those directions with the largest spread are seldom those with the most structure. However, it is computationally very inexpensive and usually worth trying.

6.1.2 Projection pursuit

Projection pursuit[19] is also a linear mapping algorithm (Eq. 118) but the criteria for choosing the optimum projection, $\hat{a}^*$, is nonlinear. Here, one directly seeks those projection axes upon which the data exhibit maximum structure. Projection pursuit maximizes a projection index of the form

$$I(\hat{a}) = s(\hat{a})\, d(\hat{a}) \quad . \tag{128}$$

The first term, $s(\hat{a})$, measures the spread of the data, as projected onto the direction $\hat{a}$, as with principal components. For $s(\hat{a})$ one takes the trimmed standard deviation from the mean

$$s(\hat{a}) = \left[\sum_{i=pN}^{(1-p)N} (\vec{x}_i \cdot \hat{a} - \bar{x}_a)^2/(1-2p)N\right]^{1/2} \tag{129a}$$

where

$$\bar{x}_a = \sum_{i=pN}^{(1-p)N} \vec{x}_i \cdot \hat{a}/(1-2p) \quad . \tag{129b}$$

Here $\vec{x}_i$ are the data points ordered on their projected values $(\vec{x}_i \cdot \hat{a})$ and p is some small fraction regarded as outliers in the projection, and thus deleted. This makes the estimate robust against extreme outliers.

The term $d(\hat{a})$ is an average nearness function of the form

$$d(\hat{a}) = \sum_{i=1}^{N} \sum_{j=1}^{N} f(r_{ij})\ 1(R-r_{ij}) \tag{130}$$

where

$$r_{ij} = |(\vec{x}_i - \vec{x}_j) \cdot \hat{a}|$$

and $1(\eta)$ is a step function

$$1(\eta) = \begin{cases} 1 \text{ if } \eta > 0 \\ 0 \text{ otherwise} \end{cases} \quad .$$

Thus, only those projected distances for which $r_{ij} < R$ contribute to the sum. The function f(r) should be monotonically decreasing for increasing r in the region $r < R$, reducing to zero at r=R.

Projections onto two dimensions are characterized by two orthogonal directions $\hat{a}$ and $\hat{b}$ $(\hat{a} \cdot \hat{b} = 0)$. For this case, Eq. 129 generalizes to

$$s(\hat{a}, \hat{b}) = s(\hat{a})\ s(\hat{b})$$

and r_{ij} becomes (131)

$$r_{ij} = \left\{ \left[(\vec{x}_i - \vec{x}_j) \cdot \hat{a} \right]^2 + \left[(\vec{x}_i - \vec{x}_j) \cdot \hat{b} \right]^2 \right\}^{1/2}$$

in Eq. 130.

The algorithm is insensitive to the explicit function form of f(r) and shows dependence only on

$$\bar{r} = \int_0^R r f(r) dr \Big/ \int_0^R f(r)\ dr \qquad \text{(one dimension)}$$

or (132)

$$\bar{r} = \int_0^R r f(r)\ r dr \Big/ \int_0^R f(r)\ r dr \qquad \text{(two dimensions)}.$$

The value of $\bar{r}$ establishes the scale of density variation to which the algorithm is sensitive and thus defines the size of the structure being sought.

The projection index, $I(\hat{a})$ (or $I(\hat{a},\hat{b})$ in two dimensions) measures the degree of structure present in the data in the particular projection. The strategy of projection pursuit is to find those projections that maximize $I(\hat{a})$, i.e.,

$$\begin{aligned} \hat{a}^* &= \max_{\hat{a}}{}^{-1}[I(\hat{a})] \qquad \text{(one dimension)} \\ (\hat{a}^*,\hat{b}^*) &= \max_{(\hat{a},\hat{b})}{}^{-1}[I(\hat{a},\hat{b})] \qquad \text{(two dimensions)} \end{aligned} \tag{133}$$

Since the mapping criteria are not linear, as with principal components, numerical hill climbing methods are required to seek the maxima. The projection index is reasonably well behaved, however, so that only a few iterations of the maximizer are usually necessary to find a solution. Also, quite often several solutions exist providing several possible highly structured projections for inspection by the researcher.

Since projection pursuit directly seeks projections with high structure, it is potentially more useful than principal components. Figure 6a shows a projection of some particle physics data on the largest principal axis. Figure 6b shows the same data projected onto the projection pursuit solution obtained at the first local maximum of the projection index, uphill from the principal components solution. Figure 6c shows the same data projected onto the plane of the two largest principal components, while Figure 6d shows the corresponding projection pursuit solution.

Although the principal axis projections indicate possible structure within the data set, the projection pursuit solutions are clearly more revealing. This is indicated by the substantial increase in the projection index (p-index), and verified by visual inspection.

Because projection pursuit is a linear mapping algorithm, it suffers from well known limitations of linear mapping. The algorithm will have difficulty in detecting clustering about highly curved surfaces in the full dimensionality. In particular, it cannot detect nested spherical clustering. It can, however, detect nested cylindrical clustering where the cylinders have parallel generators.

6.1.3 Nonlinear mapping algorithms

With nonlinear mapping, there is no specific transformation function, $y=T(\vec{x})$, defined. Each projected point $\vec{y}_i$ (usually two-dimensional) is associated with a particular data point, $\vec{x}_i$, in the full dimensionality. The positions of the $\vec{y}_i$ in the two-dimensional space are altered until they match as closely as possible some property of the $\vec{x}_i$ in the full dimensional space.

Consider the nonlinear mapping algorithm of Sammon,[20] as an example. Here, the property to be matched is the mutual interpoint distances. Let D_{ij} be the distance between two points in the projection

$$D_{ij} = |\vec{y}_i - \vec{y}_j| \tag{134a}$$

and d_{ij} be the interpoint distance in the full dimensional data space

$$d_{ij} = |\vec{x}_i - \vec{x}_j| \quad . \tag{134b}$$

A mapping error function of the form

$$E(\vec{y}_1, \vec{y}_2, \ldots, \vec{y}_N) = \frac{1}{N} \sum_{i=1}^{N} \sum_{j=i+1}^{N} \left[\frac{D_{ij} - d_{ij}}{d_{ij}} \right]^2 \tag{135}$$

is then minimized with respect to the positions of the projected points $\{\vec{y}_i\}_{i=1}^{N}$. Thus, the number of variables in the minimization is twice the number of data points. The solution positions $\{\vec{y}_i^*\}_{i=1}^{N}$ give the best two-dimensional representation of the full dimensional data with respect to the interpoint distances. The value of the mapping error at the solution gives an indication of how well the two-dimensional mapping represents the full dimensional data. A very small mapping error indicates that the data lies on a two-dimensional manifold imbedded in the full dimensional space.

There is clearly no limitation on the dimensionality of the projection subspace. A one or three-dimensional projection subspace would work just as well. Also, the values of the interpoint distances are not the only possible mapping function. Shepard and Carrol,[21] for example, suggest the monotonicity of the interpoint distances as a criterian. That is, finding the mapping that best preserves the order relationship between the interpoint distances of the data points.

The principal limitation of these nonlinear mapping techniques is the computational resources they require. Memory proportional to N^2 is needed. Each evaluation of the mapping function requires a number of operations proportional to N^2. The number of variables in the search for the minimum is $D \cdot N$, where D is the dimensionality of the projection subspace. For these reasons, the nonlinear mapping techniques cannot be applied conveniently to sample sizes larger than a few hundred. Other limitations of the nonlinear mapping algorithms are that the mapping cannot be summarized by a few parameters. This makes interpretation of the resulting map difficult. Also, the mapping only exists for the data set used in the analysis so that additional data cannot be identically mapped. These nonlinear mappings, however, are the most effective tool for

picturing the interrelationships between the data points. Data points that are found to have a particular relationship can be identified and isolated. Traditional analysis can then be used to determine the physical causes of the interrelationship.

6.2 A semi-parametric technique for model fitting (prism plot analysis)

In one very special case, multivariate data analysis reduces to univariate analysis. This is when the joint probability density function is completely factorable, that is

$$p(\vec{x}) = \prod_{i=1}^{d} p_i(x_i) \quad . \tag{136}$$

In this case, each of the measurement variables is completely independent of the others and the d-dimensional problem reduces to d one-dimensional problems. One can then analyze each of these one-dimensional problems using standard univariate techniques.

An extension of this concept, to the case where the joint probability density function can be represented as a sum of terms, each of which is completely factorable (but not necessarily by the same variables) was originally proposed by Brau, Dao, Hondous, Pless and Singer[22] (although not in this formalism) and later modified by Condon and Cowell,[23] and Van Hove.[24] With this technique, the joint probability density function is assumed to have the form

$$p(\vec{x}) = \sum_{m=1}^{M} \alpha_m f_m(\vec{x}) \tag{137a}$$

where

$$f_m(\vec{x}) = \prod_{\ell=1}^{L} f_{m\ell}(y_{m\ell}) \tag{137b}$$

and

$$y_{m\ell} = T_{m\ell}(\vec{x}) \quad . \tag{137c}$$

It is further assumed that the overlap between the functions

$$\int_R f_m(\vec{x})\, f_n(\vec{x})\, d\vec{x} \qquad (m \neq n) \tag{138}$$

is not large.

The purpose is to estimate the explicit parameters α_m and perhaps other parameters that may be associated with the functions $f_{m\ell}(y)$. The procedure

begins by making a first guess as to the values of the parameters. A weight for each data point corresponding to each term in Eq. 137a is constructed

$$w_m(\vec{x}_i) = \frac{\alpha_m f_m(\vec{x}_i)}{\sum_{n=1}^{M} \alpha_n f_n(\vec{x}_i)} \quad . \tag{139}$$

Note that $\sum_{m=1}^{M} w_m(\vec{x}_i) = 1$ for each event.

For the case where the values of the parameters are the correct ones, this weight is the probability that the event is associated with the corresponding term in the sum. That is, if each term, $f_m(\vec{x})$, in the sum is regarded as an independent density function, then $w_m(\vec{x}_i)$ is the probability that $\vec{x}_i$ was drawn from this density rather than any of the others. In addition, if a projection is made of any kinematic quantity, $y = T(\vec{x})$, and the univariate density, $\hat{p}_N(y)$ is estimated with each event weighted by $w_m(\vec{x}_i)$, then result will be the same as if

$$p_m(\vec{x}) = \frac{f_m(\vec{x})}{\int_R f_m(\vec{x})\, d\vec{x}} \tag{140}$$

were the true joint probability density rather than $p(\vec{x})$ (Eq. 137). Also, the sum

$$N_m = \sum_{i=1}^{N} w_m(\vec{x}_i) \tag{141}$$

will be the number of events drawn from $p_m(\vec{x})$. This procedure allows the isolation of the contribution from each term in $p(\vec{x})$ (Eq. 137a) to the plot of any kinematic variable.

If the first guess for the values of the various parameters is not quite correct, then the events weighted with w_m will contain contributions from all of the terms. Thus, if the experimenter has a good idea of the nature of the contribution to various kinematic variables from each of the $f_m(\vec{x})$, he can adjust the parameters, recalculate the weights, and again plot the weighted distributions. This iterative procedure can continue until all of the distributions show consistency.

This iterative procedure is not confined to simply adjusting parameters. At any point in the iterative procedure, one can introduce new terms to the sum (Eq. 137a) or completely change the form of the parameterization of a term. Thus, as well as adjusting parameters, one can build the model iteratively. This procedure is, therefore, semi-parametric in the sense that the model need

not be completely specified in advance, and can, to some degree, be developed along the way. But at all times it must conform to the functional form of Eq. 137 (as must the true joint probability density of the data).

This technique is usually applied to resonance production in three and four particle final states where the joint probability density functions (transition matrix elements squared) tend to satisfy the restrictions of Eqs. 137 and 138. Also, the contributions to various kinematical variables from each term (resonance) is usually well known. The product of functions (Eq. 137b) usually contains a Breit-Wigner term in the appropriate invariant mass, as well as production and decay angular distributions. That is

$$f_m(\vec{x}) = BW(\mu_m)\ P(\theta_p)\ D(\Omega_d)$$

where μ is the invariant mass, θ_p the production angle, and Ω_d the decay angle of the resonant particles.

As a first guess for the α_m, one takes the relative heights of each resonant peak in their corresponding mass plots. The first guess for the angular distributions is usually flat. Projections of each mass plot, as well as each angular distribution, are made weighting the events with the corresponding $w_m(\vec{x}_i)$. Those weighted distributions are then inspected for consistency. For example, if an α_m is too small, there will be peaks in the corresponding μ_m plots when the events are weighted by $w_n(\vec{x}_i)$, $n \neq m$. Conversely, if α_m is too large, there will be holes at the mass of the resonance when using these other weights. If the weighted angular distributions are not flat, then specific distributions can be incorporated into the model for the next iteration. After the model has been adjusted, the weights are again calculated and the weighted distributions again plotted. These distributions are inspected for further refinements of the model, and so on until the experimenter is satisfied with all of the weighted distributions. At that point, the model and fitting procedure are complete.

