AN APPROACH TO SUPEREFFICIENT ESTIMATION

NEIL H. WILLITS

TECHNICAL REPORT NO. 45
FEBRUARY 1981

PREPARED UNDER THE AUSPICES OF
SIAM INSTITUTE FOR MATHEMATICS AND SOCIETY

DEPARTMENT OF STATISTICS
STANFORD UNIVERSITY
STANFORD, CALIFORNIA
AN APPROACH TO SUPEREFFICIENT ESTIMATION

by

NEIL H. WILLITS
STANFORD UNIVERSITY

TECHNICAL REPORT NO. 45
FEBRUARY 1981

STUDY ON STATISTICS AND ENVIRONMENTAL FACTORS IN HEALTH (SIMS)

PREPARED UNDER SUPPORT TO SIMS FROM

ENVIRONMENTAL PROTECTION AGENCY (EPA)
DEPARTMENT OF ENERGY (DOE)
SLOAN FOUNDATION
NATIONAL SCIENCE FOUNDATION (NSF)

DEPARTMENT OF STATISTICS
STANFORD UNIVERSITY
STANFORD, CALIFORNIA
Acknowledgements

There are a number of people whose help proved indispensable in the preparation of this paper. First and foremost, I would like to thank my adviser, Tom Sager, who provided guidance in the development of the results and a keen eye in the editing of the manuscript. This assistance was all the greater, due to the fact that it was provided largely while I was at Stanford and he was at the University of Texas. Therefore, I would like to thank the United States Postal Service for transporting numerous rough drafts of the manuscript back and forth between California and Texas. I would also like to thank Paul Switzer and Tom Cover, who served on my reading committee, providing their personal insights on the text and whether it should be lengthened, shortened, or both. Special thanks are due to Paul, who also served as my foster adviser while Tom Sager was away from Stanford.

Finally, I would like to thank my secretary, Alice Lundin, who, although she didn’t type the final manuscript, did keep my other responsibilities at Stanford running smoothly while this paper was being prepared, and Keaven Anderson, who assisted me in the use of \TeX, which was used for the typesetting of the final draft.
<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>Acknowledgements</td>
<td>iii</td>
</tr>
<tr>
<td>Abstract</td>
<td>v</td>
</tr>
<tr>
<td>I. Introduction</td>
<td>1</td>
</tr>
<tr>
<td>II. A Discussion of Estimation</td>
<td>3</td>
</tr>
<tr>
<td>III. A Review of Density Estimation Methods</td>
<td>11</td>
</tr>
<tr>
<td>IV. Properties of a Superefficient Estimate</td>
<td>34</td>
</tr>
<tr>
<td>V. An Application of the Estimate of Chapter IV in the Context of Density Estimation</td>
<td>55</td>
</tr>
<tr>
<td>VI. Conclusions and Possible Refinements and Extensions</td>
<td>63</td>
</tr>
<tr>
<td>VII. Bibliography</td>
<td>66</td>
</tr>
</tbody>
</table>
AN APPROACH TO SUPEREFFICIENT ESTIMATION

by

Neil H. Willits
Stanford University

ABSTRACT

In many estimation problems, the appropriateness of a certain estimate heavily depends on the truth or falsehood of assumptions made about the underlying state of nature. In particular, the estimate \( \hat{\theta}_1 \) may be a well-chosen estimate of \( \theta \) when an assumption is valid, and yet the estimate \( \hat{\theta}_2 \) will be preferable to \( \hat{\theta}_1 \) as an estimate of \( \theta \) when the assumption is invalid. An estimation scheme is developed that estimates \( \theta \) by a convex combination of the two candidate estimates and the properties of the resultant estimate with respect to convergence in expected squared norm (mean squared error), convergence in distribution, and convergence in probability are discussed. A somewhat simplified statement of those results is that the convex combination possesses the asymptotic properties of \( \hat{\theta}_1 \) whenever the assumption is valid, and it possesses the asymptotic properties of \( \hat{\theta}_2 \) otherwise.

This estimation scheme is applied in the context of density estimation, where the assumption under scrutiny is that the underlying density is a member of a prespecified parametric family.
I. Introduction

The Rule of Accuracy, which is a bylaw of Murphy’s laws, states that “When working toward the solution of a problem, it always helps if you know the answer.” This rule may be loosely applied in the context of estimation. In fact, in a number of situations, it not only helps to know the answer, but it can help if you simply have formed a guess at the answer. These situations and their corresponding methods of estimation are generally of three types. All three assume that there is some prespecified estimate upon which it is desired to improve through some modification in its application.

The first of these situations is that in which the form of the bias of the prespecified estimate is assumed and it is desired to reduce its bias. The foremost of these bias-reduction methods is the jackknife. This method will not be discussed in this paper, for reasons that are discussed in the next chapter. However, there is a good review paper by Miller [1974] that gives insight into the method and its application.

The second of these situations is that in which it is desired to reduce the loss (generally the mean squared error) of the prespecified estimate, regardless of the value that the parameter being estimated takes. One category of estimation methods that often yields such improvement are the Empirical Bayesian methods, some of which are discussed in the next chapter.

The third and last of these situations is that in which it is desired to reduce the loss of the prespecified estimate for some fixed possibilities that the underlying parameter may take while leaving the estimate with the same loss properties for other values of the underlying parameter. The first example of such an estimate is the so-called Hodges counterexample. This example deals with the case of estimation of a normal mean based on the random variable $X$ which is distributed $\mathcal{N}(\mu, \frac{1}{n})$. The Cramèr-Rao lower bound states that the variance of an unbiased estimate of $\mu$ must be greater than or equal to $\frac{1}{n}$, and consequently, if an unbiased
estimate is asymptotically normal, then its variance must be greater than or equal to $\frac{1}{n}$. However, Hodges showed that for $\alpha \in (0, 1)$, the estimate

$$
\hat{\mu} = \begin{cases} 
X, & \text{if } |X| > n^{-\frac{1}{4}}; \\
\alpha X, & \text{if } |X| \leq n^{-\frac{1}{4}}
\end{cases}
$$

is asymptotically normal with mean $\mu$ and variance $\sigma^2(\mu)/n$, where

$$
\sigma^2(\mu) = \begin{cases} 
\alpha^2, & \text{if } \mu = 0; \\
1, & \text{if } \mu \neq 0.
\end{cases}
$$

This example doesn’t provide a counterexample to the Cramèr-Rao bound, since it is a biased estimate of $\mu$, but it does lead to the question of what functions $\sigma^2(\mu)$ are possible for asymptotically normal estimates. The answer is that $\sigma^2(\mu)$ can be less than one only on a set of zero Lebesgue measure. Since the relative asymptotic efficiency of the estimate $\hat{\mu}$ is greater than 1 when $\mu$ is zero and is 1 otherwise, $\hat{\mu}$ is called superefficient.

One problem with superefficient estimates is that the set of measure zero on which the superefficiency occurs may not be of particular interest. In the Hodges counterexample, for instance, that set consists of the single point, zero. However, when the underlying parameter is a member of a much larger space than the real line, small sets of the parameter space may be of greater interest. The area that is of particular interest to me is that of density estimation, where, for example, a set corresponding to the members of a parametric family of densities may be considered a small set relative to the space $L_2$ and may be of considerable interest, compared to the general elements of $L_2$. In view of this possible area of interest, a review of some density estimation methods, both parametric and nonparametric, is given in chapter III.

In chapter IV of this paper, a specific estimate is developed which is superefficient on such interesting, though small sets. This estimate will be predicated on a very small set of assumptions, and will consequently be widely applicable.

In chapter V of this paper, the estimate of chapter IV will be applied to a real data set, that arose in conjunction with a consulting problem.
II. A discussion of Estimation

The general purpose of this paper is to devise some methods of estimation which, under certain circumstances, possess properties which are better than those of traditional methods. Naturally, terms such as certain circumstances, better, and traditional need to be made more precise. In order to shed some light on the approach this paper takes to the problem, I will first discuss some of the prominent existing methods of estimation, both classical and more recent. This discussion is not intended to present an exhaustive review of all possible methods of estimation, but simply to examine a number of widely used methods in order to determine what problems are involved in extending and/or improving on them.

The selection of a method of estimation is unavoidably connected the selection of a class of candidate estimates from which to choose and the selection of a criterion by which to judge the candidate estimates. There is no definitively correct choice for either selection, though as the field of statistics has developed, certain classes of estimates and criteria have gained in acceptance to the extent that they may be considered classical.

It was in part the lack of such a convention in the 1920's that led to the ongoing and often bitter disagreement between R. A. Fisher and Karl Pearson about the relative merits of estimation by maximum likelihood and by the method of moments. This disagreement served a useful purpose in the development of possible criteria and properties for these two families of estimates, since it spurred Fisher and Pearson, as well as later statisticians to consider in detail the properties of these estimates. The properties of maximum likelihood estimates as they relate to the problem of density estimation are discussed in chapter III of this paper.

The central properties which interested the statisticians of this period are consistency, convergence in distribution (asymptotic normality) and efficiency. To say that an estimate \( \hat{\theta} \) of \( \theta \) (that is, a sequence of estimates \( \hat{\theta}_n \)) is consistent means that, as the sample size becomes arbitrarily large, \( \hat{\theta} \) converges to \( \theta \) in some
sense. The two most common senses in which this convergence can take place are convergence in probability and almost sure convergence.

Definition: \( \hat{\theta} \) is said to converge in probability to \( \theta \), if for any \( \epsilon > 0 \),

\[
\Pr\{|\hat{\theta} - \theta| > \epsilon\} \to 0.
\]

Definition: \( \hat{\theta} = \hat{\theta}(\omega) \) is said to converge almost surely to \( \theta \) if the set of points \( \omega \) in the underlying probability space for which \( \hat{\theta}(\omega) \) converges as a sequence to \( \theta \) has probability 1.

To say that \( \hat{\theta} - \theta \) converges in distribution implies that it converges at some rate, say \( n^k \), where \( n \) is the sample size upon which \( \hat{\theta} \) is based, and that it converges to a probability distribution \( F \). \( \hat{\theta} - \theta \) is then defined to converge in distribution to \( F \), if, for all values of \( u \) at which \( F \) is continuous,

\[
\Pr\{n^{-k}(\hat{\theta} - \theta) \leq u\} \to F(u).
\]

and is written as

\[
n^{-k}(\hat{\theta} - \theta) \overset{d}{\to} F.
\]

Due to various forms of the central limit theorem, it is often the case that \( \hat{\theta} - \theta \) will converge to a normal distribution with zero mean at the rate, \( n^{-\frac{1}{2}} \). When this is the case, it makes sense to compare estimates on the basis of the variance of this limiting distribution. This comparison is facilitated by the Cramèr Rao lower bound, which under certain regularity conditions states that, if \( \hat{\theta} \) is an estimate of \( \theta \), which is based on observations \( X_1, \ldots, X_n \) from the density \( f_\theta(u) \), and if the bias of \( \hat{\theta} \) is \( E\hat{\theta} - \theta = b(\theta) \), then the variance of \( \hat{\theta} \) will be greater than or equal to

\[
\frac{(1 + b'(\theta))^2}{E(\frac{d}{d\theta} \ln f_\theta(X))^2},
\]

where \( X \) is distributed according to \( f_\theta \).

This result gives a useful quantity to compare estimates with. In particular, if our attention is restricted to estimates \( \hat{\theta} \) which are unbiased, or which have bias which is of a smaller order of magnitude than their standard deviation, it is said that the efficiency of an estimate is the ratio of the Cramèr Rao lower bound to the variance of \( \hat{\theta} \). Therefore, when an unbiased estimate has efficiency 1, there can be no unbiased estimates which have smaller variance.

Even though the Cramèr Rao lower bound gives a means of comparing estimates, it doesn't give a very good method of comparing estimates which have differing bias properties. However, the ideas underlying a method which does were
developed by Neyman and E. S. Pearson (1933) and organized into a formal theory by Wald (1950). This method is what is today called decision theory. Unlike
two previous methods, decision theory assumes that the consequences of estimating a
parameter \( \theta \) by the number \( C \) can be quantified in terms of the loss function \( L(C, \theta) \).
Then, since \( C \) is typically a function of the random data, the objective of the deci-
sion problem is to choose an estimate \( \hat{\theta}(X_1, \ldots, X_n) = \hat{\theta} \), which makes \( E(L(\hat{\theta}, \theta)) \),
called the risk function of \( \hat{\theta} \), small. This is not a simple task, since the risk function
is a function not only of the estimate chosen, but also of the unknown parameter \( \theta \). If one relies simply on the risk criterion, the only way to compare estimates is to
say that \( \hat{\theta}_1 \) is preferable to \( \hat{\theta}_2 \) if, for all \( \theta \),
\[
E(L(\hat{\theta}_1, \theta)) \leq E(L(\hat{\theta}_2, \theta)),
\]
and, for some \( \theta_0 \), the above inequality is strict. When such a \( \hat{\theta}_1 \) exists, it is said
that the estimate \( \hat{\theta}_2 \) is inadmissible, and it is said that an estimate \( \hat{\theta} \) is admissible
among a class of estimates if no other estimate, in that class is preferable to it.

A difficulty associated with the use of admissibility as a criterion by which
to judge estimates is that there may be many admissible estimates, since different estimates can perform well for different parameter values. There are several ways
to avoid this difficulty. One way is to use the minimax estimate, which seeks to
minimize the maximum possible risk that could be suffered for a given estimate.
Thus, if \( \hat{\theta}_M \) is the minimax estimate, then
\[
\max_{\theta} E(L(\hat{\theta}_M, \theta)) = \min_{\hat{\theta}} \max_{\theta} E(L(\hat{\theta}, \theta)).
\]

An alternative approach is to assume that \( \theta \) itself is a random variable whose
distribution is assumed known, and then try to minimize
\[
E_\theta E_{\theta|\theta}(L(\hat{\theta}, \theta)).
\]
This quantity is called the Bayes risk of the estimate \( \hat{\theta} \), and the estimate has minimal
Bayes risk is called the Bayes estimate, or, more commonly, the Bayes rule.

This method may be extended to what is called generalized Bayesian estimation
by not requiring that the distribution on the parameter space be a proper probability
distribution. In particular, it may not integrate to one. Thus, a generalized Bayesian
estimate with respect to the prior \( \tau \) will be the estimate that minimizes
\[
\int E(L(\hat{\theta}, \theta)) d\tau(\theta).
\]
Bayesian and generalized Bayesian estimates are a good class from which to
select an estimate, since it can be shown that all admissible estimates have Bayes
risk that is within an arbitrary constant \( \epsilon \) of that of the Bayes estimate for an appropriately chosen prior. However, many statisticians dislike using Bayesian estimates, since these estimates have the property that the value the estimate takes will depend on the choice of a prior distribution on the parameter space, a choice which is left up to the statistician. In this sense Bayesian estimates aren’t deterministic.

One way that has been used to circumvent this objection in part is to assume that the prior distribution is a member of a parametric family that includes distributions which contribute very little prior information about the value of \( \theta \) and then to use the observed data to estimate the parameters of the prior distribution. The estimate that is used, then, is the Bayesian estimate that corresponds to the estimated parameter values. Such methods are called empirical Bayesian methods.

The field of empirical Bayesian estimation had its beginning with a paper by Charles Stein in the Proceedings of the Third Berkeley Symposium, in which it was proved that, if the random variable \( X = (X_i) \) is distributed as a multivariate normal vector with mean \( \theta = (\theta_i) \) and with the identity matrix as its covariance matrix, then \( X \) is inadmissible as an estimate of its mean \( \theta \) whenever its dimension \( d \) is three or more. This was demonstrated in the paper of James and Stein in the fourth Berkeley Symposium by showing that the estimate

\[
\hat{\theta} = X \left( 1 - \frac{d - 2}{S} \right),
\]

has uniformly lower risk, where \( S = \sum X_i^2 \). This estimate is called the James-Stein estimate.

One way in which the James-Stein estimate can be motivated is by looking at the Bayesian estimate for the situation that the \( \theta_i \)'s are assumed to be themselves independent and normally distributed with mean 0 and common variance \( \sigma^2 \). The Bayesian estimate of \( \theta \) would then be

\[
\left( \frac{\sigma^2}{\sigma^2 + 1} \right) X = X \left( 1 - \left( \frac{1}{\sigma^2 + 1} \right) \right).
\]

Since the expectation of \( 1/S \) is \( ((d - 2)(\sigma^2 + 1))^{-1} \) if \( \theta = 0 \), it is seen that the factor used to weight \( X \) in the James-Stein estimate is an unbiased estimate of the weighting factor for \( X \) in the Bayesian estimate.

The results for the case of the estimation of a multivariate normal mean have been modified and extended in a number of papers, most notably in the series of papers by Efron and Morris (1971, 1972).

Several results have also been proved for the estimation of other multivariate parameters. In the situation where \( X = (X_i) \) is multinomially distributed with
parameters \( n \) and \( p = (p_i) \), results analogous to the James-Stein results have been proved in a paper by Fienberg and Holland (1973) and in a later paper by Sutherland, Fienberg and Holland (1974). The technique that is used is to assume that the parameter, \( p_i \), has a Dirichlet prior distribution with parameters \( \lambda = (\lambda_i) \) and \( K \), where \( \lambda \) is assumed known and \( K \) is to be estimated from the data. It is worth noting that the effect of such a Dirichlet prior is to add \( K \) additional observations to the sample whose relative cell frequencies correspond to \( \lambda \). Nothing additional is actually observed, but the effect is the same as if more data had been observed. Thus, the Bayesian estimate of the mean will be

\[
\hat{\mu} = \frac{K}{n+K} \bar{\lambda} + \frac{n}{n+K} \hat{\mu},
\]

where \( \hat{\mu} = \frac{X}{n} \) is the observed mean. Note that, like the James-Stein estimate, this is a convex combination of the observed mean and the assumed mean of the prior distribution, \( \lambda \). The estimate that is suggested of \( K \) is its maximum likelihood estimate,

\[
\hat{K} = \frac{n^2 - \sum X_i^2}{\sum X_i^2 - 2n \sum X_i \lambda_i + n^2 \sum \lambda_i^2}.
\]

The estimate used is then

\[
\hat{Q}(\hat{K}, \lambda) = \frac{n}{n + \hat{K}} \hat{\mu} + \frac{\hat{K}}{n + \hat{K}} \lambda.
\]

Unfortunately, the properties of the resulting empirical Bayesian estimate are not as strong as are those of the James-Stein estimate. This stems from the fact that, in this case, the naive estimate, \( \hat{\mu} \), is admissible. Therefore, the strongest result that can be proved is that \( \hat{Q}(\hat{K}, \lambda) \) has lower risk than \( \hat{\mu} \) for a large set of underlying \( p \)'s that, as the sample size increases, expands to include any \( p \) 's for which \( \lambda_i = 0 \), for all \( i \).

The paper by Sutherland, Fienberg and Holland considers a slightly more general form of estimate and shows that a variation of this estimate, using

\[
\hat{K} = \frac{n^2 - \sum X_i^2}{\sum X_i^2 - 2(n - 1) \sum X_i \lambda_i + n(n - 1) \sum \lambda_i^2},
\]

has asymptotic properties which are preferable to the estimate of the previous paper.

Results have also been proved by James Peng which generalize the form of the James-Stein estimate to the case where \( X \) is a multivariate Poisson random variable with mean \( \lambda = (\lambda_i) \). His estimate is also a convex combination of the vector of observations and another fixed vector, and the proof of his results are similar to Stein's proofs, though no empirical Bayesian interpretation has been made.
Thus we see that in either a Bayesian or an empirical Bayesian context it is of interest to look at estimates of an underlying parameter \( p \) which are of the form 
\[
\hat{q}(\alpha) = \alpha \hat{p} + (1 - \alpha)\lambda,
\]
where \( \alpha \) and \( \lambda \) each may be either deterministic functions of the sample size \( n \) or random variables depending on the sample size. That is, they may be functions both of the sample size and of the sample itself. It would be of interest to determine what estimate of this form has minimal loss, where the loss is defined as the squared Hilbert space norm, \( ||\hat{p} - p||^2 \).

The loss of \( \hat{q}(\alpha) \) is
\[
L(\hat{q}(\alpha), p) = E||\hat{q}(\alpha) - p||^2 = E||\alpha \hat{p} + \alpha \lambda - p||^2,
\]
where \( \alpha = 1 - \alpha \),
\[
= E||\hat{p} - p||^2 + 2E\alpha(\hat{p} - p, \lambda - \hat{p}) + E\alpha^2||\lambda - \hat{p}||^2.
\]

Thus, in order to minimize \( L(\hat{q}(\alpha), p) \) over choices of \( \alpha \), (or if you like, over choices of \( \alpha \)), the partial derivative of the loss with respect to \( \alpha \) is set equal to zero. This implies that
\[
\frac{\partial}{\partial \alpha} L(\hat{q}(\alpha), p) = 0
\]
\[
= 2E(\hat{p} - p, \lambda - \hat{p}) + 2E\alpha||\lambda - \hat{p}||^2.
\]

This minimum is achieved when
\[
\alpha = \frac{\langle \hat{p} - p, \hat{p} - \lambda \rangle}{||\hat{p} - \lambda||^2}.
\]

Unfortunately, the actual use of this \( \alpha \) is most impractical, since the quantity \( \hat{p} - p \) is unknown and represents the variation in the naïve estimate, \( \hat{p} \). The best that can be hoped for is to estimate accurately the bias of \( \hat{p} \), \( E(\hat{p} - p) \) and use this value in the expression for \( \alpha \) in place of \( \hat{p} - p \) itself. The difficulty in using this approach is that the bias of \( \hat{p} \), which I will call \( p_0 - p \), typically depends on the unknown parameter, \( p \), and must be estimated from the data if this approach is to be followed. If \( p_0 - p \) is \( O(N^{-g}) \) and the limit \( N^g(p_0 - p) \) can be estimated consistently, this implies that the choice of \( \hat{p} \) was not an optimal one, since, when \( N^{-g} \) times the estimate of \( \lim N^g(p_0 - p) \) is subtracted from the estimate \( \hat{p} \), the resulting estimate will have asymptotically the same variance as \( \hat{p} \), and will have less bias than it does.

In the following chapters, I will assume that, for whatever reason, \( \hat{p} \) is thought to be a good estimate of \( p \) in the absence of information about the underlying
parameter, \( p \), in the sense that as many terms of its bias have been eliminated as is thought to be appropriate, so that the question of subtracting bias estimates may be bypassed.

An alternate approach is to look for shrinkage parameters \( \hat{\alpha} \) which don’t depend on the actual values \( \hat{p} \) and \( \lambda \), but only on their distributional properties.

As before, \( L(\hat{q}(\alpha)) \) is quadratic in \( \hat{\alpha} \) and is minimized for the value

\[
\hat{\alpha} = \frac{E(\hat{p} - p, \hat{p} - \lambda)}{E||\hat{p} - \lambda||^2} = \frac{E||\hat{p} - p||^2 + \langle p_0 - p, p - \lambda \rangle}{E||\hat{p} - \lambda||^2}.
\]

Note that the first term in the numerator of this expression is simply \( L(\hat{p}, p) \) and that the second term can be interpreted as the systematic bias of \( \hat{\alpha} \) in the direction of \( p - \lambda \). When \( p \) is not equal to \( \lambda \), then the order of magnitude of the second term will be the same as the bias of \( \hat{\alpha} \), or as the square root of the squared bias of \( \hat{p} \). Since the loss of \( \hat{p} \) may be decomposed into a squared bias term, \( ||p_0 - p||^2 \), plus a “variance” term, \( E||\hat{p} - p_0||^2 \), a reduction in the variance term could be offset by a larger increase in the bias term, making the loss greater than before. This could occur unless the squared bias of \( \hat{p} \) is of a smaller order of magnitude than the squared loss of \( \hat{p} \) or unless the limit of \( N^\sigma(p_0 - p) \) is orthogonal to the quantity \( p - \lambda \). Neither condition will be satisfied in general, since the squared bias may be of the same order of magnitude as the loss itself. Consequently, the second term will have to be estimated in order to achieve this form of optimality. It is only when \( \hat{p} \) is an unbiased estimate of \( p \), that this problem will not occur, and a direct analog of the James-Stein estimate may be hoped for.

In general, as before, the value \( p_0 - p \) appears in the expression for the optimal constant \( \hat{\alpha} \). If the lead term of this quantity can be estimated then that implies that the choice of the naive estimate, \( \hat{p} \), was less than ideal. Consequently, in the later chapters, I will work with the assumption that \( \hat{p} \) is a good estimate, in that it is not desirable to estimate the dominant term of its bias. Under this assumption, whatever improvement could be gained by decreasing the bias of \( \hat{p} \) would represent a potential loss if one were to modify the estimate in the direction opposite to \( \langle p_0 - p, p - \lambda \rangle \). Thus when assuming that it is undesirable to estimate the bias of \( \hat{p} \), all one can hope to accomplish by the use of the estimate \( \hat{q}(\alpha) \) is to leave the estimate essentially unchanged whenever \( p \neq \lambda \) and to gain the maximum possible improvement (that is, use \( \hat{\alpha} = 1 \)) whenever \( p = \lambda \). The estimate I will use in the chapter IV to achieve this objective will be of the form \( \hat{q}(\alpha) \), where

\[
\hat{\alpha} = \min\left(\frac{n^{-c}}{||\hat{p} - \lambda||^2}, 1\right).
\]
In the case that $\lambda$ is not known, but estimated by $\hat{\lambda}$, I will use the same form of estimate with

$$\tilde{\alpha} = \min\left(\frac{n-c}{||\hat{\rho} - \hat{\lambda}||^2}, 1\right).$$

The choice of the constant $c$ will be discussed later in conjunction with the discussion of the estimate's properties.
III. A review of density estimation methods.

In this chapter, I will give an overview of the most common methods of both parametric and nonparametric density estimation. My emphasis will be on those properties that will prove useful and important in the later chapters which contain my own results. As such, several very elegant and interesting sets of results will be omitted, but their omission is in no way indicative of their importance, but merely that they prove neither applicable nor illustrative in the context of this paper. Likewise, most of my discussion of density estimation will center on the most simple case in which the observations are real numbers, rather than vectors or elements of some more abstract space. There is typically no difficulty in extending the results at least to the case of vector valued observations, but in general the arguments are no more enlightening in the more general case than when the observations are univariate, so I will usually give only the results and sketchy arguments for the more general cases.

In estimating a density, it is not at all surprising that a primary consideration in the selection of an estimate is the criterion by which the performance of the estimate is to be judged. The criteria on which I will concentrate will be the pointwise mean squared error of the estimate and the weighted integral of the mean squared error with respect to the measure $\mu$. These two criteria occur frequently in the nonparametric density estimation literature. The weighted integral criterion is most frequently encountered in the situation where the measure $\mu$ is taken to be Lebesgue measure on $d$-dimensional Euclidean space. In this situation it is called the mean integrated square error, which is abbreviated as MISE. However, there is seldom a practical reason for restricting one's attention to this special case, since the proofs generally still hold whenever $\mu$ is absolutely continuous with respect to Lebesgue measure.

One might also be interested in finding nonparametric estimates which are pointwise minimum variance unbiased estimates, but under reasonable conditions,
not only do minimum variance unbiased estimates fail to exist, but also estimates which are merely unbiased fail to exist. This statement can be made precise by citing a result of Rosenblatt (1956):

Assume that $X_1, \ldots, X_n$ are independent and identically distributed observations from the underlying continuous density $p$, and $\Theta(y; X_1, \ldots, X_n)$ is a positive, Borel measurable estimate of $p(y)$ for each $y$ on the real line, which is symmetric in the arguments $X_1, \ldots, X_n$. Then on no interval of the real line is $\Theta$ a pointwise unbiased estimate of $p$.

The proof of this result is noteworthy in that it relies on the fact that, for a large enough family of distributions, the order statistics form a complete statistic for the underlying distribution. Thus, in any situation in which the class of possible underlying densities is broad enough to make the order statistics a complete statistic, it will be necessary to concentrate on finding biased estimates which perform well with respect to mean (integrated) square error rather than unbiased estimates which have small (integrated) variance.

The oldest and simplest method of nonparametric density estimation is the histogram, which was being employed, though not in its present form, as early as the 17th century.

The idea behind the histogram is first to discretize the observed data, and then to report the observed relative frequencies. If one takes the observed frequency of observations in a given element of the partition and then divides that number by the Lebesgue measure of the element, the resulting number is an estimate of the probability density (with respect to Lebesgue measure). Similarly, if one takes the observed relative frequency and divides by the $\mu$-measure of the element of the partition, the resulting number is an estimate of the Radon-Nikodym derivative of the probability measure $P$ with respect to the measure $\mu$.

This estimate is in general biased, since the density may not be constant over the partition element. In fact, if the observation space is divided according to the partition $\{S_i\}, i = 1, \ldots, K$, then the expectation of the histogram estimate evaluated at a point $u \in S_i$ is:

$$\frac{1}{\mu(S_i)} \int_{S_i} p(v) d\mu(v) = \frac{P(S_i)}{\mu(S_i)},$$

where $p$ is the underlying density, $\mu$ is the measure with respect to which the density is calculated, and $P$ is the corresponding probability measure.

Similarly, the variance of the histogram, evaluated at the point $u$ is $\frac{c(c-1)}{n\mu(S_i)^2}$, where $c = P(S_i)$, the probability of the $i$th element of the partition. If one assumes that for $v$ an element of $S_i$, $p(v) = p(u) + O(h)$ (uniformly in $v$), then the expectation simplifies to:

$$\frac{1}{\mu(S_i)} \int_{S_i} (p(u) + O(h)) d\mu(v) =$$
\[
\frac{1}{\mu(S_i)} (p(u) + O(h)) \mu(S_i) = p(u) + O(h).
\]

Thus, the mean squared error of the histogram evaluated at the point \( u \) will be

\[
O(h^2) + \frac{1-c}{n \mu(S_i)} \frac{P(S_i)}{\mu(S_i)} =
\]

\[
O(h^2) + \frac{1-c}{n \mu(S_i)} (p(u) + O(h)) =
\]

\[
O(h^2) + \frac{(1-c)p(u)}{n \mu(S_i)}
\]

plus a term of smaller order.

By choosing the partition to be relatively fine, the first term may be made small by decreasing \( h \) at the expense of making \( \mu(S_i)^{-1} \) larger.

When the observation space is \( d \)-dimensional Euclidean space, and \( p \) is continuous in a neighborhood of \( u \), then as long as the elements of the partition are “not too oblong” and if the diameters of the partition elements are \( O(h) \), then, within the set \( S_i \), \( p(v) = p(u) + O(h) \), whereas \( \mu(S_i) = O(h^d) \). Thus, under these circumstances, we have that the mean squared error of the histogram estimate evaluated at the point \( u \) is:

\[
O(h^2) + O(h^{-d}) \frac{(1-c)p(u)}{n}
\]

plus a term of smaller order,

\[
= O(h^2) + O(nh^{-d}).
\]

In this situation, the rate of convergence of the mean squared error to zero is maximized by choosing \( h \) such that \( h^2 = O(nh^{-d}) \), which is tantamount to letting \( h = O(n^{-\frac{1}{d+1}}) \). This makes the mean squared error of the order of \( n^{-\frac{d}{d+1}} \).

In its most common application, the histogram partition consists of intervals of equal length, or, in higher dimensions, hypercubes of equal volume. Without loss of generality, we may assume that the point \( u \) lies in the partition element that contains the origin. If the underlying density has a continuous second derivative in a neighborhood of \( u \), then the expectation of the histogram will be \( p(0) + \frac{h^2}{24} p''(0) \) plus a term of smaller order, or, in higher dimensions,

\[
p(0) + \frac{h^2}{24} \sum_{i=1}^d p_{ii}(0)
\]
plus a term of smaller order, where \( p_{ii} \) is the \( i \)-th second partial of \( p \).

However, the quantity to be estimated is not \( p(0) \), but \( p(u) \), so the bias of the estimate will be

\[
p(u) - p(0) = \frac{h^2}{24} \sum_{i=1}^{d} p_{ii}(0) = up'(0),
\]

plus a term of order \( h^2 \). In general, however, \( u \) is of the same order as \( h \), so the bias is also \( O(h) \).

Similarly, the variance of the histogram evaluated at \( u \) will be \( p(0)/nh^d \), plus a term of smaller order, so in order to make the order of magnitude of the mean squared error as small as possible, we choose \( h \) so that the squared bias and the variance have the same orders of magnitude, that is:

\[
O(h^2) = O((nh^d)^{-1}),
\]

which gives us that \( h = O(n^{-\frac{1}{d+2}}) \).

This choice of \( h \) makes the mean squared error of order \( n^{-\frac{d+2}{d}} \).

If in fact \( u \) happens to lie at the midpoint of one of the partition intervals (hypercubes), then the bias of the histogram will be of order \( h^2 \), rather than of order \( h \). In this case, the best choice of \( h \) will be

\[
h = O(n^{-\frac{1}{d+2}}),
\]

and this choice will make the mean squared error of order \( n^{-\frac{d+2}{d}} \). This fact will provide the motivation for Rosenblatt's estimate, which will be discussed later.

In each of the above cases, the best choice for \( h \) is of the form \( cn^k \). When \( u \) is not at the midpoint of an interval, the asymptotically best choice of \( h \) is

\[
h = \left( \frac{2nz^2p'(0)^2}{p(0)} \right)^{-\frac{1}{2}},
\]

where \( u = z h \), and \(|z| < .5\). Note that the best choice for \( h \) depends not only on where \( u \) lies in the interval \((-\frac{1}{2}, \frac{1}{2})\) but also on both \( p(0) \) and \( p'(0) \), which are unknown quantities. In the multivariate case, a similar result shows that the asymptotically best choice of \( h \) when \( u \) doesn’t lie at the midpoint of a partition interval is of the order of \( n^{-\frac{d+2}{d+2}} \).

When \( u \) does lie at the midpoint of an interval, the asymptotically best choice for \( h \) is:

\[
h = \left( \frac{p(0)}{4n \sum_{i=1}^{d} p_{ii}(0)^2} \right)^{-\frac{d+2}{d+2}},
\]

14
so the asymptotically best choice of $c$ is:

$$\left( \frac{p(0)}{4 \sum_{i=1}^{d} p_{ii}(0)^2} \right)^{-\frac{1}{2+d}},$$

which likewise depends on unknown quantities, $p(0)$ and $p_{ii}(0), i = 1, \ldots, d$. Thus in this case, as in most other cases, there are constants whose values affect the performance of the density estimate, and which are not at all easy to choose. However, regardless of the choice of the constant $c$, the order of magnitude of the mean squared error will remain the same. In the latter chapters of this paper, I will be tacitly assuming that some wise choice (or at least an acceptable choice) has been made of the constants on which the density estimate depends.