The various distributions weighted with each $w_m(\vec{x}_i)$ are the same as if one had a pure sample of resonance events from each channel. The N_m (Eq. 141) are estimates for the number of events produced by each resonance channel. The resulting values of the parameters are estimates for their true values.

When this procedure can be applied, it has several strong advantages. First, nowhere in the procedure are normalization integrals required. As noted earlier, a severe computational disadvantage with the maximum likelihood

technique was the requirement to evaluate complex multidimensional integrals for normalization. These integrals are usually evaluated with time consuming Monte Carlo methods. This technique completely avoids these Monte Carlo calculations.

Another advantage is the strong interactive coupling of the experimenter to the fitting procedure. At every stage in the analysis, the researcher is directly involved in adjusting the model and evaluating the results. Thus, his intuition can be utilized to help solve the problem.

There are also disadvantages to the procedure. First, is its limited range of applicability. It can only be used when the model and the data conform to the restrictions imposed by Eqs. 137-138. Like any parametric technique, its validity depends upon the truth of the a priori assumptions concerning the true joint probability density of the data. If the true data distribution does not conform to these assumptions, there is no way to discover this with the procedure. It is the factorization assumption (Eq. 137b) that allows one to use one-dimensional projections to build a multidimensional joint probability density. There is no way to tell if the factorization hypothesis is true by looking at the projections.

There are other less important statistical limitations to the procedure. First, since it is not deterministic (involves human interaction) there are no results concerning its statistical properties as an estimator.[25] Most important, the procedure may not be consistent. Since the weight, $w_m(\vec{x}_i)$ (Eq. 139), involves all of the $f_n(\vec{x})$ terms, not just $f_m(\vec{x})$, a deficiency in one (or several) of the other terms can cause the projections weighted with $w_m(\vec{x}_i)$ to be incorrect, even if $f_m(\vec{x})$ is correct. Thus, in principle, if a weighted projection is incorrect, there is no way to tell whether this is caused by a deficiency in the weighting channel or one of the others. It is conceivable that the experimenter might make the wrong decision and modify $f_m(\vec{x})$ to improve the projection when the problem was with $f_n(\vec{x})$, $n \neq m$. In fact, one might be able to iterate to a consistent picture (all projections look correct) that is completely incorrect. The ability to arrive at a unique (and correct) solution depends strongly on the experimenter's skill and intuition. Thus, this procedure, like most parametric procedures, must be used with great care. Keeping in mind its restrictions, this technique is, however, very useful and powerful when it can be applied.

6.3 Generalized Nonparametric multivariate techniques

This section discusses some general nonparametric methods for dealing with multidimensional data. All of these methods use some variation on the kth nearest neighbor technique, but never for direct density estimation. As pointed out earlier, direct density estimation is very difficult in high dimensional spaces. Another problem with nonparametric multivariate techniques is that the test statistics are seldom distribution free. That is, the probability density function of the test statistic $p_N(Y)$ is seldom independent of the underlying joint probability density distribution, $p(\vec{x})$, of the data. For the techniques discussed below, the test statistic distributions are nearly distribution free in that the probability density functions, $p_N(Y)$, change very little for large differences in the underlying density distributions, $p(\vec{x})$, of the data. In addition, procedures are described for estimating $p_N(Y)$ directly from the data so that for any given application the significance level of the test can be determined. In this way the tests are distribution free.

6.3.1 A nonparametric procedure for comparing multivariate point sets

Consider two samples of N_1 and N_2 observations taken on vector random variables $\vec{x}$ and $\vec{y}$ with unknown joint probability density functions $p(\vec{x})$ and $q(\vec{y})$. We wish to test the null hypothesis, H_0, that $p(\vec{x}) = q(\vec{y})$ for all $\vec{x}$ and $\vec{y}$.[26] This is the multivariate analog of the nonparametric goodness-of-fit test discussed earlier.

It is often useful in high energy physics to compare two experimental point sets to determine to what extent they are similar or different. At the most straightforward level two experiments can be compared for their compatibility. Since the test makes the comparison in the full dimensionality of the data measurement space, all of the information contained in the experiments is used.

Usually one wishes to determine if changing a property of the data has any effect or consequences on the resulting joint probability density function of the experimental measurables. This property may be external or it may be one of the measurement variables themselves. For example, an experimenter may wish to test for the presence of experimental biases in his apparatus by comparing his data to similar data taken with some of the magnet currents reversed. The null hypothesis is that there are no biases, that is, the two data sets should be the same in every way. In another application, the external property could be the spin of the beam or target. Here one would like to know if there is any property of the data that is different for different signs of the spin. Data taken

with spin up is compared to data taken with spin down. A frequent application is energy dependence of multiparticle production. Here the data from similar experiments, taken at several beam energies, are compared to see if there is any energy dependence in the Lorentz invariant amplitude for the reaction. Sometimes the varying property of the data is one of the measurement variables. For example, in electroproduction one could test the data for a dependence on the mass, q^2, of the exchanged photon.

The test described below not only gives a measure (confidence level or p-value) for the compatibility of the two point distributions, but also gives information as to those regions of the multidimensional space where the correspondence between the point sets is good and where it is bad. This information can give considerable insight as to the dynamical mechanism causing the point sets to disagree.

A common application of this algorithm is in multivariate goodness-of-fit. As discussed above, there are no general multivariate goodness-of-fit statistics. If a Monte Carlo procedure can be used to generate data points distributed from a density, $p(\vec{x})$, corresponding to some model, then this algorithm can be used to compare the Monte Carlo data to the actual data. In this way, one can obtain a confidence level for the model describing the data (in the full dimensionality of the data space) as well as determining those regions of the data space where the model gives a good description and those regions where it is poor.

This technique can also be used to design experiments. Monte Carlo data from two different models can be compared. If the comparison results in a good correspondence, then there is no way the proposed experiment can distinguish between the two models. If the correspondence is not good, then the region of the multidimensional measurement space where the two models most disagree can be identified. The values of the measurables corresponding to that region can then be used to determine how to best set up the experiment to have maximum discrimination ability between the two models.

The algorithm for testing the null hypothesis, H_0, that two multivariate point samples (classes) were drawn from the same unknown joint probability density function, proceeds in the following manner. The two samples of size N_1 and N_2 respectively, are combined into a single sample of size $N = N_1 + N_2$ with each point tagged as to the class from which it originates. The closest k points to each point are examined and the number, k_1, originating from class one (or the corresponding number, $k_2 = k - k_1$, originating from class two), is

determined. Thus, associated with each point in this combined sample is a measure of the composition of the points closest to it. The observed frequency distribution of k_1, $n(k_1)$, for all the sample points, is recorded. This frequency distribution is then compared to that expected under the null hypothesis.

There are a variety of ways of testing whether the observed distribution for k_1 conforms to that expected under H_0. One technique involves comparing the frequency distribution of k_1, $n_1(k_1)$, evaluated in the neighborhoods centered at class one points to the frequency distribution of k_1, $n_2(k_1)$, evaluated in the neighborhoods centered at class two points. Under H_0, these two distributions are expected to be the same. A useful general procedure is to compare the detailed distributions of $n_1(k_1)$ and $n_2(k_1)$ to their expected distribution, $n_0(k_1)$, under H_0. The problem of comparing two multivariate point distributions is, in this way, reduced to a univariate goodness-of-fit test.

If each of the N k-neighborhoods were mutually exclusive, then the relative frequency of the possible values of k_1 would (under the null hypothesis) conform to a binomial distribution over k_1=0, 1, 2, . . ., k with probability $p=N_1/N$; that is, $n_0(k_1)$ would be a binomial distribution with k-degrees of freedom. These neighborhoods cannot be mutually exclusive, however, since there are N neighborhoods—each containing k points—with only N total sample points. Thus, there is no reason to expect the distribution of k_1 values to be compatible with such a binomial distribution. The precise distribution in the general case is difficult to derive, but Monte Carlo calculations for a wide variety of cases indicate very little discrepancy between the true distribution and a binomial. Thus, a difference between the two multivariate samples can be measured by comparing the distribution observed for the k_1 values with the corresponding binomial distribution. Any of the univariate goodness-of-fit tests described earlier may be used for this purpose. The test statistic, Y, for comparing the two multivariate point distributions is just this univariate goodness-of-fit statistic for comparing $n_1(k_1)$ and $n_2(k_1)$ to the binomial distribution $B^k_{N_1/N}(k_1)$.

Although this procedure reduces the multivariate problem to a univariate goodness-of-fit test, its test statistic distribution is not the same as when the univariate test is applied to a standard univariate problem. This is due to the lack of independence of the values of k_1 in the univariate distribution. Because the neighborhoods are not mutually exclusive but overlap considerably, the values of k_1 for neighboring events are highly correlated. These correlations cause the probability density distribution of the test statistic, $p^{(0)}_N(Y)$, to deviate

substantially from that when the univariate sample points are all independent. This deviation usually takes the form of increased variance of $p_N^{(0)}(Y)$. That is, the expected value of the test statistic, Y, is the same as that for independent data, but the variance about that mean tends to be much larger.

As discussed in the section on univariate goodness-of-fit testing, the distribution of the test statistic under the null hypothesis, $p_N^{(0)}(Y)$, must be known or calculatable in order for the test to be useful. Specifically, one must be able to calculate the significance level, $\alpha(Y)$, for the experimentally obtained value of the test statistic

$$\alpha(Y) = \int_Y^\infty p_N^{(0)}(Y')\,dY' \quad . \tag{142}$$

For this test, it is possible to use a permutation procedure to estimate the significance level of the test, directly from the experimental data. This permutation test proceeds as follows: the two samples are combined and the points randomly re-assigned to the two sample classes in the original proportion, N_1/N_2; the comparison algorithm is applied to the two newly defined samples and the value of the test statistic is obtained. Repeated application of this random permutation procedure yields a series of test statistic values. The fraction of these values that are larger than the value, Y, obtained for the unpermuted case is an estimate of the significance level, $\alpha(Y)$, for the test.

The statistical properties of this test are discussed in detail elsewhere[26] and only the results are presented here. The test is consistent. This follows from the fact that the kth nearest neighbor technique is a consistent density estimator. The test is unbiased. Even though the kth nearest neighbor density estimator is extremely biased in high dimensionality, the bias is identical under the null hypothesis for the two samples being compared, so that the comparison is unaffected and the test is unbiased. The test is extremely robust. This is because the multivariate aspect of the test uses only order statistics and no arithmetic statistics. The test has very high efficiency. This is a somewhat surprising result. For example, this nonparametric test was found to be almost as efficient as the parametric normal theory (likelihood ratio) test on normal data for differences in location (mean), and four times more efficient for differences in scale (standard deviation), with small to moderate sample sizes.

Although the permutation procedure can always be used to estimate the significance level for the test in any application, it is seldom necessary. This

is because the test statistic distribution, $p_N^{(0)}(Y)$, is remarkably independent of the underlying data density $p(\vec{x})$. Only, when the experimenter has reason to believe that this data is especially pathological, need he apply the full permutation procedure. The performance of the test is also reasonably independent of the chosen number of near neighbors, k, so long as it is not too small ($k \geq 10$). The test statistic distribution shows its strongest dependence on the dimensionality, d, of the data space for very low dimensionality. This is because the overlapping of the neighborhoods is greater for lower dimensionality. However, for higher dimensionality, this effect diminishes so that $p_N^{(0)}(Y)$ is roughly independent of the dimensionality of the data space for $d \geq 4$.

As mentioned above, it is often useful to be able to identify those regions of the multidimensional measurement space where the two point samples most disagree (or agree) in their relative densities. This algorithm assigns such an estimate to each data point in the combined sample. Those events for which $n(k_1)$ is near zero or k are located in regions where the two distributions most disagree, while those points for which $n(k_1)$ is near $(N_1/N)k$ (its expected value under H_0) are located in regions where the agreement is best. Thus, those points for which the discrepancy is very large (or small) can be identified and isolated, and their properties can be studied independently of the rest of the sample.