In a similar way, one may wish to choose a histogram which performs well with respect to the loss function of mean integrated square error (MISE). It is an immediate consequence of the previous argument and the fact that the MISE is equal to the integral of the mean squared error of the estimate at a single point that the MISE for a histogram estimate will be:

$$\frac{1}{nh} + \frac{h^2}{12} \int (p'(u))^2 du,$$

plus terms of smaller order. As before, the best choice of $c$ depends on unknown constants, but the best choice of $k$ is free of such constants.

Another property that might be desired is that the density estimate evaluated at the point $u$, when properly normalised, converge in distribution to a nontrivial distribution of known form. This would be a useful property in that it would facilitate the testing of the hypothesis $p(u) = c_0$, or of the hypothesis $p(u) > c_0$.

When the estimate is a histogram, it is a result of one form of the Lindeberg-Feller theorem that $(\hat{p}(u) - E\hat{p}(u))/\sqrt{Var(\hat{p}(u))}$ converges in distribution to a standard normal distribution. As I have shown, in this case

$$E\hat{p}(u) = p(u) + rp'(u) + O(h^2),$$

where $u = kh + r$ and $|r| < \frac{1}{2}$. Also,

$$Var(\hat{p}(u)) = \frac{1}{nh} p(u) + O\left( \frac{1}{n} \right).$$

Having examined the simplest of density estimates in some detail, I will give a more sketchy review of some other forms of density estimation. The simplest idea that may be employed to improve upon the histogram as a density estimate
is to use the fact that the histogram has bias of a smaller order of magnitude at
the midpoints of partition intervals and to take as an estimate a function which
is piecewise polynomial with coefficients depending on the value of the histogram
at the midpoints. Then, having reduced the bias of the estimate in this way, one
should revise one's choice of the interval width, \( h \), in order to equate once again the
rates of convergence of the squared bias and of the variance. It should be noted,
however, that in order to use this method of bias reduction, it must be assumed
that some higher derivative or derivatives of the underlying density exist and, when
they fail to exist, such a method results in an increase rather than a reduction in
the order of magnitude of the pointwise bias. This method is similar to that of
Schoenberg, who, instead of using simply a piecewise polynomial interpolator of the
interval midpoints, uses a polynomial spline.

If all one wants to do is to eliminate the term of the bias which is \( O(h) \),
then this method yields an estimate which is piecewise linear between the points
\((kh, \hat{p}(kh))\), where \( \hat{p} \) is the histogram. This interpolated histogram is usually called
the frequency polygon. It has the desirable properties that it is everywhere positive
and integrates to one, (that is, it is itself a density) and the order of magnitude
of its bias at the point \( u \) is \( O(h^2) \), as long as \( p'' \) exists in a neighborhood of \( u \).
Similarly, the contribution of the squared bias to the MISE will be \( O(h^4) \) if \( p'' \) is
uniformly bounded on the line. Thus, in this case, in order to make the rate of
convergence of the mean squared error or of the MISE to zero as rapid as possible,
\( h \) should be chosen to be \( O(n^{-\frac{1}{6}}) \), which will make the MSE or the MISE of order
\( n^{-\frac{4}{3}} \). Analogous methods exist for eliminating the \( O(h) \) term of the bias when the
observation space is \( d \)-dimensional Euclidean space in which a hyperboloid is used
to interpolate between the estimated values at the midpoints of the histogram cells.
In this case it is necessary to assume that all second order partial derivatives are
bounded, in a neighborhood of the point in question for the mean squared error, or
uniformly bounded for the MISE. In this case, \( h \) should be chosen to be \( O(n^{-\frac{1}{6}}) \)
and the loss will be of the same order of magnitude as \( n^{-\frac{4}{3}} \).

If one wants to eliminate not only the bias term which is \( O(h) \), but also some
terms of higher order, this method will not be as straightforward, since the histogram
itself has bias of order \( h^2 \) even at the midpoints of intervals. When one wants merely
to eliminate the \( O(h^2) \) term of the bias, it can be shown that:

\[
E \frac{1}{24} (26 \hat{p}(kh) - \hat{p}((k - 1)h) - \hat{p}((k + 1)h)) = p(kh) + O(h^3). 
\]

The estimate can then be calculated by using the estimate which within the
interval \((kh - \frac{3}{2}, kh + \frac{3}{2})\) is the parabola which agrees with the above values at
the points \((k - 1)h, kh, \) and \((k + 1)h \). The best choice for \( h \) in this case will be
\( h = O(n^{-\frac{1}{6}}) \) and the mean squared error of the estimate will be \( O(n^{-\frac{4}{3}}) \) as long
as \( p'''' \) exists in a neighborhood of the point in question. The MISE will be of the same order of magnitude as long as \( p'''' \) uniformly exists and is bounded.

It should be noted that in this case, one can no longer conclude that the estimate will be everywhere positive, though it will still integrate to one.

It is possible to employ such a method for the elimination of the bias term of order \( h^2 \) when the observation space is of dimension greater than one if you are willing to assume the existence of all the third partial derivatives of \( p \).

Also, similar, though more complicated expressions exist for the elimination of still more terms of the bias, though these of course require one to assume the existence of still more derivatives or partial derivatives for \( p \). Such estimates are piecewise polynomial with coefficients which depend on the values taken by the histogram in the intervals near the point in question. Both the degree of the polynomials and the interval width will increase as the number of assumed (partial) derivatives increases.

As I have noted previously, a useful fact about the histogram is that the order of magnitude of its bias is smaller at the midpoint of an interval than it is anywhere else in the interval. This is precisely the fact which allows the interpolative methods to increase the rate of convergence of the mean (integrated) squared error. Another method which makes use of this fact is Rosenblatt's estimate, which first appeared in his 1956 paper. The idea behind this estimate is simply to choose the histogram interval to center at the point in question. When one is interested in a global density estimate, then at each point \( u \), the estimate is taken to be the value of the histogram estimate with the interval \((u - \frac{h}{2}, u + \frac{h}{2})\). In other words,

\[
\hat{p}(u) = \frac{1}{nh} \sum_{i=1}^{n} I_{(u - \frac{h}{2}, u + \frac{h}{2})}(X_i) = \frac{1}{nh} \sum_{i=1}^{n} I_{(\frac{X_i - u}{h})}.
\]

where \( I_S \) is the indicator function of the set \( S \).

It is a consequence of the expressions calculated for the histogram estimate that the pointwise bias of Rosenblatt's estimate has dominant term \( h^2 p''(u) \) and that, as before, its pointwise variance has dominant term \( \frac{1}{nh} p(u) \). Thus the asymptotically best choice for \( h \) will satisfy:

\[
h^5 = \frac{p(u)}{4np''(u)^2},
\]

and in order to minimize the MISE, the asymptotically best choice for \( h \) satisfies:

\[
h^5 = \frac{1}{4n \int (p''(u))^2 \, du}.
\]
In either case, the loss is of the same order of magnitude as \( n^{-\frac{4}{3}} \), the same order as for the frequency polygon. Like the interpolated histogram, Rosenblatt’s estimate can be generalized in order to estimate higher dimensional densities. In doing so, the best choice for \( h \) is \( h = O(n^{-\frac{4}{4+2}}) \), which will yield mean squared error or MISE of order \( n^{-\frac{4}{4+2}} \).

The form of Rosenblatt’s estimate can also be interpreted as the density resulting from the convolution between the empirical distribution function and a uniform distribution on the interval \((-\frac{b}{2}, \frac{b}{2})\). This may be written as:

\[
\hat{p}(u) = \frac{1}{h} \int K\left( \frac{u-v}{h} \right) dP_n(v),
\]

where \( P_n \) is the empirical distribution, and, in the case of Rosenblatt’s Estimate,

\[
K(x) = \begin{cases} 
1, & \text{when } |x| < \frac{1}{2}, \\
0, & \text{otherwise.}
\end{cases}
\]

This argument suggests that Rosenblatt’s estimate is in fact a specific case of a more general type of estimate, in which the empirical distribution function is convoluted with a general density \( K \), adjusted by the scale factor \( h \). Estimates of this generalized form are called kernel estimates and were first suggested by Parzen in 1962. He considered choices of the function \( K \) which satisfy:

1. \( \int |K(v)| \, dv < \infty \),
2. \( \sup \{|K(v)|, -\infty < v < \infty \} < \infty \),
3. \( \lim_{v \to \infty} |vK(v)| = 0 \),
4. \( K \geq 0 \),
5. \( \int vK(v) \, dv = 1 \), and
6. \( \int vK(v) \, dv = 0 \).

He showed that the kernel estimate given in (1) has bias which goes to zero whenever \( nh_n \to 0 \) and that the variance of the estimate will go to zero whenever \( nh_n \to \infty \).

In particular, if one assumes the existence of \( p'' \) in a neighborhood of the point \( u \), the bias of the kernel estimate will be:

\[
\frac{1}{2} h_n^2 p''(u) \int v^2 K(v) \, dv + o(h_n^2)
\]

and the expression for the dominant order term of the pointwise variance is:

\[
\frac{1}{nh_n} p(u) \int K^2(v) \, dv + \frac{1}{n} p^2(u) + O\left( \frac{h_n}{n} \right).
\]
Therefore, the dominant order terms of the mean squared error at the point \( u \) will be:

\[
\frac{1}{4} h_n^4 (p''(u))^2 \sqrt{\int v^2 K(v) \, dv} + \frac{1}{n h_n} p(u) \int K^2(v) \, dv.
\]

In order to minimize this quantity, \( h_n \) should satisfy:

\[
h_n^5 = \frac{p(u) \int K^2(v) \, dv}{n (p''(u))^2 (\int v^2 K(v) \, dv)^2}.
\]

If \( p''' \) is uniformly bounded, then the kernel estimate behaves well globally as well as pointwise. To be precise, the MISE will have dominant terms:

\[
\frac{1}{4} h_n^4 \int p''(v)^2 \, dv \left( \int v^2 K(v) \, dv \right)^2 + \frac{1}{n h_n} \int K^2(v) \, dv.
\]

Thus, the dominant term of the MISE is minimized for \( h_n \) satisfying:

\[
h_n^5 = \frac{\int K^2(v) \, dv}{(\int (p''(v))^2 \, dv) (\int v^2 K(v) \, dv)^2}.
\]

We see that, in each case, the best choice for the constant \( h_n \) depends on unknown quantities, \( p(u) \) and \( p''(u) \) or \( \int (p''(v))^2 \, dv \), but in any event, \( h_n \) should be taken to be \( O(n^{-\frac{1}{2}}) \), and for such a choice, the mean (integrated) square error will be \( O(n^{-\frac{3}{2}}) \).

It is worth noting that if the sixth condition fails to be satisfied, then the bias of the kernel estimate will be of order \( h_n \), as was the case with the histogram. When this happens, then, as with the histogram, the best choice for \( h \) will be of order \( n^{-\frac{1}{2}} \) and the corresponding mean squared error will be of order \( n^{-\frac{5}{2}} \). This is the reason that most attention is concentrated on the use of symmetric kernels. However, if the loss function is weighted mean squared error, and the weighting function isn’t constant, it will not necessarily be desirable to restrict attention to symmetric kernel functions.

Each of the above refinements of the histogram acts to correct the property that, since the histogram is itself not a very smooth function, functions which are smooth cannot be approximated very closely by it. The interpolated histogram makes use of Taylor’s theorem, which may be roughly paraphrased as: Functions which possess a certain number of derivatives behave locally as if they were polynomials of that order. Similarly, kernel estimates smooth the cumulative distribution function by convoluting it with a distribution which lies very close to zero. Other methods exist to facilitate the smoothing of the histogram, but first one must decide what is meant
when a function is said to be smooth. For that reason, I will next introduce the elementary definitions and properties of Hilbert spaces. I will restrict my attention to real vector spaces and real inner products, not because there is any difficulty in looking at the complex analogs, but because the real space and inner product cover most of the interesting examples without the need for additional structure.

**Definition:** A (real) **vector space** is a set $S$ for which the following hold: If $p, q,$ and $r$ are arbitrary elements of the set $S$, then the binary operator “$+$” satisfies:

1. $p + q = q + p$ is an element of $S$.
2. $(p + q) + r = p + (q + r)$.
3. There exists an element 0 of $S$ for which $0 + p = p + 0 = p$, for all $p \in S$.
4. For any real number $c$, $cp$ is an element of $S$.
5. $1p = p$.
6. $c_1(c_2p) = (c_1c_2)p$.
7. $(c_1 + c_2)p = c_1p + c_2p$, and
8. $c(p + q) = cp + cq$.

**Definition:** A real **inner product space** (or simply an inner product space) is a real vector space $S$ on which an operator (inner product), $\langle \cdot, \cdot \rangle$, is defined which maps the product space $S \times S$ into the real line. For arbitrary elements $p, q,$ and $r$ of $S$ and an arbitrary real constant $c$, the inner product has the following properties:

1. $\langle p, q \rangle = \langle q, p \rangle$
2. $\langle p + q, r \rangle = \langle p, r \rangle + \langle q, r \rangle$
3. $\langle cp, q \rangle = c\langle p, q \rangle$
4. $\langle p, p \rangle \geq 0$ for all $p \in S$, and
5. $\langle p, p \rangle = 0$ implies that $p = 0$.

On such an inner product space, a norm, $\|\cdot\|$, may be defined as $\|p\| = \sqrt{\langle p, p \rangle}$.

This norm will satisfy all the requirements for a vector space norm, namely that, for any $p$ and $q$ in $S$ and real number $c$,

1. $\|p + q\| \leq \|p\| + \|q\|$. 
2. $\|cp\| = |c|\|p\|$. 
3. $\|p\| \geq 0$, and 
4. $\|p\| > 0$ whenever $p \neq 0$.

An inner product space is said to be a **Hilbert space**, whenever it is complete. That is, any Cauchy sequence of elements of $S$ converges to an element of $S$. Thus, by definition, whenever $p_n$ is a sequence of elements of $S$ and for any $\epsilon > 0$ there exists an integer $N_\epsilon$ such that, for $i$ and $j > N_\epsilon$, $\|p_i - p_j\| < \epsilon$, then there exists an element $p$ of $S$ such that $\|p_n - p\| \to 0$.

The completeness of the inner product space will be of little use in the context of this paper, since all that is needed for the later results will be the properties of $S$ as an inner product space, but, since all of the most frequently occurring examples
are Hilbert spaces, I will refer only to Hilbert spaces, even if the results can be proved in somewhat greater generality.

The most common example of a Hilbert space is the space $L_2(\mu)$, the space of all square-integrable functions on the space $S$ with respect to the positive measure $\mu$. The inner product for this Hilbert space is:

$$(f, g) = \int_S f(v)g(v) \, d\mu(v),$$

which yields the norm

$$\|f\| = \sqrt{\int_S f^2(v) \, d\mu(v)}.$$  

In the situation of density estimation, it is convenient to take as a loss function the expected squared norm of the estimate, $\hat{\mu}$, minus the true underlying value. If the measure $\mu$ is taken to be Lebesgue measure, then the corresponding loss function will be the MISE of $\hat{\mu}$. If the measure is taken to be concentrated at the point $u$, then the corresponding loss function will be proportional to the mean squared error of $\hat{\mu}(u)$. Finally, if the measure $\mu$ is general, then the corresponding loss function will be the weighted mean squared error with respect to the measure $\mu$. Also, it is important to note that the space $S$ hasn’t been specified yet, so the Hilbert space formulation can accommodate vector-valued observations as easily as it can accommodate scalar ones by simply choosing $S$ to be a $d$-dimensional Euclidean space.

Another important example of a Hilbert space is that of Sobolev spaces. These spaces are restricted to functions which are smoother than the arbitrary member of $L_2$ in that they are required to possess a number of derivatives which are likewise members of $L_2$. The inner product on a Sobolev space is defined to be:

$$(f, g) = \sum_{i=0}^{k} \int_S f^{(i)}g^{(i)} \, d\mu,$$

where $f^{(i)}$ is the $i$-th derivative of $f$, and $\mu$ in this case is Lebesgue measure on the real line. This space is denoted by $H^k(-\infty, \infty)$, or simply $H^k$.

Of course, the “derivative” of a density at a point (or the value of the density itself at a point) is not a well defined concept in the space $L_2$, since $L_2$ considers functions which are equal almost everywhere as being the same function. Therefore the statement “$p'(u) = A$” is taken to mean that there is a version of the density $p$ which is differentiable at the point $u$, and for which the value of the derivative is $A$. Likewise, whenever I consider the mean squared error of $\hat{\mu}(u)$, I am assuming that there is some nice version of the density $p$ with respect to which $\hat{\mu}$ is being compared at the point $u$. By nice, I will typically mean either continuous or differentiable.
The introduction of Sobolev spaces suggests different sorts of loss functions which might be desirable, namely those which penalize errors not only in the estimation of the density itself but also in the estimation of the derivatives of the density. It is worth noting that functions which have small (squared) norm in the Sobolev spaces are smooth in a more restrictive sense than functions of small norm in $L_2$, since they must not only be in $L_2$, but their derivatives must likewise be in $L_2$. For example, a function in the space $H^2$ must at least be differentiable and have square integrable derivative, and a function which has small norm in $H^2$ will have a derivative that is bounded outside a set of arbitrarily small measure.

In choosing an estimate it is desired to choose an estimate which has small squared norm in the Sobolev space whenever the underlying density is chosen to have finite norm in that space, and yet approximates the observed data well, as summarized in the histogram. This can be done by looking at the seminorm:

$$
\|f\|_{k_1, k_2}^2 = \sum_{i=k_1}^{k_2} \int_S f^{(i)}^2 \, d\mu
$$

and choosing as an estimate the function of minimum seminorm which assigns the same probabilities to each of the partition intervals as the histogram. When the partition consists of a collection of fixed intervals and $k_1 = k_2 = 2$, the minimizing function is a piecewise cubic spline, and the solution is called a spline or fixed knot estimate. If the intervals of the partition aren’t fixed, but are chosen as some functions of the observed random variables, the solution is called a variable knot estimate. An example of a variable knot estimate would be to choose the endpoints of the intervals from among the order statistics of the observed sample.

Much attention in the literature has been devoted to the pointwise convergence of the mean squared error of spline or knot density estimates. For example, Wahba (Annals of Statistics, 1975) has proved that the rate of convergence of the mean squared error of an appropriately chosen variable knot estimate at a point is the same as that of the optimally converging kernel estimate, namely $n^{-2k+1}$, where $k$ is the largest order of derivative of the underlying density which is assumed to exist.

These results can be extended to estimates which have uniformly convergent mean squared error over intervals (convergent MISE on intervals) by assuming the uniform boundedness of the highest order derivative that was assumed to exist for the pointwise convergent estimate but there is a problem associated with the extension of this result to the entire space (line). The problem is that the associated calculus of variations problem doesn’t have a solution in general when one or more of the intervals are semi-infinite. This is because one can construct sequences of functions whose seminorms converge to zero and yet which don’t converge themselves. There are a number of ways to avoid this problem, including
forcing the estimate to be zero outside a range which includes all the observations or, on the other hand, only considering the estimate within some predetermined finite range. To my knowledge, the MISE convergence properties of these related estimates have not been investigated.

It is important to note two of the drawbacks of knot and variable knot density estimates. First, unlike the standard kernel estimates, they needn’t be positive. Of course, this can be corrected by estimating the density by the positive part of the knot estimate, but such an estimate is no longer as smooth as the original (albeit unrealistic) estimate. A second drawback is that if the estimate is restricted to a finite interval, then, since that interval may not contain all the observations, the estimate may not integrate to one, which again is a desirable property which the kernel estimates possess. However, despite these two considerations, the knot are appealing, both because of their Hilbert space formulation and because of their relative ease of computation, since the kernel methods require on the order of \( n \) function evaluations, whereas the knot estimates require on the order of \( n^k \) operations, where \( n^k \) is the number of intervals chosen, and \( k \) is less than 1.

This last assertion depends on the form of the inner product, since, the calculation of a spline estimate involves the inversion of a square matrix of dimension \( n^k \) by \( n^k \). In the case where the inner product is that of \( L_2 \), that matrix is tridiagonal, greatly simplifying its inversion.

Yet another class of density estimation methods also come from the Hilbert space structure that was used to generate the fixed and variable knot methods. This comes from the fact that a Hilbert space \( H \) is spanned by an orthonormal basis of functions, which may be chosen in a number of ways. Thus, one may express any function \( f \) of the Hilbert space as:

\[
f = \sum_{i=1}^{\infty} c_i \phi_i,
\]

where the functions \( \{\phi_k\} \) are some choice of an orthogonal (or orthonormal) basis for the space. The choice of an orthogonal basis for the Hilbert space \( H \) is quite arbitrary, since, given the axiom of choice, any finite orthogonal set of functions may be extended to a complete orthogonal basis. Also, naturally the choice of an orthogonal basis depends on which Hilbert space is being considered. A useful family of Hilbert spaces to consider are the ones on the set \( H \) for which the inner product is

\[
\langle f, g \rangle = \int_H f(u)g(u)w(u)\,du,
\]

where \( w \) is a function which is zero on a subset of \( H \) of measure zero. I will denote this space by \( H_w \). The orthogonal functions \( \{\phi_k\}, k = 1, \ldots, \infty \) are therefore functions for which, whenever \( i \neq j, \langle \phi_i, \phi_j \rangle = 0 \). It is usually convenient to
impose the additional constraint that \( \langle \phi_i, \phi_i \rangle = \| \phi_i \|^2 = 1 \), in which case the set of functions is called orthonormal.

There is considerable attention in the mathematical literature devoted to the study of particular choices of the weighting function \( w \) and the orthonormal basis \( \{ \phi_k \} \). Of particular interest are the following choices of \( w \) and \( \{ \phi_k \} \):

1. When \( w(u) = (1 - u)^a(1 + u)^b \) on the interval \([-1,1]\), then one choice of an orthonormal basis is the Jacobi polynomials, \( P_k^{(a,b)}(u) \). (Note: the exact form of the orthonormal functions isn't important for this discussion. I won't be showing the explicit form of the orthonormal functions, but they are commonly available in such reference works as Abramowitz and Stegun.)

2. When \( w(u) = (1 - u^2)^{-\frac{1}{2}} \) on the interval \([-1,1]\), one choice of an orthonormal basis is the Chebyshev polynomials of the first kind, \( T_k(u) \).

3. When \( w(u) = 1 \) on the interval \([-1,1]\), then one choice of an orthonormal basis is the Legendre polynomials, \( P_k(u) \).

4. When \( w(u) = e^{-u} \) on the positive half axis, then one choice of an orthonormal basis is the Laguerre polynomials, \( L_k(u) \).

5. When \( w(u) = e^{-u^2/2} \) on the entire line, then one choice of an orthonormal basis is the Hermite polynomials, \( H_k(u) \).

6. Finally, when \( w(u) = 1 \) on the interval \(( -\pi, \pi )\), and \( f(-\pi) = f(\pi) \), then one may use the functions \( \phi_k(u) = \sin(ku) \) (Fourier series) as an orthonormal basis.

If one prefers to remain in the space \( L_2 \) ( \( = H^1 \) ), then a variety of choices of orthonormal basis may be arrived at by first choosing a set of orthonormal functions, \( \phi_k(u) \), for the Hilbert space \( H_w \) and then using the functions \( \phi_k(u)\sqrt{w(u)} \) as the orthonormal basis for \( L_2 \). Whichever basis one chooses, one may write any function \( f \) in \( H \) as

\[
f = \sum_{i=1}^{\infty} c_i \phi_i
\]

\[
= \sum_{i=1}^{\infty} \langle f, \phi_i \rangle \phi_i
\]

\[
= \phi_i \sum_{i=1}^{\infty} \int f(u)\phi_i(u)w(u) \, du.
\]

Thus, the \( k^{th} \) coefficient of the orthogonal series is

\[
c_k = \int f(u)\phi_k(u)w(u) \, du
\]

\[
\int \phi_k(u)w(u) \, dF(u),
\]

24
which may be approximated by

\[ \hat{c}_k = \int \phi_k(u)w(u) d\hat{F}_n(u) \]
\[ = \frac{1}{n} \sum_{i=1}^{n} \phi_k(x_i)w(x_i). \]

The density estimation method which is generally used is to then estimate \( f \) by a partial sum of the terms \( \hat{c}_k\phi_k \). That is, \( \hat{f}(u) = \sum_{i=1}^{m(n)} \hat{c}_k\phi_k(u) \).

A more general approach is to weight the terms by some sequence \( \{\alpha_k\} \), which is chosen to depend only on \( n \), so that the estimate becomes

\[ \hat{f}(u) = \sum_{k=1}^{m(n)} \alpha_k \hat{c}_k\phi_k(u). \]

The choice of the number \( m(n) \) or the sequence \( \{\alpha_k\} \) will depend on a number of factors, including the sample size, the particular Hilbert space being used, the orthonormal basis being used, and the assumed smoothness of the underlying density, as determined relative to the basis being used.

There are many examples of uses of this general method, only some of which I will cite here. However, there is a good review paper by Wegman (Technometrics, 1972) which contains a more complete list of references to papers discussing density estimation using orthogonal series.

The first paper to suggest the use of orthogonal series for the purpose of density estimation was the 1963 paper of N.N. Cencov (Soviet Math), which treated the method in the general, Hilbert space formulation. His main result stated that, if \( \{\phi_n\} \) is an orthonormal basis for the Hilbert space \( H_w \), and if there exist constants \( A, H \) and \( m \) such that

1. \( \langle \phi_k, \phi \rangle \leq Ak^{-m} \), and
2. \( \int |\phi_k(u)w(u)|^2 p(u) d\mu(u) \leq H \),

then if the upper summation limit, \( m(n) \) is chosen to be \( O(n^{-\frac{1}{m}}) \), we get that the squared error \( ||\hat{\rho} - p||^2 \) is \( O_p(n^{-\frac{2m-1}{2m}}) \). Note that this result doesn’t conclude anything about the MISE of \( \hat{\rho} \), which is \( E||\hat{\rho} - p||^2 \), but only about the probabilistic behavior of \( ||\hat{\rho} - p||^2 \).

The next major paper was that of Schwartz (AMS, 1967) in which he suggested the use of Hermite polynomials to construct a decomposition of \( p \). His main result was that, if

25
\[1. e^{\frac{x^2}{2}} \frac{d^r}{dx^r} e^{-\frac{x^2}{2}} p(x) \in L_2, \]
\[2. p \in L_2, \text{and} \]
\[3. m(n) = O(n^{\frac{1}{2}}) \]

then the mean integrated square error of \( \hat{p} \) would be \( O(n^{-\frac{r-1}{r}}) \), where \( \hat{p}(x) = \sum_{i=1}^{m(n)} \hat{c}_i \phi_i(u) \), and \( \phi_i(u) = H_i(u)(2i! \sqrt{\pi})^{\frac{1}{4}} e^{-u^2} \).

Schwartz goes on to show that if the assumption that \( p \) is in \( L_2 \) is replaced by the assumption that it is continuous and of bounded variation, then the MISE of the same estimate is of order \( O(n^{-\frac{r-4}{r}}) \).

The papers of Kronmal and Tarter employ Fourier series as orthogonal functions and derive similar results, though they consider the more general estimates of the form \( \sum \lambda_k \hat{c}_k \phi_k \) and suggest the use of the Fejer weights, \( \lambda_k = \frac{m(n)-k}{m(n)} \) for \( k < m(n) \) and \( \lambda_k = 0 \) otherwise.

Orthogonal series density estimates suffer from some of the same drawbacks and enjoy the same advantages as knot estimates, namely that whereas they are not guaranteed to be everywhere positive and they needn’t integrate to one, they are, nonetheless, computationally less involved than kernel estimates. The non-positivity is if anything a greater problem with orthogonal series estimation than it is with knot estimation due to the so-called Gibbs phenomenon. This phenomenon is the wild oscillation of an estimate in the neighborhood of a major change in the value of the density and is most commonly demonstrated by the partial Fourier series expansions of the indicator function of an interval. The underlying difficulty stems from the fact that it is impossible to estimate a locally non-smooth function well by a smooth function.

A nonparametric density estimation method which has appeared recently and which is quite different in approach from the above methods is the method of Good and Gaskins, which appeared first in a paper in Biometrika (1971) and was expanded upon in a paper by the same authors in the Virginia Journal of Science (1972). In this method, the authors use a modified version of maximum likelihood as a criterion of estimation. The straightforward use of the maximum likelihood principle will lead to a weighted average of dirac delta functions as an estimate, which is scarcely desirable, so the authors modify the likelihood into a so-called penalized likelihood, \( w(f) \), where

\[ w(f) = \sum_i \log f(x_i) - \Phi(f), \]

where \( \Phi \) is a functional that represents the roughness of the function \( f \). Thus, under this method, the weighted sum of dirac delta functions would typically be unacceptable, since it would have \( \Phi = \infty \). This method is a generalized Bayes method, since the resulting estimate is the maximum likelihood estimate calculated.
from the posterior distribution corresponding to a prior distribution proportional to $e^{-\phi(f)}$.

This method of estimation doesn't have the problem of possible negativity that the knot or orthogonal series estimates possess, since it is won't tolerate an estimate that is negative at any of the observation points. This doesn't guarantee that the estimate will be everywhere non-negative, though this property can be easily obtained by using the substitution $g = f^2$ and choosing to estimate $f$ by the positive square root of the function $g$ that minimizes $\sum_i g(x_i) - 2\Phi(\sqrt{g})$, which will yield the same estimate as the original equation would with the non-negativity constraint imposed.

These penalized likelihood methods provide estimates which are very aesthetically pleasing, though they are very difficult to calculate in a continuous setting. Tapia and Thompson (Johns Hopkins Press, 1978) have thus concentrated on the calculation of discrete penalized maximum likelihood estimates, which overcome this difficulty by first reducing the problem to one of the estimation of a finite number of parameters and then taking the density estimate to be a spline fitting of the discrete values.

Another drawback of penalized likelihood methods is that little is known about their mean squared error or mean integrated square error convergence properties. These drawbacks make penalized likelihood estimates inappropriate nonparametric estimates to employ in conjunction with the methods of this paper, though this doesn't imply that they are inappropriate in general.

There are many other methods which may be employed to estimate densities, but the most common of other methods employed are consistent only over a dramatically smaller class of underlying densities (for example, the densities for which the estimates are consistent isn't dense in the space $L_2$). These estimates I have chosen to term parametric, though, as I have stated, this distinction is somewhat artificial. A short review of some of these estimates is presented next.

The estimation of a probability density function when the underlying density is known to be a member of a certain parametric family generally reduces to the problem of estimating the parameters associated with the density. Thus, for example, the estimation of a particular member of a location and scale family of densities generally reduces to the estimation of the location parameter $\mu$ and the scale parameter $\sigma$. Once one has estimated these two quantities by $\hat{\mu}$ and $\hat{\sigma}$, it is usually the case that the estimate

$$\hat{f}(u) = \frac{1}{\hat{\sigma}} f_0 \left( \frac{u - \hat{\mu}}{\hat{\sigma}} \right)$$

is a good estimate of $f$. The general analog to this problem is where the underlying density is assumed to be parameterized by the $k$-dimensional vector $\theta$ and called $f_\theta$. In this situation, it is the case that, under regularity assumptions, the maximum
likelihood estimate $\hat{\theta}$ of $\theta$ will have desirable estimation properties. A discussion of these results will follow shortly, but, since the problem under discussion is that of density estimation, and not merely that of parameter estimation, one would want the density $f_{\hat{\theta}}$ to possess, as an estimate of $f$, some of the properties that $\hat{\theta}$ possesses as an estimate of $\theta$. In order for this to follow, the parameterization of the Hilbert space in which the density lies must reflect the metric properties of the parameter space. In particular, if one wants the maximum likelihood estimates to be consistent in the sense of mean integrated squared error, then the mapping from the parameter space, $\mathbb{R}_k$ to the Hilbert space $L_2$ must be continuous. If one wants to maintain the rate of convergence of the maximum likelihood estimate (that is, $\frac{1}{n}$), then this mapping must be absolutely continuous. These properties do not follow in generality, but must be verified in each particular case. However, these conditions aren’t as restrictive as they might seem on the surface since, due to the invariance properties of the maximum likelihood estimates, you simply have to find a single parameterization of the family with respect to which the mappings are continuous or absolutely continuous. Thus, while this mapping isn’t absolutely continuous with respect to the scale parameter of a parametric family, there are interesting cases, such as the normal scale family, in which the mapping is absolutely continuous with respect to the logarithm of the scale parameter. I will next look at the properties of the maximum likelihood estimate of the underlying parameter, which, if the parameterization is a continuous or absolutely continuous one will imply the convergence properties of the density estimate.

When the underlying parameter, $\theta$, is an element of $p$-dimensional Euclidean space, then certain regularity conditions are necessary to obtain the desired properties of the maximum likelihood estimate. These conditions are made in order that the underlying density be smooth and that certain exchanges of partial derivatives and integrals be valid. The notation that I will use is that the form of the underlying density is $f_{\theta}$, where $\theta$ lies in some open set in $p$-dimensional Euclidean space. If $\theta$ is thought of as a vector with respect to some fixed and predetermined basis, then the $i$-th component of the vector will be denoted by $\theta_i$. The true value of $\theta$ will be denoted by $\theta_0$, and $p$-dimensional Lebesgue measure will be denoted by $\mu$. For the sake of simplicity, I will denote the partial derivative of $f_{\theta}$ with respect to $\theta_i$, evaluated at the point $\theta_0$ by

$$\frac{\partial}{\partial \theta_i} f_{\theta_0}.$$

One set of regularity conditions which are sufficient to prove asymptotic properties of the maximum likelihood estimate of a multidimensional parameter are given by S.A.D.C. Doss in his 1962 paper:

1. For almost all $x$, $\ln(f(x; \theta))$ is continuous throughout the parameter space, $\Omega$.

2. For each $i$, $\frac{\partial}{\partial \theta_i} \ln f(x; \theta)$ exists for almost all $x$. 

28
3. If \( \delta \) represents \( p \)-dimensional Euclidean distance, then there exists a positive function \( h(\theta) \) such that, for every point \( \theta' \) in \( \Omega \) for which \( 0 < \delta(\theta, \theta') \leq h(\theta') \),

\[
\int (\ln(f(x; \theta) - \ln(f(x; \theta')) f(x; \theta') \, dx < 0
\]

4. For every \( \theta \in \Omega \),

\[
\frac{\partial}{\partial \theta_i} f_\theta(x) \, dx = 0
\]

\[
= \int \frac{\partial^2}{\partial \theta_i \partial \theta_j} f_\theta(x) \, dx,
\]

for all \( i \) and \( j = 1, \ldots, p \).