Figure 7 illustrates an application of this algorithm in the study of the energy dependence of the Lorentz invariant amplitude for multiple pion production in pp collisions.[27] Here, 203 events of the reaction $pp \rightarrow pp\pi^+\pi^+\pi^-\pi^-$ at 12 GeV/c are compared to 196 events at 28 GeV/c. For this application, the test statistic was formed in the following manner; the two k_1-distributions, $n_1(k_1)$ and $n_2(k_1)$, were summed to a single distribution $n(k_1) = n_1(k_1) + n_2(k_1)$, and the resulting sum compared to that predicted by the null hypothesis, $n_0(k_1)$. This comparison was done using a χ^2 test statistic

$$Y = \sum_{k_1=0}^{k} \left[n(k_1) - n_0(k_1) \right]^2 / n_0(k_1) \qquad (143)$$

Figure 7 shows frequency histograms of the number of 12 GeV/c events (class 1) in a 20 event neighborhood (k=20) about every event in the combined sample for various coordinate combinations. Figure 7a compares the two samples in the six-dimensional subspace of scaled momentum components parallel to the incident beam. The frequency histogram, for this case, deviates considerably

from the binomial distribution (open circles) expected for the null hypothesis, indicating that these two samples differ considerably in their multidimensional shapes. Figure 7b shows the results of comparing these two samples in the 12-dimensional momentum subspace transverse to the incident beam direction. Figures 7c and 7d show the comparison in the two six-dimensional cylindrical coordinate subspaces of transverse momentum. In contrast to the scaled longitudinal momenta, the frequency histograms for these cases do not deviate significantly from the expected binomial distribution. The comparison is slightly better in the azimuthal angle subspace than the subspace of the transverse momenta squared, however, neither deviates strongly from the expected binomial distribution.

These results show that the 12-dimensional shape of the differential cross-section transverse to the beam direction, $d^{12}\sigma/d(\vec{p}_\perp)^{12}$, either independent of, or at most, varies slowly with energy for this reaction in this energy range. By contrast, the shape of the six-dimensional differential cross section parallel to the incident beam, $d^6\sigma/dx^6$, is changing considerably; thus, the energy dependence of the dynamics manifests itself mostly, if not completely, in the longitudinal variables.

Inspection of Fig. 7a shows that the main source of disagreement in the longitudinal variables is an excess of events for high values of k_1 ($15 \leq k_1 \leq 20$). This means that there is a region of the six-dimensional scaled longitudinal phase space with a strong excess of 12 GeV/c events. This can be interpreted as a dynamical production mechanism that has a substantial cross section at 12 GeV/c which becomes very small at 28 GeV/c. These events, for which $15 \leq k_1 \leq 20$, can be isolated and studied separately using traditional techniques to identify the nature of this production mechanism.

Figure 8 illustrates the use of this algorithm for multidimensional goodness-of-fit testing. Here the same data is compared, in the six-dimensional subspace of azimuthal angles, to models that predict no azimuthal angle dependence in the Lorentz invariant amplitude. For this purpose, 970 Monte Carlo events of the reaction $pp \rightarrow pp(4\pi)$ at 23 GeV/c were generated according to peripheral phase space. These Monte Carlo events were compared to the data, at 23 GeV/c, in the six-dimensional azimuthal angle subspace: Figure 8 shows the results of the comparison. The frequency of data events within a 20-event neighborhood of each point in the combined sample is clearly compatible with the corresponding binomial distribution. Thus, to the statistical accuracy of this test, the

shape of $d^6\sigma/d\phi^6$ is compatible with that predicted solely by momentum conservation. In particular, jets in the transverse plane would require the data to approximately lie on a lower dimensional manifold in this six-dimensional space and would be easily detectable.

6.3.2 Multivariate tests for independence

An important property of multivariate data is the degree to which its joint probability density function, $p(\vec{x})$, factors into a product of density functions, $p_j(\vec{x}_j)$, each defined over exclusive orthogonal subspaces of the full dimensional space. That is,

$$p(\vec{x}) = \prod_{j=1}^{q} p_j(\vec{x}_j) \tag{144}$$

where $\vec{x} = \{\vec{x}_j\}_{j=1}^{q}$ and $2 \leq q \leq d$. That is, $\dim(\vec{x}_j) < d$ and $d = \sum_{j=1}^{q} \dim(\vec{x}_j)$. If the joint probability density does factor in this manner, then the vectors defined over each of the q-different subspaces are said to be stochastically independent. That is, the distribution of points in each subspace is totally independent of the distributions in the other subspaces.

A special case of this factorization was discussed earlier, namely where q=d and thus,

$$p(\vec{x}) = \prod_{j=1}^{d} p_j(x_j) \quad . \tag{145}$$

In this case, the vector components x_j are said to be totally independent. One can also speak of the pairwise independence of pairs of coordinates. A pair of coordinates (x_i, x_j) is said to be pairwise independent if their marginal two-dimensional joint probability density

$$p(x_i, x_j) = \int_R p(\vec{x}) \prod_{k \neq i,\, j}^{d} dx_k \tag{146a}$$

factors

$$p(x_i, x_j) = p_i(x_i)\, p_j(x_j) \quad . \tag{146b}$$

It is important to keep in mind the distinction between these three types of independence. It is clear that total independence (Eq. 145) implies both pairwise independence (Eq. 146) for all pairs, and stochastic independence (Eq. 144), but pairwise independence implies neither total nor stochastic independence. Even if all of the coordinates $(x_i, x_j \; i \neq j)$ are pairwise independent, this does not imply that the joint probability density function, $p(\vec{x})$, is either totally or stochastically independent. Stochastic independence implies pairwise independence

for those pairs where each coordinate comes from a different stochastically independent subspace. However, the converse is not true. Even if all pairs of coordinates between two subspaces are pairwise independent, this does not imply that the two subspaces are stochastically independent.

It is also important to keep in mind the distinction between a pairwise dependence or relationship between two coordinates and a correlation between them. In Statistics, a correlation is defined to be a linear dependence or relationship as measured by the linear correlation coefficient

$$C_{ij} = V_{ij}/\sqrt{V_{ii} V_{jj}} \quad , \tag{147}$$

where

$$V_{ij} = \frac{1}{N} \sum_{k=1}^{N} x_k^{(i)} x_k^{(j)} - \left[\frac{1}{N} \sum_{k=1}^{N} x_k^{(i)}\right]\left[\frac{1}{N} \sum_{k=1}^{N} x_k^{(j)}\right]$$

is the sample covariance matrix. The correlation coefficient can have values between ±1. A value of zero implies no correlation or linear relationship. Positive values imply a positive slope to the linear relationship while negative values imply the opposite slope.

A pair of coordinates $(x_i x_j)$ can have a pairwise dependence and be uncorrelated ($C_{ij} = 0$). Consider a pair of coordinates such that the data points all lie on the circumference of a circle, i.e., $x_i^2 + x_j^2 = a^2$ where a is a constant. This pair of coordinates is clearly related and show a pairwise dependence; however, since this relationship is purely quadratic, they are uncorrelated as can be easily verified by calculating their correlation coefficient (Eq. 147). Therefore, lack of correlation is necessary but not sufficient to insure pairwise independence.

For the special case where $p(\vec{x})$ is a multivariate normal distribution (Eq. 3), noncorrelation implies pairwise independence and pairwise independence implies total independence. Thus, a necessary and sufficient condition for total independence is that all of the pairwise correlation coefficients (Eq. 147) be consistent with zero; or, put another way, that the sample covariance matrix be diagonal. For two subspaces to be stochastically independent, it is necessary and sufficient that the correlation coefficients for all pairs, where one coordinate comes from one subspace and the other coordinate from the other subspace, be consistent with zero. These results are due to the fact that only linear relationships are possible with multivariate normal distributions.

The correlation coefficient (Eq. 147) is the most widely used statistic for measuring pairwise dependence. In fact, the term correlation is often used interchangeably with dependence. It should be kept in mind, however, that the correlation coefficient only measures linear dependence which is seldom the total dependence. In high energy physics especially, very few multivariate distributions are normal so that a small correlation does not necessarily mean a small dependence. A general test for independence must be able to detect nonlinear relationships between coordinate pairs, and since pairwise independence does not imply stochastic independence, it must be able to detect stochastic dependence directly. In the next section, we discuss generalized tests for pairwise independence and in the following section an even more generalized test for stochastic dependence.

6.3.2.1 The mutual information measure for pairwise dependence

The mutual information measure[28] uses the entropy of a probability density function. The entropy, H, of a probability density function is defined as

$$\begin{aligned} H &= -\int_R p(\vec{x}) \log \left[p(\vec{x})\right] d\vec{x} \\ &= -E_x\left[\log p(\vec{x})\right] . \end{aligned} \tag{148}$$

The entropy measures the spread of the distribution. For example, for a univariate normal distribution

$$p(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2}(x^2/\sigma^2)} , \tag{149}$$

the entropy is

$$H(\sigma) = \frac{1}{2}\left[1 + \log\left(\sqrt{2\pi}\sigma\right)\right] . \tag{150}$$

This definition is related to the intuitive notion of entropy. A distribution that is very narrow tends to constrain the random variable to local regions where p(x) is large, tending to restrain the range of values taken by x. A very broad distribution allows the random variable to take on many more different values, causing a "more random" distribution.

The mutual information between two coordinates (x_i, x_j) is defined as

$$M_{ij} = H_i + H_j - H_{ij} .$$

Here

$$H_i = -\int_R p(x_i) \log\left[p(x_i)\right] dx_i \tag{151}$$

and

$$H_{ij} = -\int_R p(x_i, x_j) \log\left[p(x_i, x_j)\right] dx_i\, dx_j$$

where $p(x_i, x_j)$ is the marginal bivariate joint probability density defined in Eq. 146a and $p(x_i)$ is the marginal univariate probability density

$$p(x_i) = \int_R p(\vec{x}) \prod_{k \neq i}^{d} dx_i \quad , \tag{152a}$$

or alternatively

$$p(x_i) = \int_R p(x_i, x_j)\, dx_j \quad . \tag{152b}$$

In all cases, R symbolizes the allowed range of values for the random variables.

The mutual information can take on values in the range

$$0 \leq M_{ij} \leq \text{minimum}\left[H_i, H_j\right] \quad . \tag{153}$$

A normalized mutual information can be defined as

$$m_{ij} = M_{ij}/\text{minimum}\left[H_i, H_j\right] \tag{154}$$

so that $0 \leq m_{ij} \leq 1$. A small value of m_{ij} indicates small pairwise dependence between x_i and x_j, while a value near its maximum indicates a large pairwise dependence. Unlike the correlation coefficient, however, the mutual information measure is sensitive to all types of relationships and not just linear ones. A value of m_{ij} consistent with zero[29] indicates pairwise independence between the coordinates x_i and x_j.

For nonparametric applications, the marginal probability densities $p(x_i)$, $p(x_j)$, and $p(x_i, x_j)$ must be estimated from the data sample. Any of the techniques described earlier may be used for that purpose. Since these are one and two-dimensional densities, these estimates do not encounter the difficulties present with higher dimensional density estimates.

6.3.2.2 An algorithm for the direct measure of stochastic independence

In this section we discuss methods for directly testing for stochastic independence.[30] For simplicity of discussion, we will consider the special case of q=2 in Eq. 144. Generalizations for arbitrary values of q are straightforward.

We wish to test the null hypothesis, H_0, that the unknown joint probability density distribution of the data can be factored into two independent probability density functions, each defined over orthogonal subspaces of the full dimensional space. That is,

$$H_0: \; p(\vec{x}) \equiv p(x_1 x_2 \ldots x_d) = p_A(x_1 x_2 \ldots x_M)\, p_B(x_{M+1} x_{M+2} \ldots x_d) \qquad (1 \leq M < d). \tag{155}$$

If a set of measurables $(x_1, x_2, \ldots, x_d)$ can be found for which such a factorization occurs, there are two important consequences. First, the nature of the particular set of measurables can give considerable insight into the dynamics of the production process. Many theories of multiparticle production either make predictions concerning the factorability of the Lorentz invariant amplitude, or need to make assumptions concerning such factorability properties in order to calculate predicted experimental results. This algorithm allows one to test directly for such factorization properties.

A second important consequence is that if the subspaces are stochastically independent then the d-dimensional problem can be separated into an M-dimensional problem and an independent (d-M)-dimensional problem. Thus, the dimensionality has been reduced with no loss of information. Since the problems in data analysis increase dramatically with increasing dimensionality, this is always a great advantage.

The algorithm compares the interrelationships between the points in one subspace to the interrelationships in the other subspace. Specifically, the identities of the k closest neighbors to each data point are found and listed separately in each of the two subspaces. For each point, these two lists are compared for coincidences. Namely, the number of data points, k_c, that the two lists have in common are counted. The number of such coincidences between the two subspaces is evaluated for each event. This number, k_c, can have values from zero to k.

If the two subspaces are stochastically independent (null hypothesis), then those coincidences that do occur will be totally accidental in nature. The probability distribution of the number of such accidentals can be shown to be a binomial distribution, namely,

$$p_N^{(0)}(k_c) = B^k_{k/N}(k_c) \quad , \tag{156}$$

with expected value

$$E[k_c] = k^2/N \quad . \tag{157}$$

A test statistic can then be formed by performing a univariate goodness-of-fit test between the experimental distribution of the k_c values, $n(k_c)$, obtained from all of the data points, and this binomial distribution. A departure of the experimental distribution, $n(k_c)$, from the binomial distribution indicates stochastic dependence between the two subspaces.