5. For all \( \theta \in \Omega \), the information matrix \( I(\theta) \) is positive definite and has finite determinant, where

\[
I_{ij}(\theta) = -\int \frac{\partial^2}{\partial \theta_i \partial \theta_j} \ln(f_\theta(x)) f_\theta(x) \, dx.
\]

6. There is a set of positive functions, \( g_i(\theta), i = 1, \ldots, p \), which are continuous and differentiable for all \( \theta \in \Omega \) and for which,

\[
\frac{\partial}{\partial \theta_i} g_j(\theta) \frac{\partial}{\partial \theta_j} \ln f_\theta(x)
\]

are uniformly continuous in \( \Omega \).

These conditions are not especially intuitive. In particular, the third and sixth conditions assume the existence of certain bounding functions which must depend on the underlying family of distributions. However, the first three conditions are sufficient to guarantee that the maximum likelihood estimate of the parameter vector is consistent, and, combined with the latter three conditions, the following theorem (Doss's Theorem 3) may be proved:

Under the above conditions, the solution of the system of likelihood equations is asymptotically distributed as a \( p \)-variate normal distribution with mean zero and variance-covariance matrix \( \frac{1}{n} I(\theta_0) \). Moreover, it is a jointly asymptotically efficient estimate of the unknown parameter vector.

Therefore, by this theorem we get that the maximum likelihood estimate of the parameter vector is:

(i) convergent in squared Euclidean norm at the rate \( \frac{1}{n} \),

(ii) convergent in distribution to a multivariate normal distribution at the rate \( \frac{1}{\sqrt{n}} \), and, consequently,
(iii) convergent in probability at the rate $\frac{1}{\sqrt{n}}$, and
(iv) any absolutely continuous function of the parameter vector is also convergent in probability, distribution and squared norm at the same rates.

The principal problem with maximum likelihood estimates of a density is that their properties can be arbitrarily bad when sample sizes are small. For that reason, it is desirable, when possible, to take the sample size into consideration when producing an estimate. One way in which this may be done is to use the unique minimum variance unbiased estimate (UMVU) of the underlying density. One may recall that due to a theorem of Rosenblatt, it is not in general possible to find unbiased estimates of a density, due to the fact that in general, the order statistics form a complete and sufficient statistic for the underlying distribution. However, if one limits the family of possible distributions to a parametric family for which the order statistics aren’t complete and sufficient, this may no longer be the case.

The trouble in finding UMVU density estimates is that, as with the maximum likelihood estimates, they must be calculated separately for each distinct parametric family of distributions. However, with the UMVU estimate, there is the added problem that this estimate is not guaranteed to exist. However, this work has been carried out for certain parametric families of particular interest.

Most of the work in this area has been associated with the problem of finding unbiased estimates of tail probabilities for distributions of certain parametric form. However, if one can find an unbiased estimate $\hat{F}(x)$ of the cumulative distribution function $F(x)$, for all $x$, then its derivative $\hat{f}(x) = \frac{d}{dx}\hat{F}(x)$ will be an unbiased estimate of the underlying density wherever it exists.

The principal results in this area stem from a paper by Tate (1959) and one by Ghurye and Olkin (1969) and are summarized in a technical report by Olkin (1978). The two main families for which results exist are the Gamma–Weibull family and the Normal–Lognormal family.

The result for the gamma family is that, if $X_1, \ldots, X_n$ are $\gamma$-distributed with density

$$f(u) = \frac{1}{\Gamma(p)\theta^p} u^{p-1} e^{-u/\theta},$$

then the unbiased estimate of $f(u)$ is

$$\hat{f}(u) = \frac{1}{B(p, p(n-1)) \sum X_i} u^{p-1} \left(1 - \frac{u}{\sum X_i}\right)^{np-p-1},$$

where

$$B(p, q) = \frac{\Gamma(p)\Gamma(q)}{\Gamma(p+q)}$$

is the Beta function, and
\[ \Psi(x) = \begin{cases} x, & \text{if } x > 0. \\ 0, & \text{if } x \leq 0. \end{cases} \]

Since \( f(u) \) is a function of a complete and sufficient statistic for the gamma family, it is not only unbiased, but also the UMVU estimate of the density.

This \( \Psi \) function will recur frequently, and is defined in general for a symmetric matrix argument as

\[ \Psi(M) = \begin{cases} \det(M), & \text{if } M \text{ is positive definite, and} \\ 0, & \text{otherwise.} \end{cases} \]

Since a gamma random variable with shape parameter 1 is exponentially distributed, this result implies that the UMVU estimate of an exponential density evaluated at the point \( u \), from observations \( X_1, \ldots, X_n \) is

\[ \frac{n - 1}{\sum X_i} \left( 1 - \frac{u}{\sum X_i} \right)^{n-1}, \]

whenever \( u < \sum X_i \), and is zero otherwise.

The UMVU estimate of the Weibull density can be obtained by using the fact that if \( X \) is distributed with the Weibull density

\[ \frac{P}{\theta} u^{p-1} e^{-u^p/\theta}, \]

then \( X^p \) will be exponentially distributed with parameter \( \theta \). Thus, the UMVU estimate of the cumulative distribution function of \( X \) evaluated at the point \( u \) will be the same as the UMVU estimate of the cumulative distribution function of \( X^p \), evaluated at the point \( u^p \) (with the understanding that if \( u < 0 \), the estimate will be zero regardless of the sign of \( u^p \)). Thus the UMVU of the Weibull density evaluated at \( u \) will be the derivative of this with respect to \( u \), which, by the chain rule is \( pu^{p-1} \) times the UMVU estimate of the density of \( X^p \) evaluated at \( u^p \).

The corresponding UMVU density estimates for the Normal–Lognormal family of distributions are somewhat more complex, due to the fact that in this case there are two parameters which need to be taken into account and which may be assumed to be either known or unknown. The situation is also complicated by the fact that the corresponding results for the multivariate analogs of these distributions have been explored in some detail. The history of this research is well outlined in the technical report by Olkin.

In the univariate case, when \( \sigma \) is known, the UMVU estimate of the normal density function evaluated at the point \( u \) is

\[ \sqrt{\frac{n}{n-1}} \frac{1}{\sqrt{2\pi\sigma}} e^{-(u-x)^2/2(n-1)\sigma^2}, \]

31
where $\bar{x}$ is the arithmetic mean of the observations.

When $\sigma$ is unknown, the UMVU estimate is

$$\sqrt{\frac{n}{n-1}} \frac{1}{\nu B\left(\frac{n-2}{2}, \frac{1}{2}\right)} \Psi\left(1 - \frac{n(u - \bar{x})^2}{(n-1)\nu^2}\right)^{\frac{n-4}{4}},$$

where $\nu^2 = \sum (x_i - \bar{x})^2$, and $\Psi$ is the same function as before.

In the multivariate case, when $\Sigma$ is known, the UMVU estimate of the density function is

$$\frac{1}{\sqrt{|\Sigma|}} \left(\frac{2\pi(n-1)}{n}\right)^{\frac{n}{2}} \exp\left\{-\frac{n}{2(n-1)}(u - \bar{x})\Sigma^{-1}(u - \bar{x})'\right\}.$$

When $\Sigma$ is unknown, the UMVU estimate is

$$\frac{\Gamma\left(\frac{n-1}{2}\right)}{\Gamma\left(\frac{n-\frac{p-1}{2}}{2}\right)n^{\frac{p}{2}}(n-1)^{\frac{p}{2}}\sqrt{|\Sigma|}} \Psi\left(1 - \frac{(u - \bar{x})S^{-1}(u - \bar{x})'}{n-1}\right)^{\frac{n-\frac{p-3}{2}}{2}},$$

where the matrix $S$ is $\frac{1}{n}$ times the matrix of central crossproducts. That is,

$$S_{ij} = \frac{1}{n} \sum_{k=1}^{n} (x_{ik} - \bar{x}_i)(x_{jk} - \bar{x}_j).$$

There are also corresponding results for the cases where the covariance matrix is assumed to be an unknown constant $\sigma^2$ times a known covariance matrix $\Sigma_0$, and for the cases where the mean is assumed to be linearly restricted and the covariance matrix is either known, known up to a constant multiple, and completely unknown. The details of these results are contained in the Olkin technical report.

As I have said before, I prefer to consider a density estimate to be truly nonparametric only when it is consistent over a set of possible underlying densities which is dense in the Hilbert space in question (typically $L_2$). By a slight abuse of definition, then, any density estimate which fails to be consistent over such a set is considered "parametric", even though it may not be parametrizable in the usual sense. Such an estimate is the maximum likelihood estimate for a unimodal density, which was developed by Prakasa Rao (1969) and contained in the book by Barlow, Bartholomew, Bremner and Brunk (1972).

The problem with maximum likelihood estimation of a general density is that no global maximum likelihood exists over the set of all densities. This is because, as a function approaches $\frac{1}{n} \sum \delta(X_i)$ in distribution, where $\delta$ is the Dirac delta function,
the associated likelihood will go to \( \infty \). As I have previously mentioned, one way to avoid this problem is through a generalized prior distribution placed on the space of all functions, which yields the penalized maximum likelihood method, developed by Good and Gaskins. An alternative approach is to restrict attention to a smaller class of candidate functions, all of which are smooth. This entails that they can't approach the formal derivative of the empirical cumulative function in distribution. When the space in question is the real line, several such classes of functions are the strictly unimodal functions, where the mode, \( m \), is alternately known, or unknown, as is the probability, \( p \), of an observation less than the mode.

When both \( m \) and \( p \) are known, then the maximum likelihood estimate of the distribution function is the function that, for \( x \) greater than the known mode, \( m \), is the least concave majorant of the empirical cumulative distribution function which passes through the point \((m, p)\), where \( p \) is the probability that \( X \) is less than \( m \), and, for \( x \) less than \( m \), it is the greatest convex minorant of the empirical cumulative distribution function which passes through \((m, p)\).

When \( p \) is unknown, the maximum likelihood estimate of the density is simply found by calculating the maximum likelihood estimate of the density with \( p \) assumed known and equal to \( p_0 \) and then maximizing the resulting likelihood over choices of \( p_0 \).

When \( m \) is unknown, the likelihood can approach infinity, since any function that puts a point mass on one of the data points and positive likelihood everywhere else, will have infinite likelihood. Due to this, it is appropriate to choose the mode to be at one of the data points, and to maximize the likelihood which is calculated by omitting that point from the usual likelihood product. This will not result in a full density, but will provide pointwise consistent estimates of the density at any point other than the true mode.

This density estimate (with \( m \) known) has been shown to be consistent and convergent in probability by Prakasa Rao, who also showed that it converges in distribution at a rate of \( n^{\frac{1}{2}} \) to a somewhat complicated distribution related to the heat equation.

This rate of convergence is somewhat disappointing, since it is slower than the rate of convergence of the optimal kernel estimate, which under mild regularity conditions is \( n^{\frac{3}{2}} \), but this result is not at all surprising, when one considers that the maximum likelihood estimate for the unimodal case is a piecewise constant estimate, and that its rate of convergence is the same as for other piecewise constant estimates, like the histogram. In other words, it is a hopeless task to adequately represent a continuous function by one which isn't continuous on predetermined intervals.

In the next chapter, I will discuss how such convergence results may be used to produce estimates which converge at a rapid rate when parametric assumptions are satisfied, and yet which are still consistent when those assumptions are violated.
IV. Properties of a superefficient estimate

In chapter II, a number of traditional estimation methods were discussed, as well as some Empirical Bayesian methods. The Empirical Bayesian estimates may be thought of as modifications of traditional estimates that mimic the form of Bayesian estimates, but which don’t rely on the Bayesian assumption of an underlying prior distribution on the parameter space. The resulting estimates have improved loss properties over the traditional estimates in some sense, and it would be desirable to extend the Empirical Bayesian techniques to more general estimation situations. These estimates would likewise take the form of a linear combination between a "traditional" estimate and a "special" estimate (which may be constant). Unfortunately, it isn’t possible to achieve the same sort of improved loss properties as in the Empirical Bayesian situations that were discussed in chapter II. This is because it is no longer being assumed that the traditional estimate is an unbiased estimate. In this chapter, very little will be assumed about the forms of the traditional and special estimates, since the theorems proved will rely only on the loss properties of these estimates. Because of this, I will simply refer to them as the traditional estimate and the special estimate.

Following the reasoning of the discussion of chapter II, our objective for improvement over the traditional estimate \( \hat{p} \) must be modified to that of trying to improve over \( \hat{p} \) on a small subset of the set of all possible values of \( p \) while retaining the asymptotic properties of \( \hat{p} \) for all other values of \( p \). It is because \( \hat{p} \) isn’t assumed to be unbiased that this modification is necessary. If it is possible to estimate consistently the dominant term of the bias of \( \hat{p} \), then an improved “traditional estimate” will result from subtracting the estimated bias from \( \hat{p} \). However, it will be assumed that such improvements of \( \hat{p} \) have already been carried out to the extent that it’s desirable to do so.

An estimate will be constructed which is designed to have convergence properties better than those of \( \hat{p} \) whenever the true value of \( p \) is in a particular subset \( A \).
\( \Lambda \) will first be assumed to consist of a single point \( \lambda \), and will then be generalized to the situation where, when \( p = \lambda \in \Lambda \), \( \hat{\lambda} \) is an estimate of \( \lambda \) that is preferable to \( \hat{p} \). In the context of density estimation, \( \Lambda = \{ \lambda \} \) may be interpreted as a density that corresponds to a theoretical model under scrutiny or to a density that was known to be the true underlying density at some time in the past. On the other hand, the general space \( \Lambda \) may be interpreted as the collection of the elements of a parametric family of densities and the special estimate, \( \hat{\lambda} \) may be interpreted as a parametric density estimate. In chapter V, an application of the estimate developed in this chapter will be presented, demonstrating how it may be used to choose a density estimate out of a nested series of alternatives.

The estimate considered will be a convex linear combination of the traditional estimate \( \hat{p} \) and either the known fixed point \( \lambda \) or the special estimate, \( \hat{\lambda} \).

That is,

\[
\hat{q}(\alpha) = \alpha \hat{p} + \alpha \lambda,
\]

or when \( \lambda \) isn’t fully known,

\[
\hat{q}(\alpha) = \alpha \hat{p} + \alpha \hat{\lambda},
\]

where \( \alpha = 1 - \alpha \).

This is similar in form to the Empirical Bayesian estimates discussed in chapter II, where the fixed point of the James-Stein estimate was the vector of all zeroes. However, since the bias is unknown, it is desired that \( \alpha \) be near enough to one to make the loss properties of \( \hat{q}(\alpha) \) asymptotically equivalent to those of \( \hat{p} \) whenever \( p \not\in \Lambda \). It will also be assumed that the special estimate, \( \hat{\lambda} \) is a well chosen estimate of \( p \) when \( p \in \Lambda \), so it will be desired that \( \alpha \) be near to one under those circumstances.

The estimate that will be considered is \( \hat{q}(\alpha) \), where

\[
\alpha = \frac{\psi}{\|\hat{p} - \lambda\|^2},
\]

and

\[
\psi = \min\left( N^{-c}, \|\hat{p} - \lambda\|^2 \right)
\]

whenever \( \Lambda \) consists of a single fixed point, and

\[
\tilde{\alpha} = \frac{\psi}{\|\hat{p} - \hat{\lambda}\|^2},
\]

where

\[
\psi = \min\left( N^{-c}, \|\hat{p} - \hat{\lambda}\|^2 \right)
\]
when \( \hat{\lambda} \) is an estimate of an unknown \( \lambda \in \Lambda \). The choice of the constant \( c \) will depend on the asymptotic behavior of \( \hat{\theta} \) and \( \hat{\lambda} \) and will be discussed later in conjunction with the relevant results.

Note that the first estimate is somewhat similar in form to the James-Stein estimate, since the denominator of the shrinkage constant of that estimate is also \(||X - \lambda||^2 = \sum X_i^2\). Also, the positive part rule,

\[
X \left( 1 - \frac{k - 2}{s} \right) = \max \left( 0, 1 - \frac{k - 2}{s} \right)
\]

is quite similar in form to \( \hat{\vartheta}(\alpha) \). However, since the constant \( c \) will be chosen somewhat arbitrarily within a range of values, no Bayesian or Empirical Bayesian interpretation will be placed on the estimate \( \hat{\vartheta}(\alpha) \).

The function \( N^{-c} \) in the definition of \( \psi \) is chosen out of convenience in the form of the proof. It plays the role of a dividing line between the values that the squared norm \(||\hat{\theta} - \lambda||^2 \) typically takes when \( p \in \Lambda \) and the values that it typically takes when \( p \notin \Lambda \). The proofs in this chapter can be generalized to use a more general function in place of \( N^{-c} \), but it’s thought that, for the purpose of the simplicity of this paper, the gain from this generality would be more than offset by the loss of clarity in the proofs.

It is scarcely surprising that the properties of \( \hat{\vartheta}(\alpha) \) will depend on the properties of \( \hat{\theta} \) and \( \hat{\lambda} \). Consequently, the theorems in this chapter may be summarized somewhat imprecisely by the following statement:

**When** \( p \in \Lambda \), \( \hat{\vartheta}(\alpha) \) **has the asymptotic properties** of \( \hat{\lambda} \) **and when** \( p \notin \Lambda \), \( \hat{\vartheta}(\alpha) \) **has the asymptotic properties** of \( \hat{\theta} \).

As reviewed in chapter II, the properties of estimates that will be of interest are convergence in probability, convergence in expected squared norm (mean squared error or mean integrated squared error), and convergence in distribution. Theorems 1 through 3 will deal with the problem of convergence in expected squared norm, theorem 4 will deal with convergence in distribution, and theorems 5 and 6 will deal with convergence in probability.

It can easily be shown that for estimates \( \hat{\vartheta}(\alpha) \) of the form

\[
\hat{\vartheta}(\alpha) = \alpha \hat{\theta} + \beta \lambda,
\]

\[
L(\hat{\vartheta}(\alpha), p) - L(\hat{\theta}, p) \leq E \frac{\psi^2}{||\hat{\theta} - \lambda||^2} + 2E|\psi \alpha_0|,
\]

where \( \alpha_0 = (\hat{\theta} - p, \hat{\theta} - \lambda)/||\hat{\theta} - \lambda||^2 \).

Assuming that we know that the rate of convergence of the loss of the traditional estimate, \( \hat{\theta} \) is \( N^{-2\gamma} \), (that is, \( E||\hat{\theta} - p||^2 = O(N^{-2\gamma}) \)), we have the following lemma:
Lemma 1: If \( E||\hat{p} - p||^2 = O(N^{-2g}) \), then \( \Pr\{||\hat{p} - \lambda||^2 < \epsilon\} = O(N^{-2g}) \), for all positive \( \epsilon \) such that \( \epsilon < \frac{1}{2}||p - \lambda||^2 \).

Proof:
Note that by the triangle inequality,
\[
||\hat{p} - p|| \geq ||p - \lambda|| - ||\hat{p} - \lambda||,
\]
so \( ||\hat{p} - \lambda||^2 < \epsilon \) implies that
\[
||\hat{p} - p|| > ||p - \lambda|| - \sqrt{\epsilon} > \frac{1}{2}||p - \lambda||,
\]
which is a constant.
Thus,
\[
\Pr\{||\hat{p} - \lambda||^2 < \epsilon\} < \Pr\{||\hat{p} - p||^2 > \frac{1}{4}||p - \lambda||^2\}.
\]
However,
\[
E||\hat{p} - p||^2 = \int_{(||\hat{p} - p||^2 \leq \frac{1}{4}||p - \lambda||^2)} ||\hat{p} - p||^2 dP + \int_{(||\hat{p} - p||^2 > \frac{1}{4}||p - \lambda||^2)} ||\hat{p} - p||^2 dP
\]
\[
\geq \frac{1}{4}||p - \lambda||^2 \Pr\{||\hat{p} - p||^2 > \frac{1}{4}||p - \lambda||^2\}
\]
\[
> \frac{1}{4}||p - \lambda||^2 \Pr\{||\hat{p} - \lambda||^2 < \epsilon\}.
\]
Since the left hand side of this string of inequalities is assumed to be of order \( O(N^{-2g}) \), we have
\[
\frac{1}{4}||p - \lambda||^2 \Pr\{||\hat{p} - \lambda||^2 < \epsilon\} = O(N^{-2g}),
\]
which implies that
\[
\Pr\{||\hat{p} - \lambda||^2 < \epsilon\} = O(N^{-2g}).
\]

Using this lemma, the following theorem may be proved:

Theorem 1:
If
1. \( E||\hat{p} - p||^2 = O(N^{-2g}) \), and
2. \( \hat{q}(\alpha) = \alpha \hat{p} + \bar{\alpha} \lambda \), where
\[
\bar{\alpha} = \min\left(\frac{N^{-c}}{||\hat{p} - \lambda||^2}, 1\right) \text{ and } c > g,
\]
then, when \( p \neq \lambda \),
\[
\frac{L(\hat{q}(\alpha), p) - L(\hat{p}, p)}{L(\hat{p}, p)} \to 0.
\]
proof:
As mentioned above,

\[ L(\hat{q}(\alpha), p) - L(\hat{p}, p) \leq E \frac{\psi^2}{\|\hat{p} - \lambda\|^2} + 2E|\psi\alpha_0|. \quad (1) \]

The first term of (1) is by definition equal to

\[ E \min \left( \|\hat{p} - \lambda\|^2, \frac{N^{-2c}}{\|\hat{p} - \lambda\|^2} \right). \]

Conditioning this expectation on whether \(\|\hat{p} - \lambda\|^2 \leq N^{g-c}\) or not,

\[ E \frac{\psi^2}{\|\hat{p} - \lambda\|^2} = \]

\[ E\left( \min \left( \|\hat{p} - \lambda\|^2, \frac{N^{-2c}}{\|\hat{p} - \lambda\|^2} \right) | \|\hat{p} - \lambda\|^2 \geq N^{g-c} \right) \Pr\{\|\hat{p} - \lambda\|^2 \geq N^{g-c}\} \]

\[ + \]

\[ E\left( \min \left( \|\hat{p} - \lambda\|^2, \frac{N^{-2c}}{\|\hat{p} - \lambda\|^2} \right) | \|\hat{p} - \lambda\|^2 < N^{g-c} \right) \Pr\{\|\hat{p} - \lambda\|^2 < N^{g-c}\} \leq \]

\[ \frac{N^{-2c}}{\|\hat{p} - \lambda\|^2} \]

\[ E(\|\hat{p} - \lambda\|^2 | \|\hat{p} - \lambda\|^2 < N^{g-c}) \Pr\{\|\hat{p} - \lambda\|^2 < N^{g-c}\} \leq \]

\[ N^{-(g-c)} + N^{g-c} \Pr\{\|\hat{p} - \lambda\|^2 < N^{g-c}\}. \]

However, by the Lemma, \(\Pr\{\|\hat{p} - \lambda\|^2 < N^{g-c}\} = O(N^{-2g})\), since \(c\) is greater than \(g\), so the first term of (1) is less than or equal to

\[ N^{-(g+c)} + N^{g-c}O(N^{-2g}) = o(N^{-2g}). \]

Now look at the second term of (1), namely

\[ E|\psi\alpha_0| = E \frac{|\psi\langle\hat{p} - p, \hat{p} - \lambda\rangle|}{\|\hat{p} - \lambda\|^2} \]

\[ = E \min \left( |\langle\hat{p} - p, \hat{p} - \lambda\rangle|, N^{-c} \frac{|\langle\hat{p} - p, \hat{p} - \lambda\rangle|}{\|\hat{p} - \lambda\|^2} \right). \]

38
Decomposing this expression by conditioning on whether \(\|\hat{p} - \lambda\|^2\) is less than \(N^{-(c-g)}\) or not, (2) is equal to

\[
E\left( \min \left( \frac{\|\hat{p} - p, \hat{p} - \lambda\|}{\|\hat{p} - \lambda\|^2} \right) \|\hat{p} - \lambda\|^2 < N^{-(c-g)} \right) \\
\Pr\{\|\hat{p} - \lambda\|^2 < N^{-(c-g)}\} + \\
E\left( \min \left( \frac{\|\hat{p} - p, \hat{p} - \lambda\|}{\|\hat{p} - \lambda\|^2} \right) \|\hat{p} - \lambda\|^2 \geq N^{-(c-g)} \right) \\
\Pr\{\|\hat{p} - \lambda\|^2 \geq N^{-(c-g)}\}
\]

(3)

The first term of (3) is less than or equal to

\[
E(\|\hat{p} - p\|\|\hat{p} - \lambda\|\|\hat{p} - \lambda\|^2 < N^{-(c-g)})\Pr\{\|\hat{p} - \lambda\|^2 < N^{-(c-g)}\} < \\
N^{-\frac{1}{2}(c-g)}E(\|\hat{p} - p\|\|\hat{p} - \lambda\|^2 < N^{-(c-g)})\Pr\{\|\hat{p} - \lambda\|^2 < N^{-(c-g)}\}.
\]

However, by the triangle inequality, \(\|\hat{p} - p\| \leq \|p - \lambda\| + \|\hat{p} - \lambda\|\), so the first term of (3) is less than or equal to

\[
N^{-(\frac{1}{2}c-g)}(\|p - \lambda\| + \\
E(\|\hat{p} - \lambda\|\|\hat{p} - \lambda\|^2 < N^{-(c-g)})\Pr\{\|\hat{p} - \lambda\|^2 < N^{-(c-g)}\} \leq \\
N^{-(\frac{1}{2}c-g)}(\|p - \lambda\| + N^{-(\frac{1}{2}c-g)})\Pr\{\|\hat{p} - \lambda\|^2 < N^{-(c-g)}\} = \\
N^{-(\frac{1}{2}c-g)}(\|p - \lambda\| + o(1))O(N^{-2g}) = O(N^{-2g-\frac{1}{2}(c-g)}) = o(N^{-2g}).
\]

Similarly, the second term of (3) is less than or equal to

\[
E\left( N^{-c}\frac{\|\hat{p} - p, \hat{p} - \lambda\|}{\|\hat{p} - \lambda\|^2} \|\hat{p} - \lambda\|^2 > N^{-(c-g)} \right) \\
\Pr\{\|\hat{p} - \lambda\|^2 > N^{-(c-g)}\} \leq \\
N^{-c+\frac{1}{2}(c-g)}\int_{\|\hat{p} - \lambda\|^2 > N^{-(c-g)}} \|\hat{p} - p\| \, dP \leq \\
N^{-c+\frac{1}{2}(c-g)}\int \|\hat{p} - p\| \, dP,
\]

which, by the Cauchy-Schwartz inequality, is less than or equal to

\[
N^{-c+\frac{1}{2}(c-g)}\sqrt{\int \|\hat{p} - p\|^2 \, dP}.
\]
By the assumed rate of convergence of $L(\hat{\beta}, \beta)$, this last expression is

$$N^{-c+\frac{1}{2}(c-g)}O(N^{-g}) = O(N^{-2g-\frac{1}{2}(c-g)}) = o(N^{-2g}).$$

Thus, since each term of (3) is of order $o(N^{-2g})$, as is the first term of (1), we have that (1) is $o(N^{-2g})$. Thus,

$$L(\hat{q}(\alpha), p) - L(\hat{\beta}, p) = o(N^{-2g}),$$

so

$$\frac{L(\hat{q}(\alpha), p) - L(\hat{\beta}, p)}{L(\hat{\beta}, p)} \rightarrow 0.$$

It is now desired to extend these results to the situation where $\lambda$ isn't known, but is an unknown point in the set $\Lambda$ and which must be estimated from the observed data. The proof of the next theorem is virtually identical to that of Theorem 1, but will be included in its entirety for the sake of completeness. As with Theorem 1, a preparatory lemma is first necessary:

**Lemma 2:** If $E\|\hat{p} - p\|^2 = O(N^{-2g})$ and $E\|\hat{\lambda} - \lambda\|^2 = O(N^{-2g})$, where $\gamma > g$, then whenever $\epsilon < \frac{\|\hat{p} - \lambda\|}{3}, \Pr\{\|\hat{p} - \hat{\lambda}\|^2 < \epsilon^2\} = O(N^{-2g}).$

**proof:** By the triangle inequality,

$$\|\hat{p} - \hat{\lambda}\| \geq \|\hat{p} - \lambda\| - \|\hat{\lambda} - \lambda\| \geq \|p - \lambda\| - \|\hat{\lambda} - \lambda\| - \|\hat{p} - p\|,$$

so

$$\|\hat{p} - \hat{\lambda}\|^2 < \epsilon^2$$

implies that

$$\|\hat{\lambda} - \lambda\| + \|\hat{p} - p\| \geq 2\epsilon,$$

which implies that either

$$\|\hat{\lambda} - \lambda\| \geq \epsilon \text{ or } \|\hat{p} - p\| \geq \epsilon.$$

Thus,

$$\Pr\{\|\hat{p} - \hat{\lambda}\|^2 < \epsilon^2\} \leq \Pr\{\|\hat{\lambda} - \lambda\| \geq \epsilon \text{ or } \|\hat{p} - p\| \geq \epsilon\} \leq \Pr\{\|\hat{\lambda} - \lambda\| \geq \epsilon\} + \Pr\{\|\hat{p} - p\| \geq \epsilon\}.\]
However,
\[
E\|\hat{\lambda} - \lambda\|^2 = \int \|\hat{\lambda} - \lambda\|^2 dP \\
\geq \int_{\|\hat{\lambda} - \lambda\|^2 \geq \epsilon^2} \|\hat{\lambda} - \lambda\|^2 dP \\
\geq \epsilon^2 \Pr\{|\|\hat{\lambda} - \lambda\|^2 \geq \epsilon^2\}.
\]

The left hand side of this expression is by assumption \(O(N^{-2\gamma})\), so the right hand side must be \(O(N^{-2\gamma})\) as well. Thus, \(\Pr\{|\|\hat{\lambda} - \lambda\|^2 \geq \epsilon^2\} = O(N^{-2\gamma})\).

Similarly, \(E\|\hat{\lambda} - \lambda\|^2 \geq \epsilon^2 \Pr\{|\|\hat{\lambda} - \lambda\|^2 \geq \epsilon^2\} = O(N^{-2\gamma})\) as well.

Thus, \(\Pr\{|\|\hat{\lambda} - \hat{\lambda}\|^2 < \epsilon^2\} \leq O(N^{-2\gamma}) + O(N^{-2\gamma}) = O(N^{-2\gamma})\).

Now I am ready to prove this analog of Theorem 1:

**Theorem 2:** If \(\hat{\hat{q}}(\alpha) = \alpha \hat{p} + \alpha \lambda\), where

1. \(E\|\hat{\lambda} - \lambda\|^2 = O(N^{-2\gamma})\)
2. \(0 < g < \gamma\)
3. \(\hat{\alpha} = \frac{\psi}{\|\hat{p} - \hat{\lambda}\|^2}\), where \(\psi = \min(N^{-c}, \|\hat{p} - \hat{\lambda}\|^2)\) and finally,
4. \(g < c\),
5. \(\hat{\alpha} = \psi\|\hat{p} - \hat{\lambda}\|^2\),

then, when \(p \neq \lambda\),
\[
\frac{L(\hat{\hat{q}}(\alpha), p) - L(\hat{p}, p)}{L(\hat{p}, p)} \rightarrow 0.
\]

**Proof:**

As in the proof of Theorem 1, the difference in loss between \(\hat{\hat{q}}(\alpha)\) and \(\hat{p}\) may be written as follows:
\[
L(\hat{\hat{q}}(\alpha), p) - L(\hat{p}, p) = E\hat{\alpha}^2\|\hat{p} - \hat{\lambda}\|^2 - 2E\langle \hat{\hat{q}} - p, \hat{\lambda} \rangle \leq
\]
\[
E\frac{\psi^2}{\|\hat{p} - \hat{\lambda}\|^2} + 2E\frac{|\psi(\hat{p} - p, \hat{\lambda} - \hat{\lambda})|}{\|\hat{p} - \hat{\lambda}\|^2}.
\tag{4}
\]

The first term of (4) is equal to:
\[
E\left(\min\left(\frac{N^{-c}}{\|\hat{p} - \hat{\lambda}\|^2}, \|\hat{\lambda} - \hat{\lambda}\|^2\right)\right).
\tag{5}
\]
Conditioning on whether \( \| \hat{\beta} - \hat{\lambda} \|^2 \) is less than \( N^{-(c-g)} \) or not, (5) is equal to:

\[
E \left( \min \left( \frac{N^{-2c}}{\| \hat{\beta} - \hat{\lambda} \|^2}, \| \hat{\beta} - \hat{\lambda} \|^2 \right) \bigg| \| \hat{\beta} - \hat{\lambda} \|^2 \geq N^{-(c-g)} \right) \right) + \\
\Pr\{\| \hat{\beta} - \hat{\lambda} \|^2 \geq N^{-(c-g)}\} \\
E \left( \min \left( \frac{N^{-2c}}{\| \hat{\beta} - \hat{\lambda} \|^2}, \| \hat{\beta} - \hat{\lambda} \|^2 \right) \bigg| \| \hat{\beta} - \hat{\lambda} \|^2 < N^{-(c-g)} \right) \right) \leq \\
\Pr\{\| \hat{\beta} - \hat{\lambda} \|^2 < N^{-(c-g)}\} \\
E \left( \frac{N^{-2c}}{\| \hat{\beta} - \hat{\lambda} \|^2} \bigg| \| \hat{\beta} - \hat{\lambda} \|^2 \geq N^{-(c-g)} \right) \Pr\{\| \hat{\beta} - \hat{\lambda} \|^2 \geq N^{-(c-g)}\} + \\
E\{\| \hat{\beta} - \hat{\lambda} \|^2 \| \hat{\beta} - \hat{\lambda} \|^2 \bigg| \| \hat{\beta} - \hat{\lambda} \|^2 < N^{-(c-g)} \} \Pr\{\| \hat{\beta} - \hat{\lambda} \|^2 < N^{-(c-g)}\} \leq \\
N^{-(c+g)} + N^{-(c-g)} \Pr\{\| \hat{\beta} - \hat{\lambda} \|^2 < N^{-(c-g)}\}. \\
\]