The binomial distribution result for the accidental rate is invariant to how the list of points associated with each sample point in the two subspaces was prepared. For example, one could form a list in each subspace of the k points farthest away from the sample point. Alternatively, we could compare the list of the farthest away in one subspace to the closest in the other subspace. Under the null hypothesis, the distribution of coincidences should conform to the binomial distribution (Eq. 156) for all of these cases. The power of the test to discriminate against various classes of alternate hypotheses can be improved by forming a test statistic from a combination of our goodness-of-fit tests, shown in Table 2.

Table 2

Case	Subspace A	vs.	Subspace B
1)	k-closest		k-closest
2)	k-closest		k-farthest
3)	k-farthest		k-closest
4)	k-farthest		k-farthest

It is easy to see that for purely linear relationships (correlations) the experimental value of $\bar{k}_c$ will be larger, smaller, smaller, and larger, respectively, than $E[k_c]$ (Eq. 157) for these four cases.

Any of the goodness-of-fit tests described earlier can be used for comparing each of the four $n(k_c)$ distributions to the corresponding binomial distribution (Eq. 156). Another test statistic that has proven useful is

$$Y = \sqrt{N} \sum_{i=1}^{4} \frac{\left(\bar{k}_c^{(i)} - E[k_c]\right)^2}{\sqrt{\hat{V}_i(k_c) + V(k_c)}} \tag{158a}$$

where

$$\hat{V}_i(k_c) = \frac{1}{N} \sum_{j=0}^{k} \left(k_{cj}^{(i)} - \bar{k}_c^{(i)} \right)^2 p_N^{(i)}(k_{cj}) \tag{158b}$$

is the experimental sample variance, and $V(k_c)$ is the variance of the predicted binomial distribution under H_0,

$$V(k_c) = \frac{k^2}{N} \left(1 - \frac{k}{N} \right) \quad . \tag{158c}$$

The sum in Eq. 158a is over the four distributions corresponding to the cases listed in Table 2.

As was the case for comparing two multivariate points sets, there exists for this test a permutation procedure for estimating the probability density function of the test statistic, $p_N^{(0)}(Y)$, under the null hypothesis, directly from the data. For this permutation, the identities of the data points in one of the subspaces are randomly re-assigned. The identities of the data points in the other subspace may, but need not, be given a different random re-assignment. The test for stochastic independence is applied to the re-assigned samples and a value for the test statistic obtained. Repeated application of this permutation procedure yields a series of test statistic values that closely approximate the null probability density function for the test statistic $p_N^{(0)}(Y)$. In particular, the number of permuted test statistic values greater than the value obtained from the experimental (unpermuted) test is an estimate of the significance level for the test.

The statistical properties of this test are quite similar to those for the test that compares multivariate point sets. This is not surprising since basic to both is the kth nearest neighbor technique. Specifically, the test is consistent, unbiased, and very robust. The test is somewhat less efficient than tests using correlation coefficients when there are only linear dependencies (correlations) involved. It is also slightly less efficient than the mutual information tests when there are only pairwise dependencies in the data. It is, of course, much more efficient than either when the dependence between subspaces is not linear and is more complicated than simply pairwise dependencies. Most important, unlike the simpler tests, this test provides a necessary and sufficient test for stochastic independence.

6.4 A multivariate goodness-of-fit test

This section describes an algorithm for general multivariate goodness-of-fit testing.[31] That is, given a mathematical model, $f(\vec{x})$, defined over the multidimensional data space, this algorithm tests the hypothesis that the true underlying data probability density distribution, $p(\vec{x})$, is compatible with $f(\vec{x})$,

$$H_0: \; p(\vec{x}) = f(\vec{x}) \quad .$$

In addition, this test does not require that $f(\vec{x})$ be normalized so that the calculation of

$$\int_R f(\vec{x}) \, d\vec{x}$$

is not necessary, avoiding the computational problems of multidimensional integration. Since any goodness-of-fit test can also be used for estimation, this procedure can be used to estimate the parameters of multidimensional models. Unlike the maximum likelihood and moments methods, where the computational expense is often dominated by multidimensional integrations, this procedure is computationally very fast. However, like most goodness-of-fit tests that are used for estimation, this test has generally lower efficiency than the direct parametric estimators.

Most maximum likelihood estimates involve iterating from some starting values for the parameters to a set of solution values that maximize the likelihood function. All of these iterative schemes converge much faster to a solution, the closer the parameter starting values are to the solution values. Because it is computationally very fast, the solution values from this algorithm can be used as the starting point for a likelihood maximizer. Since this starting point will generally be very close to the maximum likelihood solution, considerable computation will be saved. Also, if at its solution this algorithm indicates a very poor goodness-of-fit, the experimenter may wish to avoid the likelihood estimate altogether since this lack-of-fit indicates that the parametric assumptions upon which the maximum likelihood technique is based are not valid (i.e., the model doesn't fit).

As discussed earlier, univariate goodness-of-fit tests are constructed by forming a dissimilarity measure between the density as predicted by the model, $f(x)$, and a nonparametric estimate of the density, $\hat{p}_N(x)$, from the data. Such a procedure is not possible in the general multivariate case because of the difficulty (discussed earlier) in nonparametric multidimensional density

estimation. This algorithm, like the multivariate algorithms discussed above, achieves its success by avoiding direct density estimation.

The procedure begins by finding the k nearest neighbors to each point in the data sample. Let $S_k(\vec{x})$ be the region in the d-dimensional space containing the k nearest neighbors to a point at $\vec{x}$, and let $V_k(\vec{x})$ be the volume of this region. Consider the quantity

$$v_k(\vec{x}) = \int_{S_k(\vec{x})} \frac{1}{f(\vec{x}')}\, p(\vec{x}')\, d\vec{x}' \tag{159}$$

where $f(\vec{x})$ is the model and $p(\vec{x})$ is the true probability density function of the data. Under the null hypothesis, H_0, $p(\vec{x})/f(\vec{x})$ is a constant, C, so that

$$v_k^{(0)}(\vec{x}) = \int_{S_k(x)} C\, d\vec{x}' = CV_k(\vec{x}) \quad . \tag{160}$$

The integral in Eq. 159 can be estimated by

$$\hat{v}_k = \frac{1}{k+1} \sum_{i=0}^{k} \frac{1}{f(\vec{x}_i)} \tag{161}$$

where the summation is over the data point located at $\vec{x}$ (i=0) and the k-nearest neighbors to $\vec{x}$ (i=1, k). Under H_0, the quantity $\hat{v}_k$ should be proportional to V_k and the univariate probability density of $\hat{v}_k$, $p_N(\hat{v}_k)$, should be compatible with the univariate probability density of V_k, $p_N(V_k)$.

For a model independent estimate of $V_k(\vec{x})$, $\hat{V}_k$, one can take the smallest spherical volume centered at $\vec{x}$ containing the k nearest points to $\vec{x}$. The two univariate probability densities $p_N(\hat{v}_k)$ and $p_N(\hat{V}_k)$ can then be compared using a standard univariate goodness-of-fit test. Thus, a multivariate goodness-of-fit test has again been reduced to a univariate goodness-of-fit test.

When the null hypothesis is not true $f(\vec{x}) \neq p(\vec{x})$, then $v_k(\vec{x})$ (Eq. 159) will not be proportional to the volume $V_k(\vec{x})$ and its variation will take on a different shape, giving rise to a different univariate probability density for $\hat{v}_k$, $p_N(\hat{v}_k)$. This will result in a bad correspondence between $p_N(\hat{v}_k)$ and $p_N(\hat{V}_K)$ in the univariate goodness-of-fit test.

The test therefore consists of finding the k nearest neighbors to each data point in the full dimensionality. The volume, $\hat{V}_k$, of a d-dimensional sphere whose radius is the distance from the point to its kth closest neighbor is calculated. The quantity $\hat{v}_k$ (Eq. 161) is also calculated for the data point and its

k closest neighbors. The test statistic for this multivariate goodness-of-fit test is then taken to be a univariate goodness-of-fit test statistic between the distribution of $\hat{V}_k$ and $\hat{v}_k$ over all of the data points.

The fact that the shape of the volume for $\hat{V}_k$ is taken to be spherical is mainly for calculational convenience and is not essential. Any volume may be used that contains the k closest points and no others. The spherical volume works well except when the data space has boundaries that are important. That is, the data density is high near a boundary. In this case, the spherical shape will severely bias the volume estimate towards values that are much too large since a considerable fraction of the sphere will lie outside the allowed region for the data. A solution would be to use only that volume of the sphere that lay inside the allowed data space. This, however, requires a detailed knowledge of the shape of the boundaries and, except in the simplest cases, considerable computation.

Quite often, one would like to do a goodness-of-fit test or estimate parameters without being required to supply information concerning the details of the boundaries of the data space. Note that both the maximum likelihood and moments estimators require such information since these boundaries form the region of integration for the multidimensional normalization integrals.

If the density of data points tends to be small near the boundaries, then the algorithm described above, using spherical volumes, is adequate since the effect of the bias near the boundaries will not be severe. However, if this is not the case, a different shape for the volume of the k nearest neighbors is required.

A volume shape that always just contains the k nearest neighbors and never exceeds the data boundaries, no matter what their shape (so long as its convex), can be conveniently calculated for the special case where k=d. In this case, the d+1 points (data point plus its d closest neighbors) can be considered to be the vertices of a d-dimensional simplex. A simplex is the simplest geometrical solid for a given dimensionality (i.e., a triangle for d=2, a tetrahedron for d=3, etc.). This simplex is, in fact, the smallest nonconcave volume that

contains these data points and no others. The volume of such a simplex is

$$\hat{V}_k = \text{determinant} \begin{bmatrix} x_1^{(1)} x_1^{(2)} \cdots x_1^{(d)} & 1 \\ x_2^{(1)} x_2^{(2)} \cdots x_2^{(d)} & 1 \\ \cdot \quad \cdot \qquad \cdot & \cdot \\ \cdot \quad \cdot \qquad \cdot & \cdot \\ \cdot \quad \cdot \qquad \cdot & \cdot \\ x_{d+1}^{(1)} x_{d+1}^{(2)} \cdots x_{d+1}^{(d)} & 1 \end{bmatrix} \tag{162}$$

where $x_i^{(j)}$ is the jth coordinate of the ith vertex point.

The simplex volume shape has several disadvantages. First, the number of nearest neighbors is constrained to be equal to the dimensionality. This is a disadvantage for very low dimensionality. However, in low dimensionality the boundary effects are considerably less severe than for high dimensionality, reducing the bias from spherical volumes. Second, the variance of the volume estimates is larger with the simplex than with the spherical volume. Also, the distribution of the test statistic, $p_N^{(0)}(Y)$, becomes more dependent on the under lying data density, $p(\vec{x})$, and in some cases becomes badly biased when using the simplex volume. For these reasons, spherical volumes should always be used unless boundary effects are important.

When used as a goodness-of-fit test, this algorithm is somewhat less efficient than comparing Monte Carlo events generated from the model to the data events, using the procedure for comparing multivariate point sets. Also, this algorithm does not provide as much information concerning those regions of the multidimensional space where the fit is good and where it is bad. These limitations are due to the fact that this algorithm does not require $f(\vec{x})$ to be normalized. The loss of this information, as well as information concerning the boundaries of the data space, causes the reduction in efficiency. It also results in a great increase in computational economy. In order to generate Monte Carlo events from a model, the data space boundaries and the normalization must be either explicitly or implicitly determined. Also, such Monte Carlo's are usually computationally very expensive. This algorithm trades a loss in statistical efficiency for a great gain in computational efficiency.

As for estimation, this algorithm can form the first step in a two-step procedure for multivariate goodness-of-fit. First, this computationally fast

procedure is applied. If the result is a poor goodness-of-fit, then no further processing is necessary. If this test shows a marginal or good goodness-of-fit, then the more expensive procedure of generating Monte Carlo events from the model and comparing point sets can be applied.

Another limitation with this test associated with the lack of normalization information is that, in general, the test is not consistent. The procedure tests for goodness-of-fit of the model to the data only in regions where there are data points. It is possible that the model fits the data well in regions where there are data, but predicts large data densities in regions where there is no actual data. The test is completely insensitive to this case. The converse is not true. The test is very sensitive to the case where there is data in regions where the model predicts small or zero densities. Although it is unlikely that the model will fit the data well where the data points exist, and still not be correct, it should be kept in mind that this situation is possible. This situation is easily detected when the Monte Carlo points are compared to the data points.

Another disadvantage with this test, as compared to the Monte Carlo generation method, is that there is no analog of the permutation procedure for estimating the null distribution of the test statistic, $p_N^{(0)}(Y)$, directly from the data. This null distribution is reasonably (but not completely) distribution free. It is usually sufficient to determine the test statistic distribution with a Monte Carlo procedure once and for all using a model that allows quick and easy Monte Carlo generation, and is at least a crude approximation to those models being tested. Since the test statistic is nearly distribution free, the null distribution obtained in this manner will serve as a good approximation for most applications.