However, for

\[
N > \left( \frac{\| p - \lambda \|^2}{9} \right)^{g-c}, N^{-(c-g)} < \frac{\| p - \lambda \|^2}{9},
\]

so, applying Lemma 2, equation (5) is less than or equal to

\[
N^{-(c+g)} + N^{-(c-g)} O(N^{-2g}) = O(N^{-(c+g)}) = o(N^{-2g}).
\]

The second term of (4) is equal to

\[
E \left( \min \left( \frac{N^{-c}||\hat{\beta} - p, \hat{\lambda}||}{\| \hat{\beta} - \hat{\lambda} \|^2}, \|\hat{\beta} - p, \hat{\lambda}|| \right) \right) \leq \\
E \left( \min \left( \frac{N^{-c}||\hat{\beta} - p||}{\| \hat{\beta} -\lambda \|}, \|\hat{\beta} - p||\|\hat{\beta} - \hat{\lambda}|| \right) \right). \\
\]

(6)

Conditioning on whether \( \| \hat{\beta} - \hat{\lambda} \| \) is less than \( N^{-(c-g)/2} \) or not, (6) is equal to

\[
E \left( \min \left( \frac{N^{-c}||\hat{\beta} - p||}{\| \hat{\beta} - \hat{\lambda} \|}, \|\hat{\beta} - p||\|\hat{\beta} - \hat{\lambda}|| \right) \bigg| \| \hat{\beta} - \hat{\lambda} \| \geq N^{-(c-g)/2} \right) \right) + \\
\Pr\{\| \hat{\beta} - \hat{\lambda} \| \geq N^{-(c-g)/2}\} \\
E \left( \min \left( \frac{N^{-c}||\hat{\beta} - p||}{\| \hat{\beta} - \hat{\lambda} \|}, \|\hat{\beta} - p||\|\hat{\beta} - \hat{\lambda}|| \right) \bigg| \| \hat{\beta} - \hat{\lambda} \| < N^{-(c-g)/2} \right) \right) \right) \leq \\
\Pr\{\| \hat{\beta} - \hat{\lambda} \| < N^{-(c-g)/2}\}. \\
\]

(7)
The first term of (7) is less than or equal to:

\[
E\left( N^{-c} \frac{||\hat{p} - p||}{N^{-(c-g)/2}} \left| ||\hat{p} - \lambda|| \geq N^{-(c-g)/2} \right\} \leq \right.
\]

\[
N^{-(c+g)/2} \sqrt{\int (||\hat{p} - \hat{\lambda}|| \geq N^{-(c-g)/2}) ||\hat{p} - p||^2 dP} \leq
\]

\[
N^{-(c+g)/2} \sqrt{L(\hat{p}, p)} = o(N^{-2g}).
\]

The second term of (7) is less than or equal to:

\[
N^{-(c-g)/2} \int (||\hat{p} - \hat{\lambda}|| < N^{-(c-g)/2}) ||\hat{p} - p|| dP \leq
\]

\[
N^{-(c-g)/2} \int (||\hat{p} - \hat{\lambda}|| < N^{-(c-g)/2}) \cap (||\hat{p} - p|| < ||p - \lambda||^2) ||\hat{p} - p|| dP +
\]

\[
N^{-(c-g)/2} \int (||\hat{p} - p|| \geq ||p - \lambda||^2) ||\hat{p} - p|| dP \leq
\]

\[
N^{-(c-g)/2} ||p - \lambda||^2 \Pr\{||\hat{p} - \hat{\lambda}|| < N^{-(c-g)/2}\} +
\]

\[
N^{-(c-g)/2} \frac{2}{||p - \lambda||} \int ||\hat{p} - p||^2 dP =
\]

\[
||p - \lambda||^2 N^{-(c-g)/2} O(N^{-2g}) + \frac{2}{||p - \lambda||} N^{-(c-g)/2} L(\hat{p}, p) =
\]

\[
O(N^{-2g+(c-g)/2}) + \frac{2}{||p - \lambda||} N^{-(c-g)/2} O(N^{-2g}) =
\]

\[
= o(N^{-2g}).
\]

Therefore, \( L(\hat{q}(\alpha), p) - L(\hat{p}, p) = o(N^{-2g}) \), so

\[
\frac{L(\hat{q}(\alpha), p) - L(\hat{p}, p)}{L(\hat{p}, p)} \rightarrow 0.
\]

The problem with mean squared error or expected squared norm as a criterion is that the asymptotic behavior of the expected squared norm needn’t have any particular relation to the observed squared norm. That is, a large error in an estimate on a set of small probability can change the convergence or at least the
rate of convergence in expected squared norm of an estimate. Since the proposed estimate depends not on a measure of expected squared norm, but only on the observed squared norm, its resulting behavior will be the same as was anticipated only if the observed squared norm is of the same order of magnitude as the expected squared norm. Thus, when \( p = \lambda \), neither the estimate that is a convex combination of \( \hat{p} \) and \( \lambda \) nor the estimate that is a convex combination of \( \hat{p} \) and \( \hat{\lambda} \) will be able to remove the error in \( \hat{p} \) whenever it is of a larger order of magnitude than \( \sqrt{E||\hat{p} - p||^2} \). This statement is made precise in the following theorem.

**Theorem 3:**

If \( \hat{q}(\alpha) = \alpha \hat{p} + (1-\alpha)\hat{\lambda} \), where

1. \( \alpha = \min\left(\frac{N^{-c}}{||\hat{p} - \lambda||^2}, 1\right) \),
2. \( E||\hat{p} - p|| = O(N^{-2g}) \), and
3. \( c > g \)

then, assuming that \( p = \lambda \), for any \( \delta > 0 \), the portion of \( \mathcal{L}(\hat{q}(\alpha), p) \) that occurs on the set \( \{||\hat{p} - p||^2 < N^{-(c+\delta)}\} \) is less than or equal to \( \mathcal{L}(\hat{\lambda}, p) \) for \( N \) sufficiently large.

In order to prove this theorem, the following Lemma is first necessary:

**Lemma 3:**

If \( ||\hat{p} - p||^2 < ||\hat{\lambda} - p||^2 \), then, for any \( \alpha \in (0, 1) \),

\[
||\hat{q}(\alpha) - p||^2 < ||\hat{\lambda} - p||^2.
\]

**Proof of the Lemma:**

\[
||\hat{\lambda} - p||^2 - ||\hat{q}(\alpha) - p||^2 = -||\hat{p} - \hat{\lambda}||^2 \alpha^2 - 2(\hat{p} - \hat{\lambda}, \hat{\lambda} - p)\alpha.
\]

Call this quantity \( f(\alpha) \). \( f(\alpha) \) is quadratic in \( \alpha \) and the coefficient of \( \alpha^2 \) is negative, so \( f(\alpha) \) is positive if and only if \( \alpha \) lies in the open interval whose endpoints are its two roots. Since \( f(0) = 0 \), 0 is one of the endpoints of this interval. Likewise, since

\[
f(1) = ||\hat{\lambda} - p||^2 - ||\hat{p} - p||^2,
\]

which is greater than zero by assumption, the second root of \( f(\alpha) \) must be greater than 1. Thus, for \( \alpha \in (0, 1) \), \( f(\alpha) > 0 \). This is equivalent to the statement of the lemma. 

44
Proof of Theorem 3:
If $N > (4)^{1/6}$, then $4N^{-(c+\epsilon)} < N^{-c}$. Note that $\alpha = 0$ whenever

$$||\hat{p} - \hat{\lambda}||^2 \leq N^{-c},$$

so that $\alpha \neq 0$ implies that $||\hat{p} - \hat{\lambda}|| > N^{-\frac{c}{2}}$.

Thus,

$$\int_{\{||\hat{p} - p||^2 < N^{-(c+\epsilon)}\}} ||\hat{q}(\alpha) - p||^2 dP =$$

$$\int_{\{||\hat{p} - p||^2 < N^{-(c+\epsilon)}\}} ||\hat{\lambda} - p||^2 dP.$$

Also, for $N > (4)^{1/6}$ and $\alpha \neq 0$,

$$||\hat{\lambda} - \lambda|| \geq ||\hat{p} - \hat{\lambda}|| - ||\hat{p} - p|| >$$

$$N^{-\frac{c}{2}} - N^{-\frac{c+\epsilon}{2}} > N^{-\frac{c+\epsilon}{2}} > ||\hat{p} - p||.$$

Thus, by Lemma 3,

$$\int_{\{||\hat{p} - p||^2 < N^{-(c+\epsilon)}\}} ||\hat{q}(\alpha) - p||^2 dP \leq$$

$$\int_{\{||\hat{p} - p||^2 < N^{-(c+\epsilon)} \text{ and } \alpha = 0\}} ||\hat{\lambda} - p||^2 dP +$$

$$\int_{\{||\hat{p} - p||^2 < N^{-(c+\epsilon)} \text{ and } \alpha \neq 0\}} ||\hat{\lambda} - p||^2 dP =$$

$$\int_{\{||\hat{p} - p||^2 < N^{-(c+\epsilon)}\}} ||\hat{\lambda} - p||^2 dP \leq$$

$$\int ||\hat{\lambda} - p||^2 dP = L(\hat{\lambda}, p)$$

The interpretation of this theorem is important in that it provides the main criterion by which to choose the constant $c$. Since it makes a statement only about the error that occurs on the set where $||\hat{p} - p||^2$ is smaller than $N^{-(c+\epsilon)}$, it is desirable to have that set be as large as possible in order to have as little additional error as possible occurring on the set where the squared norm of $\hat{p} - p$ is large. Thus, choices of $c$ that are close to $g$ (or $\gamma$) will improve the behavior of $\hat{q}(\alpha)$ more than larger choices in the situation that $p = \lambda$. However, when $p \not\in \Lambda$, giving a positive
weight to \( \hat{\lambda} \) will typically increase the loss of the estimate. It is only as the weight given to \( \hat{\lambda} \) goes to zero that this increase will become negligible. Since the weight is proportional to \( N^{-c} \), the smaller \( c \) is, the slower the increase in loss will become negligible.

I will now look at the properties of \( \hat{q}(\alpha) \) when \( \hat{p} \) and \( \hat{\lambda} \) are assumed to converge in distribution at some rate to a known distribution. It will be necessary for the proof to assume also that the expected losses of the estimates also converge at the same rates. This assumption will not necessarily hold, but will be satisfied in many well-behaved situations. Since the notion of convergence in distribution is defined only for a finite dimensional entity and nothing has been assumed about the form of the Hilbert space in which \( p \) lies, the following theorem will deal with the convergence in distribution of a vector valued linear functional of \( \hat{q}(\alpha) \). In keeping with the principal example of density estimation, in which the most interesting linear functional is the one that evaluates the density at an arbitrary point \( u \), I will denote the functional, evaluated at the point \( p \) by \( p(u) \). However, the fact that the evaluation functional is univariate is not necessary to the proof of the theorem.

**Theorem 4:**

If

(i). \( N_\theta(\hat{p}(u) - p(u)) \overset{d}{\rightarrow} F_1 \),
(ii). \( E||\hat{p} - p||^2 = O(N^{-2\theta}) \),
(iii). \( N_\gamma(\hat{\lambda}(u) - \lambda(u)) \overset{d}{\rightarrow} F_2 \),
(iv). \( E||\hat{\lambda} - \lambda||^2 = O(N^{-2\gamma}) \),
and,
(v). \( \hat{q}(\alpha) = \alpha \hat{p} + \bar{\alpha} \hat{\lambda} \),

where

\[
\bar{\alpha} = \min\left( \frac{N^{-c}}{||\hat{p} - \hat{\lambda}||^2}, 1 \right),
\]

\( g < c < 2g \), and \( \gamma > g \), then

I. When \( p \neq \lambda \), \( N_\theta(\hat{q}(\alpha; u) - p(u)) \overset{d}{\rightarrow} F_1 \)

II. When \( p = \lambda \), \( N_\gamma(\hat{q}(\alpha; u) - p(u)) \overset{d}{\rightarrow} F_2 \).

When \( \gamma = \infty \), property II. is of the form zero times infinity. This should be interpreted as meaning that II. holds for arbitrarily large values of \( \gamma \), leading to the appropriate conclusion in this case, namely that, when \( \gamma \) is infinite, II holds for arbitrarily large values of \( \gamma \).

proof of I:

First note that if \( |\hat{\lambda}(u) - \lambda(u)| \leq 1 \), \( |\hat{p}(u) - p(u)| \leq 1 \), and

\[
\bar{\alpha} \leq \frac{N^{-g\epsilon}}{2 + |p(u) - \lambda(u)|},
\]

46
then
\[ N^g \tilde{z}(u) - \hat{p}(u) \leq \epsilon. \]

This implies that
\[
\Pr\{N^g |\hat{\lambda}(u) - \hat{p}(u)| > \epsilon\} \leq \Pr\{|\hat{\lambda}(u) - \lambda(u)| > 1\} + \Pr\{\|\hat{p}(u) - p(u)\| > 1\} + \Pr\{\alpha > \frac{N^{-g\epsilon}}{1 + |p(u) - \lambda(u)|}\}. 
\]

However, for any \( \epsilon' > 0 \), there exists a \( K_{\epsilon'} \) such that
\[ F_1(-K_{\epsilon'}, K_{\epsilon'}) > 1 - \epsilon', \]
and
\[ F_2(-K_{\epsilon'}, K_{\epsilon'}) > 1 - \epsilon'. \]

Thus, for \( N \) large enough, \( \Pr\{|\hat{p}(u) - p(u)| > 1\} \leq \frac{1}{1 - F_1(-K_{\epsilon'}, K_{\epsilon'})} \)
\[ = \frac{1}{1 - F_2(-K_{\epsilon'}, K_{\epsilon'})} \]

Also, for the same \( N \), \( \Pr\{|\hat{\lambda}(u) - \lambda(u)| > 1\} \leq \frac{1}{1 - F_2(-K_{\epsilon'}, K_{\epsilon'})} \)

Thus, \( \Pr\{|\hat{p}(u) - p(u)| > 1\} \to 0 \), and \( \Pr\{|\hat{\lambda}(u) - \lambda(u)| > 1\} \to 0 \).

Now,
\[
\Pr\{\alpha > \frac{N^{-g\epsilon}}{2} + |p(u) - \lambda(u)|\} \leq \Pr\left\{\frac{N^{-g\epsilon}}{\|\hat{p} - \hat{\lambda}\|^2} > \frac{N^{-g\epsilon}}{2} + |p(u) - \lambda(u)|\right\} = \Pr\left\{\|\hat{p} - \hat{\lambda}\|^2 < N^{-(c-g)} \left(\frac{2 + |p(u) - \lambda(u)|}{\epsilon}\right)\right\}.
\]

Since \( c > g \), we know that for \( N \) large enough,
\[ N^{-(c-g)} \left(\frac{2 + |p(u) - \lambda(u)|}{\epsilon}\right) < \frac{\|p - \lambda\|^2}{4}. \]

For such an \( N \),
\[
\Pr\{\alpha > \frac{N^{-g\epsilon}}{2 + |p(u) - \lambda(u)|}\} \leq \Pr\left\{\|\hat{p} - \hat{\lambda}\| < \frac{\|p - \lambda\|}{2}\right\}.
\]

47
However, since
\[ \|\hat{p} - \hat{\lambda}\| \geq \|p - \lambda\| + \|\hat{p} - p\| - \|\hat{\lambda} - \lambda\|, \]
we see that \(\|\hat{p} - \hat{\lambda}\| < \frac{\|p - \lambda\|}{2}\) implies that either
\[ \|\hat{p} - p\| > \frac{\|p - \lambda\|}{4}, \]
or
\[ \|\hat{\lambda} - \lambda\| > \frac{\|p - \lambda\|}{4}. \]

Thus,
\[
\Pr\left\{\hat{\alpha} > \frac{N^{-g} \epsilon}{2 + |p(u) - \lambda(u)|}\right\} \leq \Pr\left\{\|\hat{p} - p\|^2 > \frac{\|p - \lambda\|^2}{16}\right\} + \Pr\left\{\|\hat{\lambda} - \lambda\|^2 > \frac{\|p - \lambda\|^2}{16}\right\}.
\]

These two terms are of order \(N^{-2g}\) and \(N^{-2\gamma}\) respectively, due to the assumptions that \(E\|\hat{p} - p\|^2 = O(N^{-2g})\) and \(E\|\hat{\lambda} - \lambda\|^2 = O(N^{-2\gamma})\), and due to the assumption that \(p \not= \lambda\).

Thus,
\[
\Pr\left\{\hat{\alpha} > \frac{N^{-g} \epsilon}{2 + |p(u) - \lambda(u)|}\right\} \to 0,
\]
which implies that
\[
\Pr\{N^g\hat{\alpha}|\hat{\lambda}(u) - \hat{p}(u)| > \epsilon\} \to 0,
\]
for all positive \(\epsilon\). In other words, \(N^g\hat{\alpha}(\hat{\lambda}(u) - \hat{p}(u))\) converges to zero in probability.