Both this algorithm and the one that compares point distributions leave to the researcher's discretion the choice of the coordinate variables and metric, and the number of nearest neighbors, k. These algorithms are reasonably insensitive to the choice of k, provided that it is not too small. In order for the tests to be consistent, k should be a function of the total sample size such that

$$\lim_{N \to \infty} k(N) = \infty \quad , \qquad \text{and} \qquad \lim_{N \to \infty} \frac{k(N)}{N} = 0 \quad .$$

Experimentation has shown that the choice of k is not important so long as $k > 10\text{-}20$. Clearly, k should be small compared to the total sample size, N.

These algorithms are somewhat more sensitive to choice of measurement variables and metric. Unfortunately, there are no good guidelines for their choice. For infinite sample size these algorithms are clearly invariant to

changes in coordinate variables and metric since these changes simply alter the shape of the volume element containing the evaluation point. Since these volumes are infinitesimally small, their shape doesn't matter.

For finite sample sizes, however, the shape does matter. Changes in the volume shapes that result in changes of the identities of the nearest neighbors will have an effect on the performance of the algorithms. Fukunaga and Hostetler[8] show that for those data distributions that can be made spherically symmetric by a linear transformation, the optimum metric is the inverse covariance matrix of the underlying distribution, $p(\vec{x})$. If this covariance matrix is estimated by the data sample covariance matrix, then this is equivalent to scaling each of the coordinates so that they have equal variance along the principal axes of the data.

If one has no a priori information concerning the data, then this is probably the best procedure. Another reasonable procedure is to simply scale the data to have equal variance along the original measurement coordinates. On the other hand, different experimental measurement accuracy or different characteristic length of density variation can dictate unequal scales among the various coordinates. Changing the scale of a coordinate changes its relative importance in determining the goodness-of-fit. Thus, if the researcher has information as to which coordinates are most important, they should be given larger scales.

The number and specific choices of coordinate variables also affect the performance of these algorithms. Increasing the number of coordinates only improves the performance when those variables contain information concerning the hypothesis under test. In fact, coordinates that do not contain such information (noise coordinates) dilute the power of the tests. This is because these dimensions add statistical variance to the volume estimates without providing information helpful to the estimation. Even a coordinate that does contain some additional information may not help because the increase in statistical variance that it introduces hurts more than the information increase helps. The precision of these tests can be increased greatly if the researcher's knowledge and intuition lead him to a judicious choice of coordinate variables.

For goodness-of-fit testing, the model itself can be used to help choose optimum coordinate variables. Clearly those coordinates that enter directly into the model are the ones that will tend to have the most bearing on the problem. However, there may be strong dependencies in the data, not predicted by the model. In this case, choosing only coordinate variables that appear directly

in the model will dilute the power of the goodness-of-fit to discriminate against the model.

There are other considerations that affect the best choice for coordinate variables. For the algorithm discussed in this section only, the nature of the data space boundaries is very important. For some variables, the data densities approach zero at the boundaries or there are no boundaries at all. For example, since angular variables are simply periodic, they have no boundaries or unallowed regions. Also, in multiparticle production, the limited transverse momenta plus energy conservation usually prevent large data population near kinematic boundaries. On the other hand, t-channel variables such as four-momentum transfers squared usually have the highest densities near some of their boundaries.

Choosing good coordinate variables and a good metric usually requires a compromise among all of these considerations. Intelligent choices based on experience and intuition can substantially improve the performance of the algorithms, provided that these are correct. The researcher always has the option, of course, of applying the algorithms with many different choices and, therefore, empirically determining which are the best.

ACKNOWLEDGMENTS

Helpful discussions with William H. Rogers, Sam Steppel and John W. Tukey are gratefully acknowledged. I would also like to acknowledge L. Van Hove whose comment[32] "Speaking generally, one can only regret that some people are still content to analyze data at a level of superficiality which practically guarantees in advance that no really new and instructive conclusions will emerge," formed the motivation for a substantial portion of this work.

FOOTNOTES AND REFERENCES

1) See Reference 2 for an excellent and relatively complete discussion of the foundations of the statistical techniques used in high energy particle physics.

2) W.T. Eadie, D. Drijard, F.E. James, M. Roos and B. Sadoulet, Statistical Methods in Experimental Physics, North-Holland, Amsterdam-London, 1971.

3) The smallest possible variance for an estimate is called the minimum variance bound and is related to the information content of the estimator by the Cramer-Rao inequality (see Reference 2, Section 7.4, pp 130).

4) If the total number of counts, N, is itself considered a random variable, then the distribution of counts in a histogram bin is Poisson

$$p(\hat{n}_1, \hat{n}_2, \ldots, \hat{n}_M) = \prod_{i=1}^{M} e^{-\bar{n}_i} \frac{\bar{n}_i^{\hat{n}_i}}{\hat{n}_i!}$$

with $N = \sum_{i=1}^{M} \hat{n}_i$, for which $E[\hat{n}_i] = \bar{n}_i$ and $V[\hat{n}_i] = \bar{n}_i$.

5) M. Rosenblatt, "Remarks on some nonparametric estimates of a density function," Ann. Math. Statist., 27, pp 832-837 (1956).

6) E. Parzen, "On the estimation of a probability density function and the mode," Ann. Math. Statist., 33, pp 1065-1976 (1962).

7) E. Fix and J. L. Hodges, Jr., "Nonparametric discrimination: consistency properties," Report Number 4, Project Number 21-49-004, USAF School of Aviation Medicine, Randolf Field, Texas, Feb. 1951.

8) K. Fukunaga and L. D. Hostetler, "Optimization of k-nearest-neighbor density estimates," IEEE Trans. Info. Theory, Vol. IT-19, pp 320-326 (1973).

9) D. O. Loftsgaarden and C. P. Quesenberry, "A nonparametric density function," Ann. Math. Statist., Vol. 36, pp 1049-1051 (1965).

10) J. W. Tukey, Exploratory Data Analysis, Vol. I, Chapter 7, Addison-Wesley, Reading, Mass., 1970 (limited preliminary edition).

11) A. E. Beaton and J. W. Tukey, "The fitting of power series, meaning polynomials, illustrated on bandspectroscopic data," Technometrics, May 1974.

12) See Reference 2, Chapter 7 for derivations of these results.

13) J. W. Tukey, Exploratory Data Analysis, Vol. III, Chapter 24, Addison-Wesley, Reading, Mass., 1971 (limited preliminary edition).

14) Reference 2, pp 268-269.

15) Reference 2, pp 269-271.

16) Reference 13, Chapter 26.

17) M. A. Fisherkeller, J. H. Friedman, and J. W. Tukey, "PRIM-9: An interactive multidimensional data display and analysis system," Proceedings of the Second Annual AEC Scientific Computer Information Exchange Meeting, pp 3-33, May 2-3, 1974, also Stanford Linear Accelerator Center report, SLAC-PUB-1408, April 1974.

18) M. F. Hodous and I. A. Pless, "Advanced Graphical displays used in the analysis of high energy physics data," Proceedings of the Second Annual AEC Scientific Computer Information Exchange Meeting, p 59, May 2-3, 1974.

19) J. H. Friedman and J. W. Tukey, "A projection pursuit algorithm for exploratory data analysis," IEEE Trans. Computers, Vol. C-23, pp 881-890 (1974).

20) J. W. Sammon, Jr., "A nonlinear mapping for data structure analysis," IEEE Trans. Computers, Vol. C-18, pp 401-409 (1969).

21) R. N. Shepard and J. D. Carroll, "Parametric representation of nonlinear data structures," in Multivariate Analysis, P. Krishnaiah, Ed. New York: Academic Press (1966).

22) J. E. Brau, F. T. Dao, M. F. Hodous, I. A. Pless, and R. A. Singer, Phys. Rev. Letters, 27, p. 1481 (1971).

23) P. E. Condon and P. L. Cowell, "Channel Likelihood: an extension of maximum likelihood to multibody final states," Phys. Rev. D, Vol. 9, pp 2558-2562 (1974).

24) L. Van Hove, "The extended prism plot technique." Presented at the Topical Conference on Multidimensional Analysis of Hadron Collisions, CERN, Geneva, 11-14 February (1974).

25) Substituting $\alpha_m = N_m/I_m$ with

$$I_m = \int_R f_m(\vec{x})\, d\vec{x}$$

into Eqs. 139 and 141, one has

$$N_m = \frac{N_m}{I_m} \sum_{i=1}^{N} \frac{f_m(\vec{x}_i)}{\sum_{n=1}^{M} N_n f_n(\vec{x}_i)/I_n} \qquad (m=1, N) \ .$$

Condon and Cowell (Ref. 23) show that the values of $\{N_m\}_{m=1}^{N}$ that solve this set of simultaneous equations are equivalent to the maximum likelihood solutions with Eq. 137a as the model.

26) J. H. Friedman, S. Steppel, and J. W. Tukey, "A nonparametric procedure for comparing multivariate point sets," Stanford Linear Accelerator Center, Computation Group Technical Memo No. 153, November 1973.

27) J. H. Friedman, "Measurement of multivariate scaling and factorization in exclusive multiparticle production," Phys. Rev. D, Vol. 9, pp 3053-3059 (1974).

28) P. M. Lewis, IEEE Trans. Info. Theory, IT-8, pp 171-178 (1962).

29) The minimum value for the mutual information estimate expected for the null hypothesis depends upon the sample size and approaches zero asymptotically.

30) J. H. Friedman, "A general test for stoachistic independence," Stanford Linear Accelerator Center, Computation Group Technical Memo No. 154, November 1973.

31) J. H. Friedman and S. Steppel, "A general multivariate goodness-of-fit test," Stanford Linear Accelerator Center, Computation Group Technical Memo No. 159, July 1974.

32) L. Van Hove, "Particle production in high energy hadron collisions," Physics Reports, Vol. 1, pp 374 (1971).

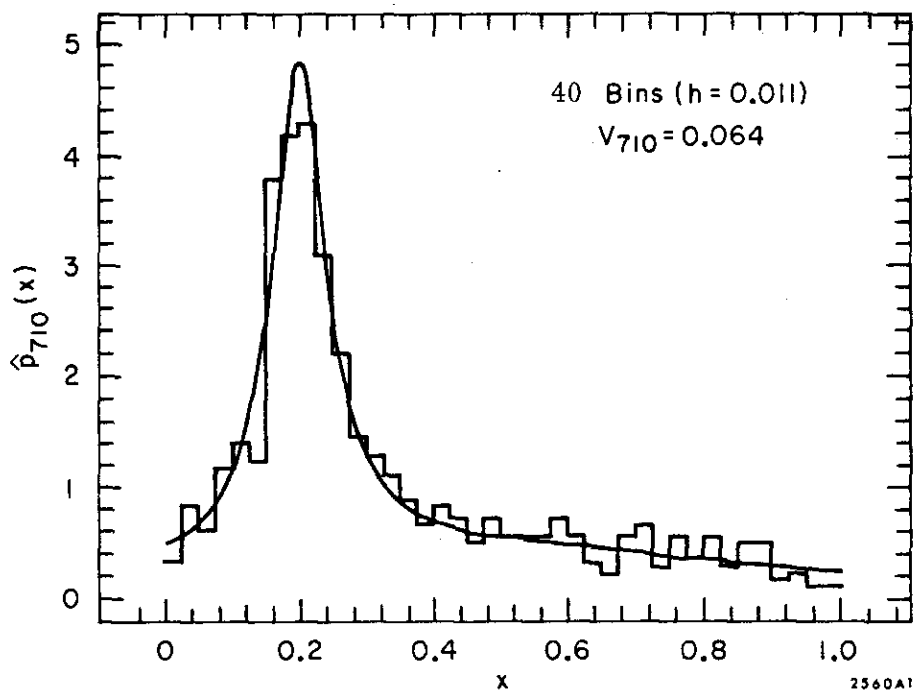

FIG. 1a Histogram density estimate.

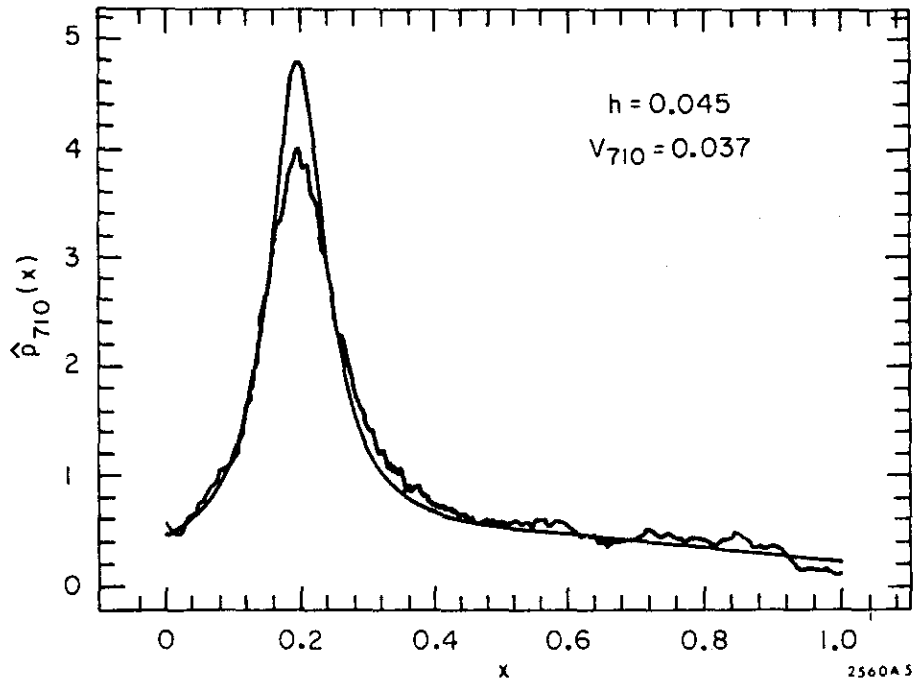

FIG. 1b Rosenblatt density estimate.

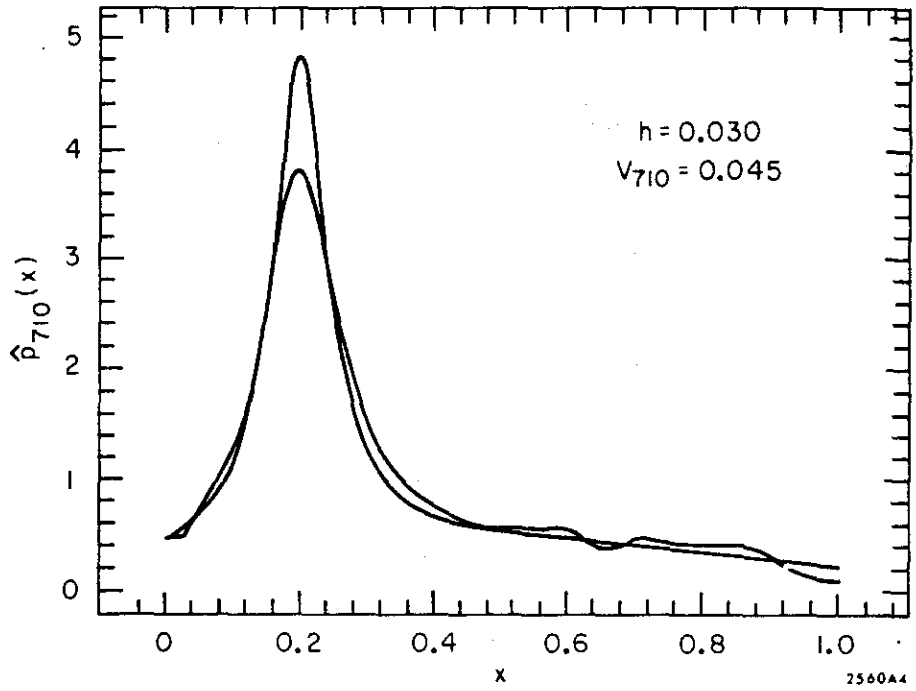

FIG. 1c Parzen (normal kernel) density estimate.

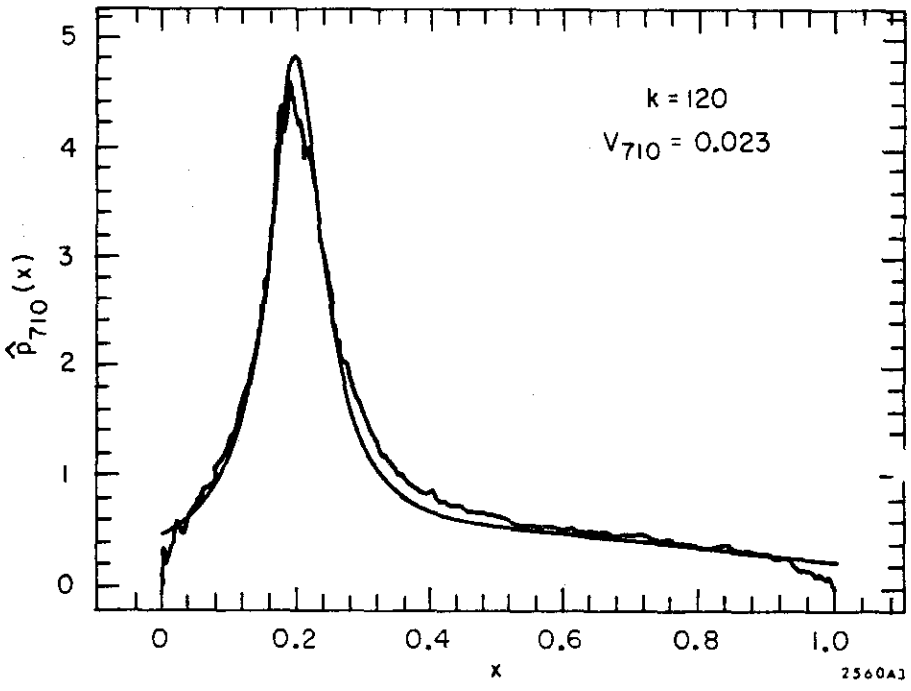

FIG. 1d k-th nearest neighbor density estimate.

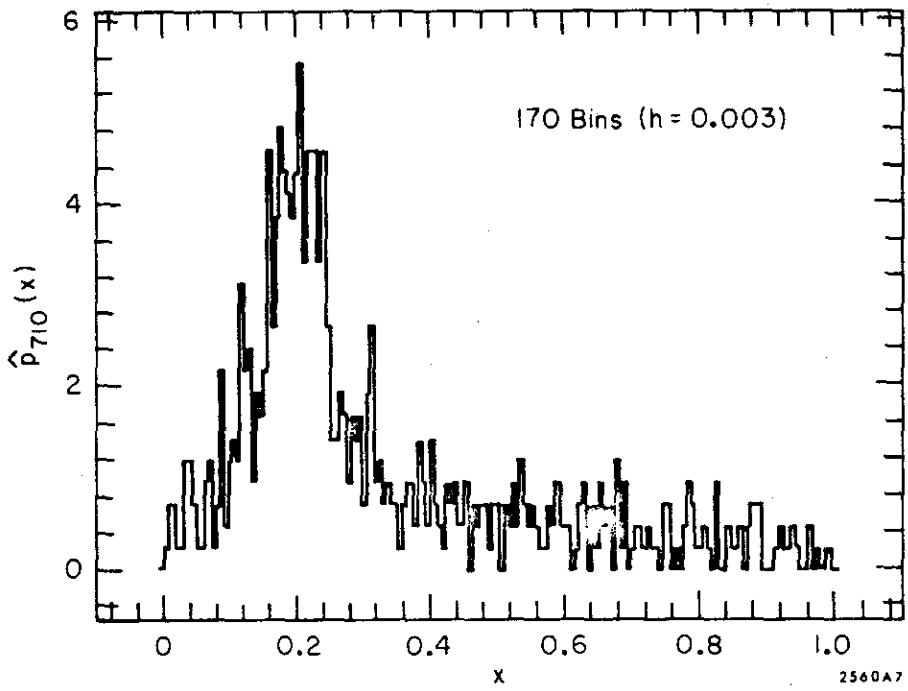

FIG. 2a Histogram density estimate.

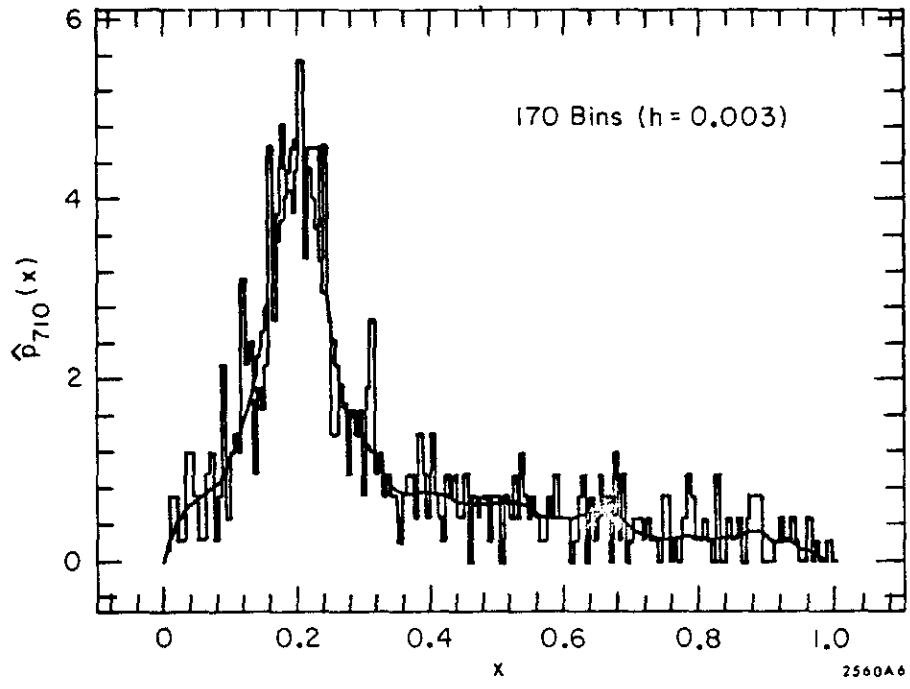

FIG. 2b Histogram density estimate (smooth superimposed).

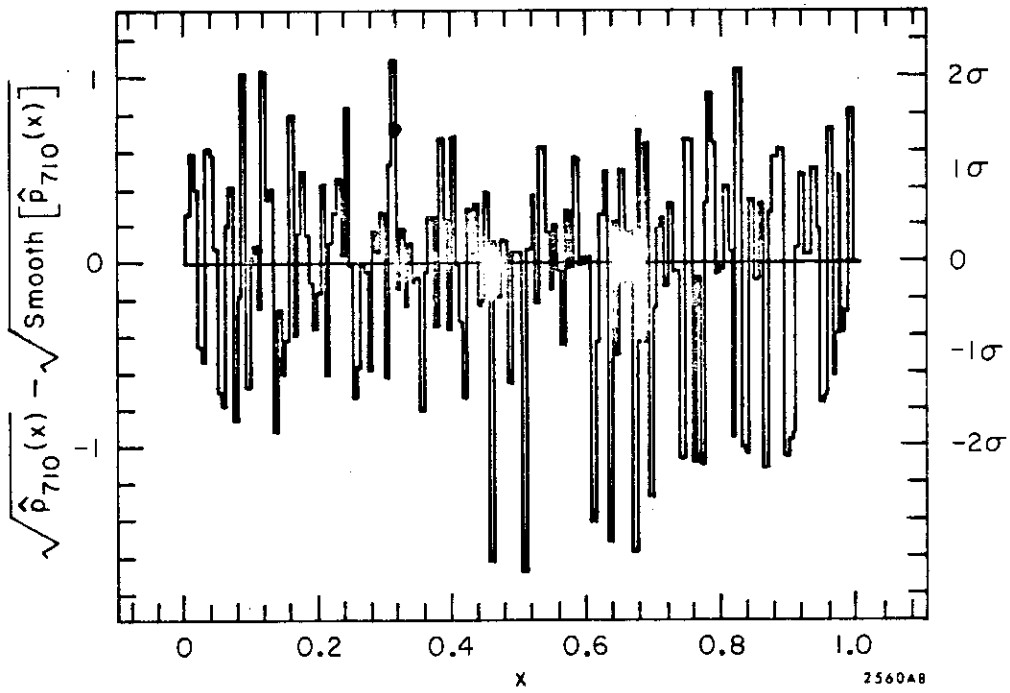

FIG. 2c Residuals between rootogram and smooth of rootogram.

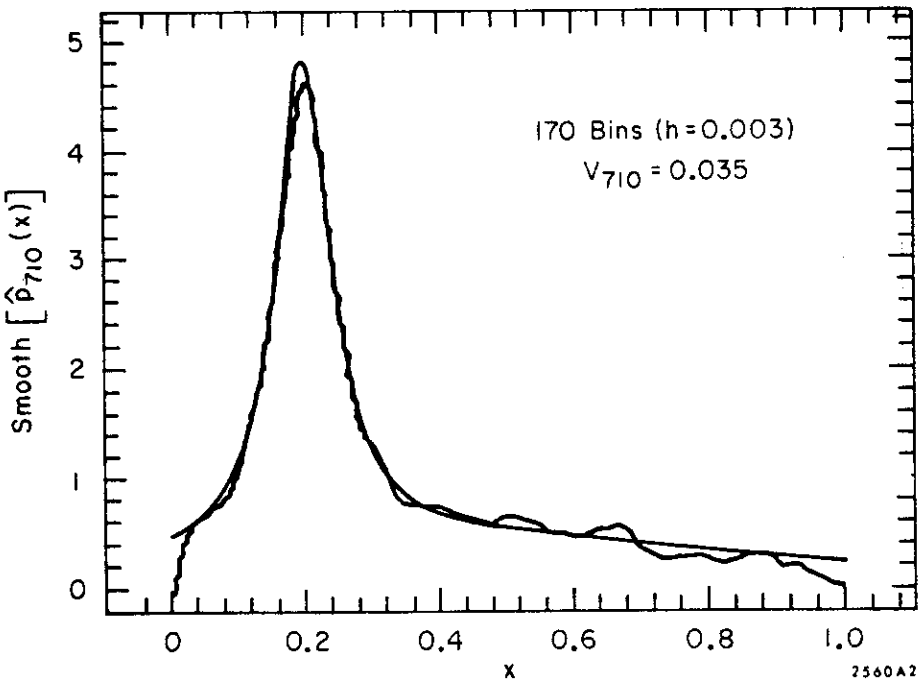

FIG. 2d Comparison of histogram smooth to the data density.

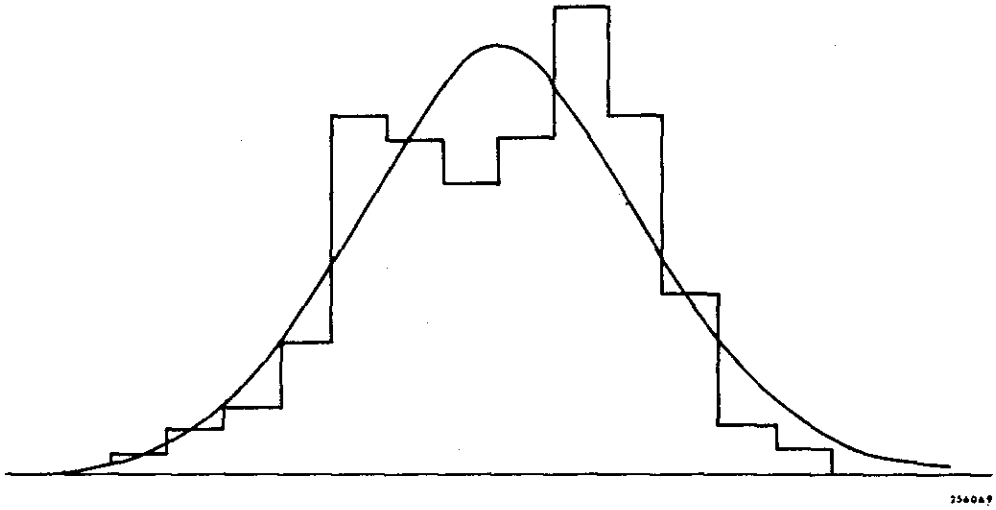

FIG. 3a Traditional histogram representation.

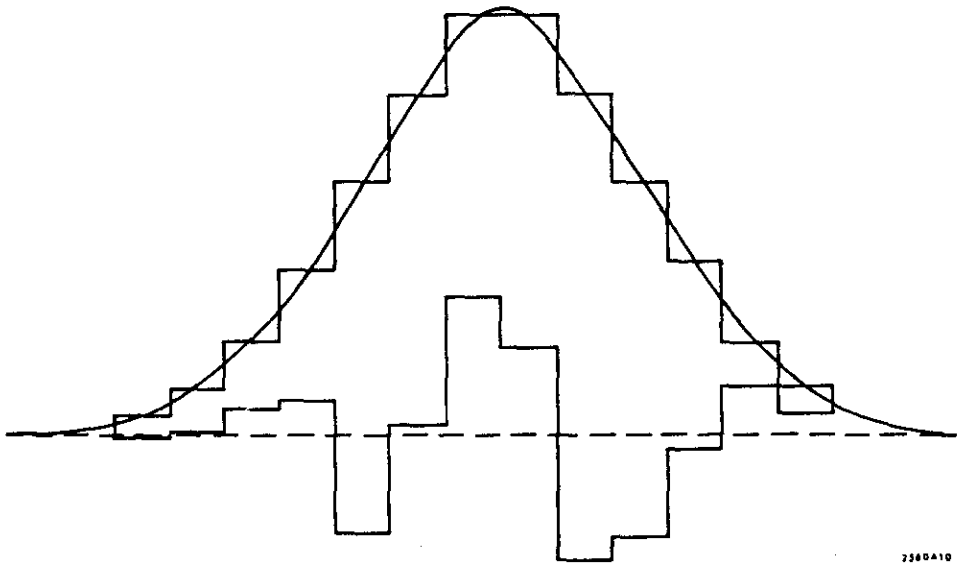

FIG. 3b Histogram aligned with comparison curve.

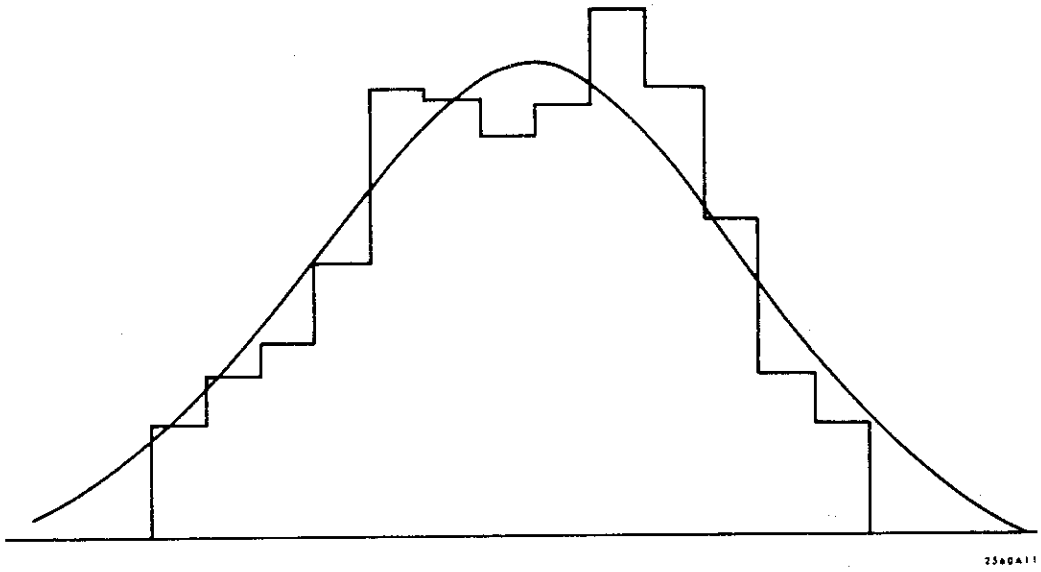

FIG. 3c Standard rootogram representation.

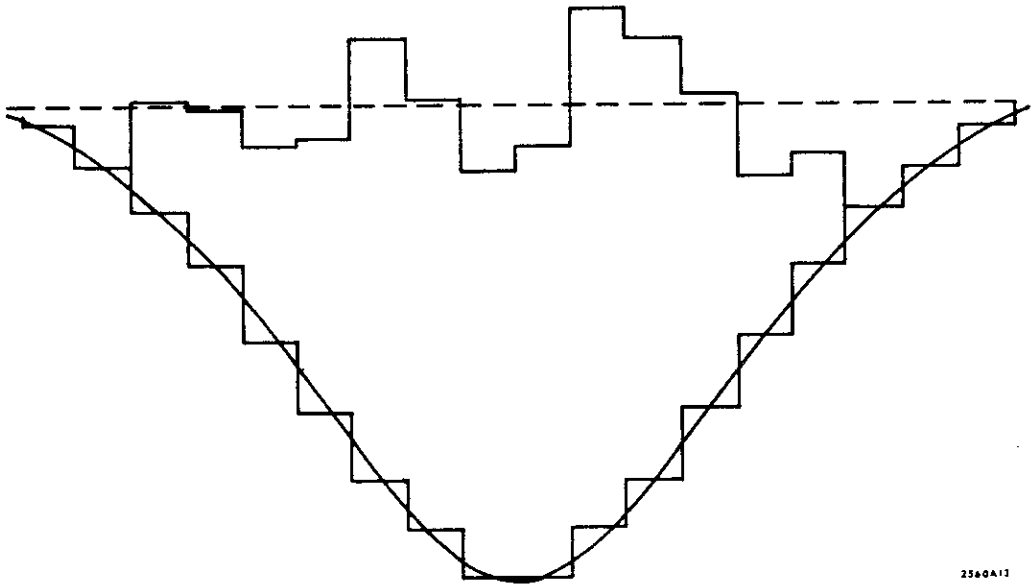

FIG. 3d Hanging rootogram.

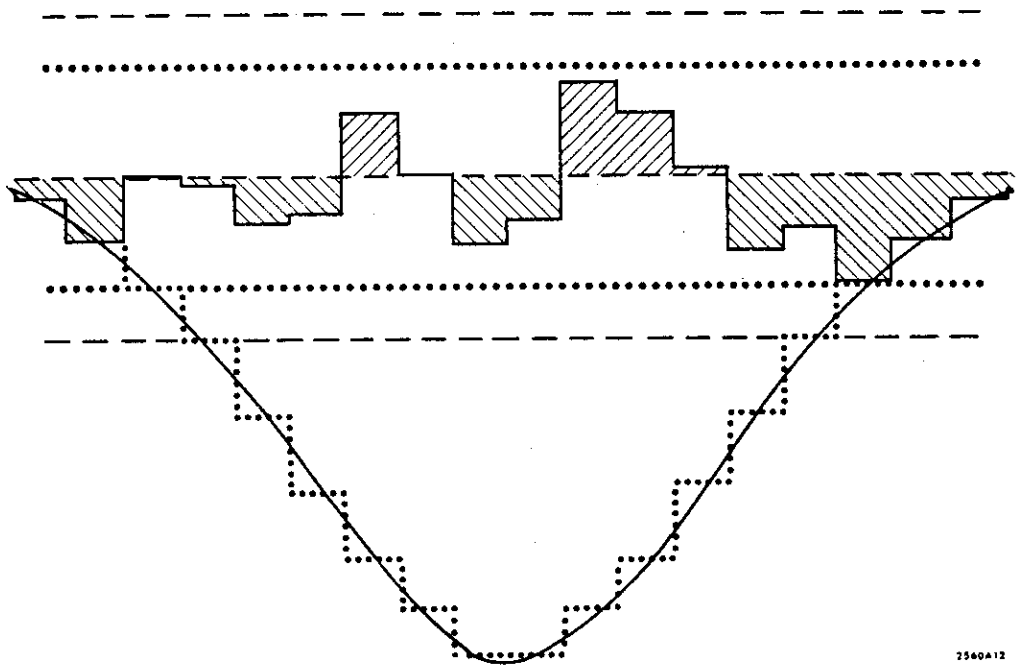

FIG. 3e Hanging rootogram with residuals emphasized.

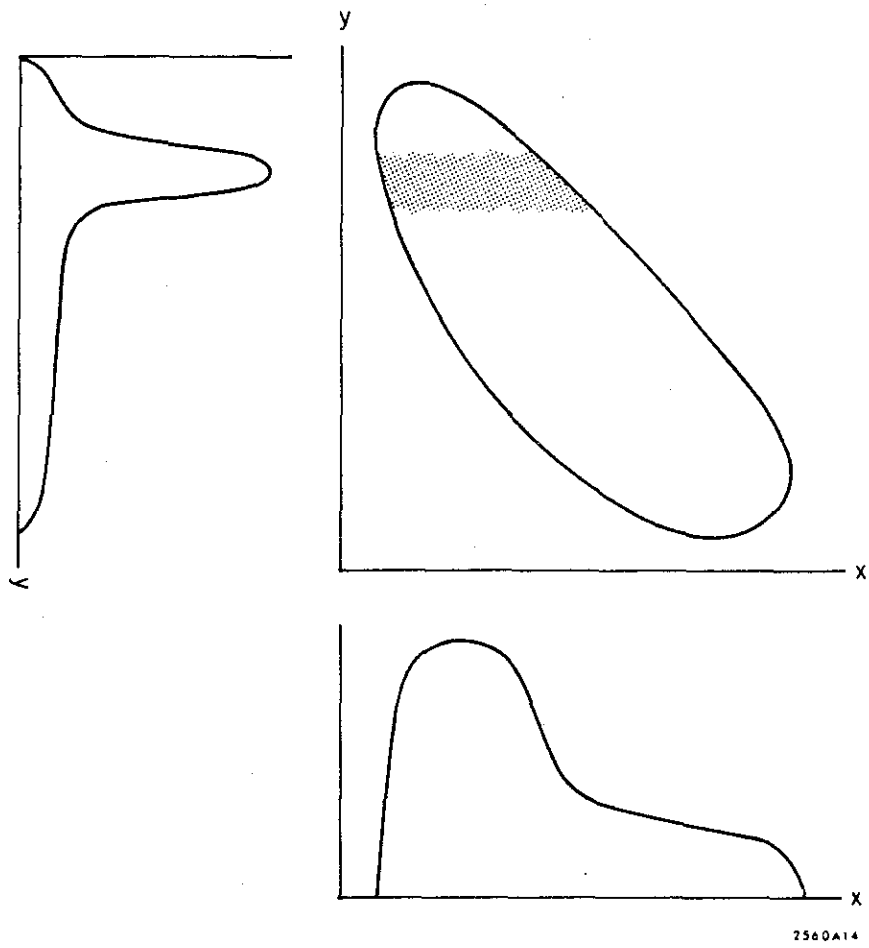

FIG. 4 Example of reflections caused by non-rectangular boundaries.

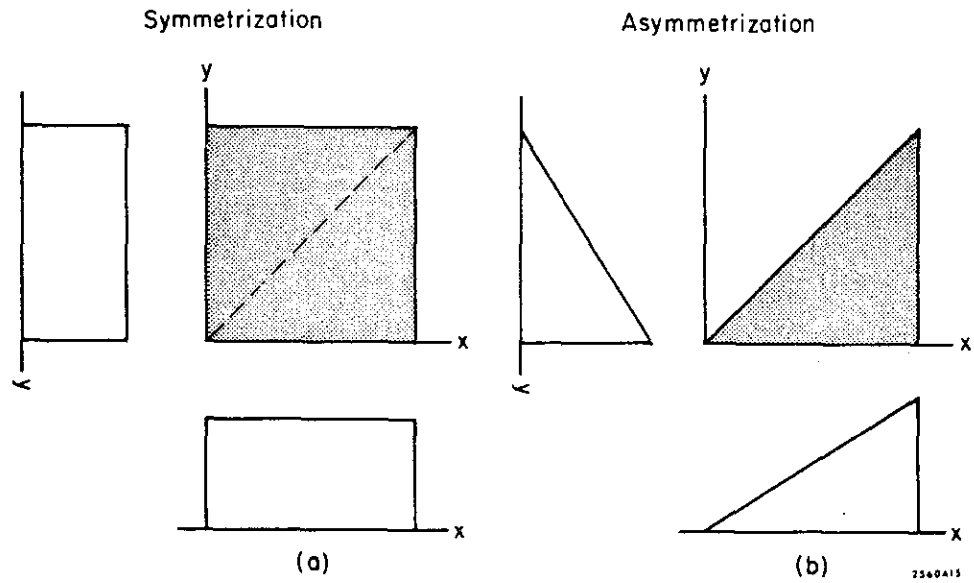

FIG. 5

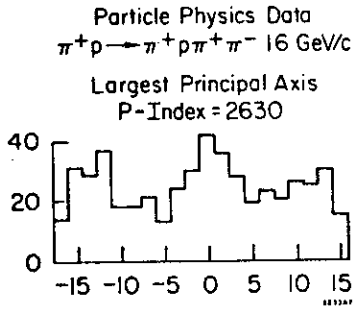

FIG. 6a

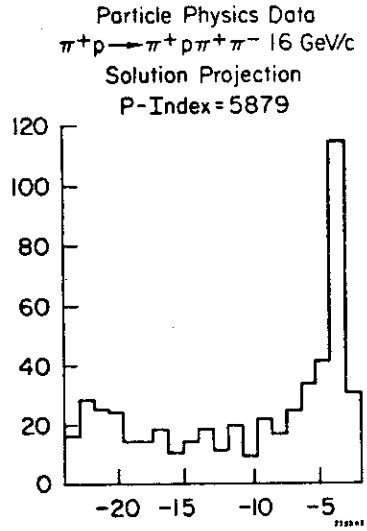

FIG. 6b

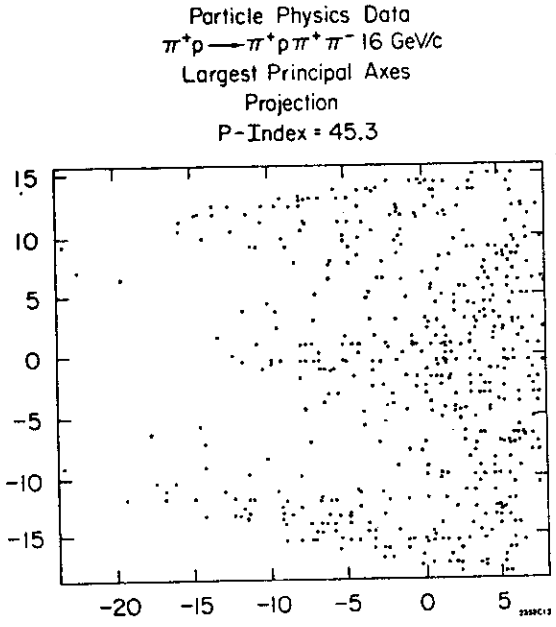

FIG. 6c

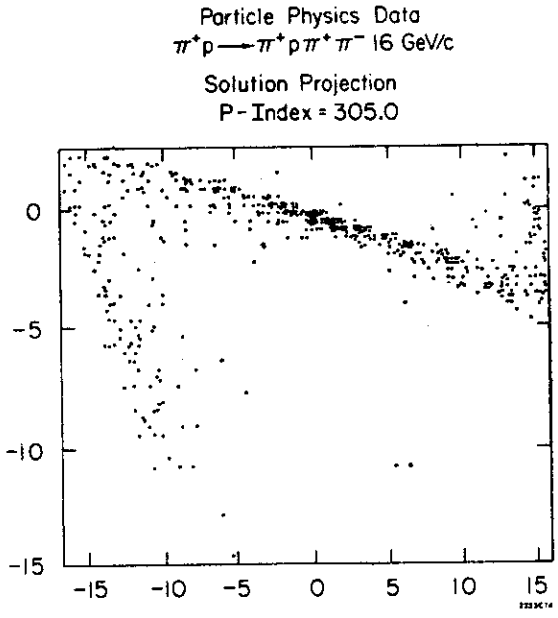

FIG. 6d

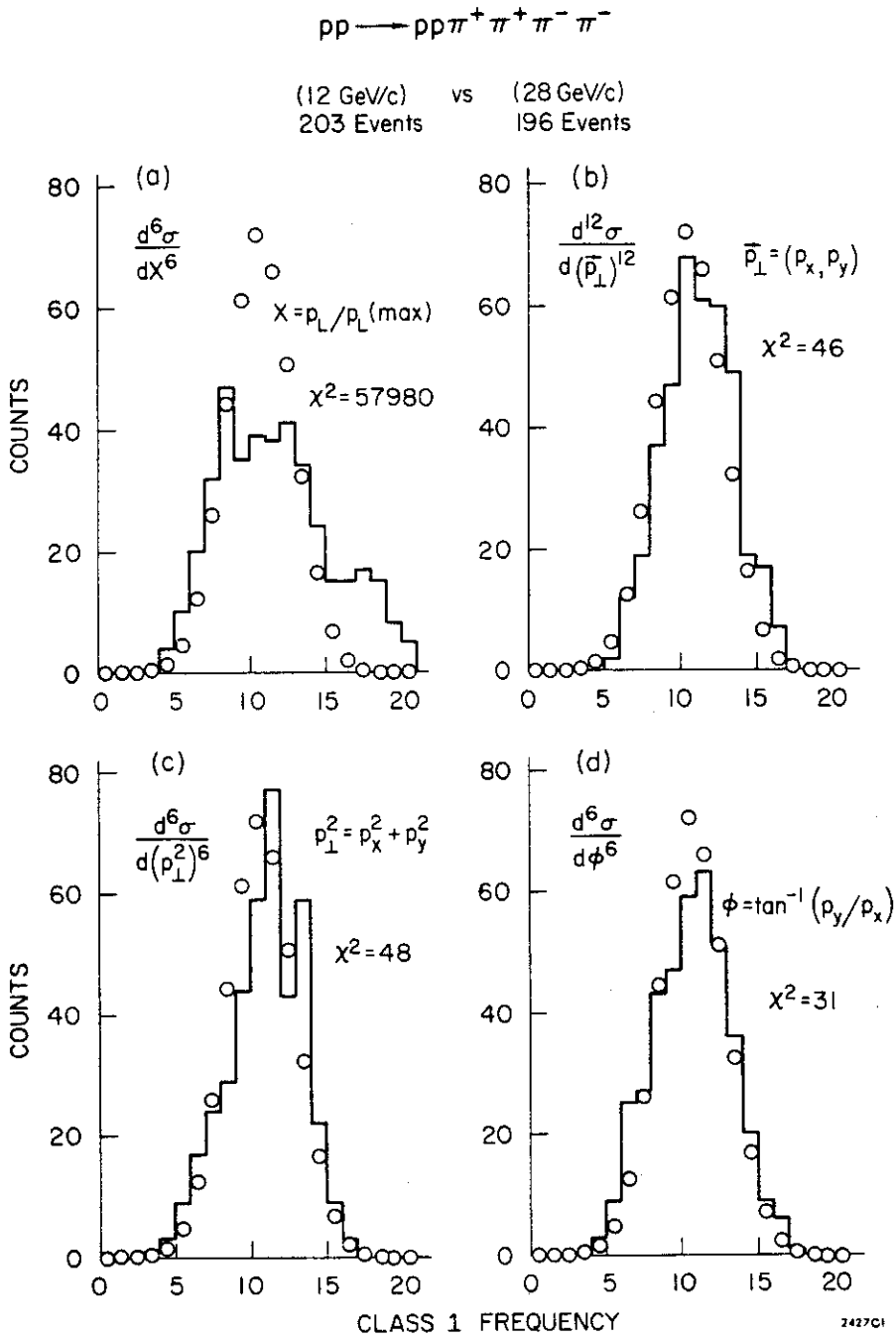
pp ⟶ pp π⁺ π⁺ π⁻ π⁻
(12 GeV/c) 203 Events
vs
(28 GeV/c) 196 Events
(a)
d⁶σ/dX⁶
X = p_L/p_L(max)
χ² = 57980
(b)
d¹²σ/d(p_⊥)¹²
p_⊥ = (p_x, p_y)
χ² = 46
(c)
d⁶σ/d(p_⊥²)⁶
p_⊥² = p_x² + p_y²
χ² = 48
(d)
d⁶σ/dφ⁶
φ = tan⁻¹(p_y/p_x)
χ² = 31
COUNTS
COUNTS
0
20
40
60
80
0
5
10
15
20
CLASS 1 FREQUENCY
2427C1

FIG. 7

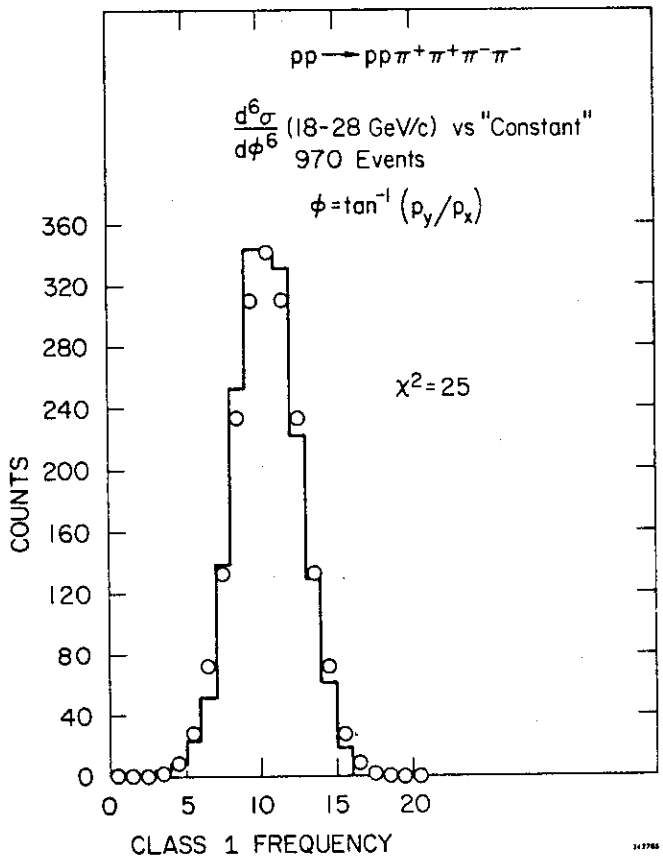
$pp \rightarrow pp\pi^+\pi^+\pi^-\pi^-$
$\frac{d^6\sigma}{d\phi^6}$ (18-28 GeV/c) vs "Constant"
970 Events
$\phi = \tan^{-1}(p_y/p_x)$
$\chi^2 = 25$
COUNTS
360
320
280
240
200
160
120
80
40
0
0
5
10
15
20
CLASS 1 FREQUENCY

FIG. 8