Thus, by Slutsky’s Theorem, since \(N^g(\hat{p}(u) - p(u)) \overset{d}{\to} F_1\) by assumption,
\[
N^g(\hat{p}(u) - p(u)) + N^g\hat{\alpha}(\hat{\lambda}(u) - \hat{p}(u)) \overset{d}{\to} F_1.
\]

The left hand side of this final expression is
\[
N^g(\alpha \hat{p}(u) + \hat{\alpha}\hat{\lambda}(u) - p(u)) = N^g(\hat{q}(\alpha; u) - p(u)).
\]

Thus, \(N^g(\hat{q}(\alpha; u) - p(u)) \overset{d}{\to} F_1\), as desired.
Proof of II: As above, since \( p = \lambda \),

\[
N^\gamma (\hat{q}(\alpha; u) - p(u)) = N^\gamma (\hat{\lambda}(u) - \lambda(u)) + N^\gamma \alpha (\hat{p}(u) - \hat{\lambda}(u)).
\]

If it can be shown that \( N^\gamma \alpha (\hat{p}(u) - \hat{\lambda}(u)) \) converges to 0 in probability, then, applying Slutsky's Theorem again, since

\[
N^\gamma (\hat{\lambda}(u) - \lambda(u)) \xrightarrow{d} F_2,
\]

we would have

\[
N^\gamma (\hat{q}(\alpha; u) - p(u)) \xrightarrow{d} F_2.
\]

However,

\[
\Pr\{N^\gamma \alpha |\hat{p}(u) - \hat{\lambda}(u)| > \varepsilon\} \leq \Pr\{\alpha \neq 0\} = \Pr\{|\hat{p} - \hat{\lambda}| > N^{-\frac{c}{4}}\}.
\]

Notice that, since \( p = \lambda \),

\[
|\hat{p} - \hat{\lambda}| > N^{-\frac{c}{4}}
\]

implies that

\[
N^{-\frac{c}{4}} < |\hat{p} - \hat{\lambda}| < |\hat{p} - p| + |\lambda - \lambda|,
\]

which implies that either

\[
|\hat{p} - p|^2 > \frac{N^{-c}}{4},
\]

or

\[
|\hat{\lambda} - \lambda|^2 > \frac{N^{-c}}{4}.
\]

Thus, \( \Pr\{\alpha \neq 0\} \leq \)

\[
\Pr\left\{|\hat{p} - p|^2 > \frac{N^{-c}}{4}\right\} + \Pr\left\{|\hat{\lambda} - \lambda|^2 > \frac{N^{-c}}{4}\right\}.
\]

However, since \( E|\hat{p} - p|^2 = \)

\[
\int_{|\hat{p} - p|^2 \leq \frac{1}{4} N^{-c}} |\hat{p} - p|^2 dP + \int_{|\hat{p} - p|^2 > \frac{1}{4} N^{-c}} |\hat{p} - p|^2 dP \geq \frac{N^{-c}}{4} \Pr\left\{|\hat{p} - p|^2 > \frac{N^{-c}}{4}\right\},
\]

49
which implies that

$$\text{Pr}\left\{ \|\hat{p} - p\|^2 > \frac{N^{-c}}{4} \right\} \leq 4N^c E\|\hat{p} - p\|^2 = o(1).$$

Similarly,

$$\text{Pr}\left\{ \|\hat{\lambda} - \lambda\|^2 > \frac{N^{-c}}{4} \right\} \leq 4N^c E\|\hat{\lambda} - \lambda\|^2 = o(1).$$

Hence $\text{Pr}\{\alpha \neq 0\} \to 0$, which implies that

$$\text{Pr}\{N^\gamma \alpha |\hat{p}(u) - \hat{\lambda}(u)| > \epsilon\} \to 0,$$

so $N^\gamma \alpha |\hat{p}(u) - \hat{\lambda}(u)|$ converges to zero in probability, as desired. Thus, by Slutsky's Theorem, $N^\gamma (\hat{\gamma}(\alpha; u) - p(u))$ converges in distribution to $F_2$.

The final two theorems deal with the convergence in probability of the estimate $\hat{\gamma}(\alpha)$, both in the case where $\lambda$ is known and in the case where it is estimated by $\hat{\lambda}$.

**Theorem 5:** If

1. For all $\epsilon > 0$ there exist $N_\epsilon$ and $K_\epsilon$, such that, $N > N_\epsilon$ implies that

$$\text{Pr}\{\|\hat{p} - p\|^2 > K_\epsilon N^{-2\gamma}\} < \epsilon,$$

and if

2. For all $\epsilon_1 > 0$ there exist $N_{\epsilon_1}^*$ and $K_{\epsilon_1}^*$, such that $N > N_{\epsilon_1}^*$ implies that

$$\text{Pr}\{\|\hat{\lambda} - \lambda\|^2 > K_{\epsilon_1}^* N^{-2\gamma}\} < \epsilon_1,$$

where $\gamma > g$ and $\lambda \neq p$, then for any $\epsilon$ and $\delta > 0$, and for

$$N > \max\left(N_\epsilon, N_{\epsilon_1}^*, \left(\frac{3}{\delta\|p - \lambda\|}\right)^{1/(\epsilon - g)}, \left(\frac{9K_\epsilon}{\|p - \lambda\|^2}\right)^{1/2g}\right),$$

$$\text{Pr}\{\|\hat{\gamma}(\alpha) - p\|^2 > KN^{-2g}\} < \epsilon + \epsilon_1,$$

where

$$\hat{\gamma}(\alpha) = \alpha \hat{p} + \hat{\alpha} \hat{\lambda}, \hat{\alpha} = \min\left(\frac{N^{-c}}{\|\hat{p} - \hat{\lambda}\|^2}, 1\right), g < c,$$

and $K$ satisfies $\sqrt{K} = \sqrt{K_\epsilon} + \delta$.

**Corollary 1:**

If $\|\hat{p} - p\|^2 = O_p(N^{-2g})$ and $\|\hat{\lambda} - \lambda\|^2 = O_p(N^{-2\gamma})$, where $\lambda \neq p$ and $\gamma > g$, then

$$\|\hat{\gamma}(\alpha) - p\|^2 = O_p(N^{-2g}).$$
Note that the corollary follows immediately from Theorem 5 by definition, but that Theorem 5 is somewhat stronger, since it implies that in satisfying the probabilistic convergence, one may use nearly the same constant $K_\epsilon$ as in the definition of $O_\epsilon$ convergence when $\epsilon$ is augmented by an arbitrarily small amount.

**Proof of Theorem 5:**

Note the following chain of implications:

$$||\hat{q}(\alpha) - p||^2 > K N^{-2g}$$

whenever

$$||\hat{p} - p + \hat{\alpha}(\hat{\lambda} - \hat{p})|| > \sqrt{K} N^{-g},$$

which implies that either

1. $||\hat{p} - p|| > (\sqrt{K} - \delta) N^{-g} = \sqrt{K_\epsilon} N^{-g},$ or
2. $\hat{\alpha}||\hat{\lambda} - \hat{p}|| > \delta N^{-g}$ and $||\hat{p} - p|| \leq \sqrt{K_\epsilon} N^{-g}.$

Statement (2) implies that, since

$$\hat{\alpha} = \min\left(\frac{N^{-c}}{||\hat{p} - \hat{\lambda}||^2}, 1\right),$$

$$||\hat{p} - \hat{\lambda}|| < N^{(g-c)/\delta}$$

and

$$||\hat{p} - p||^2 \leq K_\epsilon N^{-2g}.$$ 

Since, by assumption,

$$N \geq \left(\frac{3}{||p - \lambda||\delta}\right)^{1/(c-g)},$$

$$||\hat{p} - \hat{\lambda}|| < \frac{1}{\delta}\left(\frac{3}{||p - \lambda||\delta}\right)^{-1} = \frac{||p - \lambda||}{3}.$$ 

Likewise, since it assumed that

$$N \geq \left(\frac{9K_\epsilon}{||p - \lambda||^2}\right)^{1/2g},$$

$$||\hat{p} - p||^2 \leq \frac{||p - \lambda||^2}{9},$$

51
so
\[ \|\hat{p} - p\| \leq \frac{\|p - \lambda\|}{3}. \]

Thus, (2) implies that
\[ \|\hat{\lambda} - \hat{p}\| < \frac{\|p - \lambda\|}{3}, \]
and
\[ \|\hat{p} - p\| \leq \frac{\|p - \lambda\|}{3}, \]
so
\[ \|\hat{\lambda} - \lambda\| \geq \|p - \lambda\| - \|\hat{p} - p\| - \|\hat{\lambda} - \hat{p}\| \geq \frac{\|p - \lambda\|}{3}. \]

Thus, \(\|\hat{q}(\alpha) - p\|^2 > KN^{-2g}\) implies that either

1. \(\|\hat{p} - p\|^2 > K\varepsilon N^{-2g}\)
2. \(\|\hat{\lambda} - \lambda\| > \frac{\|p - \lambda\|}{3}\).

Since it was assumed that
\[ N > \max \left\{ \left( \frac{3}{\|p - \lambda\|} \right)^{1/(c-g)}, \left( \frac{9K\varepsilon}{\|p - \lambda\|^2} \right)^{1/2g} \right\}, \]
it follows that
\[ \Pr\{\|\hat{q}(\alpha) - p\|^2 > KN^{-2g}\} \leq \Pr\{\|\hat{p} - p\|^2 > K\varepsilon N^{-2g}\} + \Pr\{\|\hat{\lambda} - \lambda\|^2 > \frac{\|p - \lambda\|^2}{9}\}, \]
which, since \(N > N\varepsilon\), is less than or equal to
\[ \varepsilon + \Pr\{\|\hat{\lambda} - \lambda\|^2 > \frac{\|p - \lambda\|^2}{9}\}. \]

Since
\[ N \geq \left( \frac{9K\varepsilon^*_1}{\|p - \lambda\|^2} \right)^{1/2g}, \]
this probability is less than or equal to
\[ \varepsilon + \Pr\{\|\hat{\lambda} - \lambda\|^2 > K\varepsilon^*_1 N^{-2g}\}, \]
which, since \(N > N\varepsilon^*_1\) is less than or equal to \(\varepsilon + \varepsilon_1\).
Note that the following theorem is an analog to Theorem 5 in that it shows that, when \( p = \lambda \), \( \hat{\alpha}(\alpha) \) possesses the \( O_p \) convergence properties of \( \hat{\lambda} \), with the added result that the same \( K^*_{\epsilon} \) may be used with arbitrarily small impact on the \( \epsilon \) of the definition. Therefore, the corollary (Corollary 2) will follows directly from Theorem 6.

**Theorem 6:**

If

(I). for all \( \epsilon > 0 \), there exist \( N_{\epsilon} \) and \( K_{\epsilon} \) such that, for \( N > N_{\epsilon} \),

\[
\Pr\{||\hat{p} - p||^2 > K_{\epsilon} N^{-2\gamma}\} < \epsilon,
\]

and

(II). for all \( \epsilon > 0 \), there exist \( N_{\epsilon}^* \) and \( K_{\epsilon}^* \) such that, for \( N > N_{\epsilon}^* \), where \( \gamma > g \) and \( \lambda = p \), and, for

\[
N > \max\left\{(4K_{\epsilon_1})^{1/(2\gamma - c)}, N_{\epsilon_1}, (4K_{\epsilon})^{1/(2\gamma - c)}\right\},
\]

\[
\Pr\{||\hat{\alpha}(\alpha) - p||^2 > K_{\epsilon}^* N^{-2\gamma}\} < \epsilon + \epsilon_1,
\]

where \( \hat{\alpha}(\alpha) = \alpha \hat{p} + \bar{\alpha} \hat{\lambda}, \bar{\alpha} = \min\left(\frac{N - \epsilon}{\|\hat{p} - \hat{\lambda}\|^2}, 1\right) \), and \( c < 2g \).

**Corollary 2:**

If \( ||\hat{p} - p||^2 = O_p(N^{-2g}) \) and \( ||\hat{\lambda} - \lambda||^2 = O_p(N^{-2\gamma}) \), then, when \( p = \lambda \), \( ||\hat{\alpha}(\alpha) - p||^2 = O_p(N^{-2\gamma}) \).

Proof of Theorem 6:

Note that when \( p = \lambda \),

\[
||\hat{\alpha}(\alpha) - p||^2 > K_{\epsilon}^* N^{-2\gamma}
\]

may be rewritten as

\[
||\hat{\lambda} - \lambda + \alpha(\hat{p} - \hat{\lambda})||^2 > K_{\epsilon}^* N^{-2\gamma},
\]

which implies that either:

1. \( ||\hat{\lambda} - \lambda||^2 > K_{\epsilon}^* N^{-2\gamma} \) or
2. \( \alpha \neq 0 \) and \( ||\hat{\lambda} - \lambda||^2 \leq K_{\epsilon}^* N^{-2\gamma} \).

53
Statement (2) implies that \( \| \hat{\lambda} - \lambda \|^2 > N^{-c} \) and \( \| \hat{\lambda} - \lambda \|^2 \leq K_\epsilon^* N^{-2\gamma} \).
Since, by assumption, \( N \geq (4K_\epsilon^*)^{1/(2\gamma - c)} \), we have that
\[
K_\epsilon^* N^{-2\gamma} \leq \frac{1}{4} N^{-c},
\]
and, consequently,
\[
\| \hat{\lambda} - \lambda \|^2 > \frac{1}{4} N^{-c}.
\]
Since by assumption,
\[
N > (4K_\epsilon) \frac{1}{2} N^{-c},
\]
\[
\frac{1}{4} N^{-c} > K_\epsilon N^{-2g},
\]
and (2) implies that
\[
\| \hat{\lambda} - \lambda \|^2 > K_\epsilon N^{-2g}.
\]
Thus, \( \| \hat{\hat{\lambda}} - \hat{\lambda} \|^2 > K_\epsilon^* N^{-2\gamma} \) implies that either
(1): \( \| \hat{\lambda} - \lambda \|^2 > K_\epsilon^* N^{-2\gamma} \) or
(2): \( \| \hat{\hat{\lambda}} - \hat{\lambda} \|^2 > K_\epsilon N^{-2g} \).
Since we assumed that \( N > N_\epsilon^* \) and \( N > N_\epsilon \),
\[
\Pr\{ \| \hat{\hat{\lambda}}(\alpha) - \hat{\lambda} \|^2 > K_\epsilon^* N^{-2\gamma} \} \leq
\]
\[
\Pr\{ \| \hat{\hat{\lambda}} - \hat{\lambda} \|^2 > K_\epsilon^* N^{-2\gamma} \} + \Pr\{ \| \hat{\hat{\lambda}} - \hat{\lambda} \|^2 > K_\epsilon N^{-2g} \} < \epsilon + \epsilon_1
\]
It is worthwhile to comment briefly on the role of the constant \( c \) in the definition of \( \hat{\hat{\lambda}}(\alpha) \). It should be noted that with the exception of Theorem 3, the value that \( c \) takes within the interval \([g, 2g]\) doesn’t matter as far as the asymptotic results are concerned. Of course, one problem with the utilization of asymptotic results is that for a particular finite sample size, one doesn’t know how nearly the results hold. In general, the closer one comes to \( 2g \) in choosing \( c \), the better the behavior of \( \hat{\hat{\lambda}}(\alpha) \) will be when \( p \notin \Lambda \), while the poorer its behavior will be when \( p = \lambda \). This choice must therefore be made while taking into consideration the relative seriousness of errors depending on whether \( p = \lambda \) or not. An example is given in the next chapter in which these considerations are examined in conjunction with the choice of \( c \).

In chapter IV, much attention was given to the estimate \( \hat{q}(\alpha) \), which was defined to be

\[
\alpha \hat{p} + \bar{\alpha}\lambda,
\]

where

\[
\bar{\alpha} = 1 - \alpha = \min\left( \frac{N^{-c}}{\|\hat{p} - \lambda\|^2}, 1 \right)
\]

when \( \lambda \) is assumed to be a fixed and known point, and

\[
\hat{q}(\alpha) = \alpha \hat{p} + \bar{\alpha}\hat{\lambda},
\]

where

\[
\bar{\alpha} = \min\left( \frac{N^{-c}}{\|\hat{p} - \hat{\lambda}\|^2}, 1 \right)
\]

when \( \hat{\lambda} \) is an estimate of the point \( \lambda \) that is preferable to the estimate \( \hat{p} \) whenever \( p = \lambda \).

The theorems of chapter IV demonstrated the properties of \( \hat{q}(\alpha) \) with respect to convergence in probability, convergence in distribution, and convergence in expected squared norm. This last criterion encompasses mean squared error, in the case that \( p \) is univariate, as well as mean integrated squared error, when \( p \) is a density.

In this chapter, I will give an example of the application of the estimate \( \hat{q}(\alpha) \) to a data set that arose in a consulting situation, and I will indicate a number of other interesting situations in density estimation in which the use of \( \hat{q}(\alpha) \) might be appropriate. As it arose, the consulting problem was one of hypothesis testing, rather than of density estimation, but it was one in which a nested series of
alternatives was of interest. I will be examining the related question of how to estimate the underlying density if it is desired to make the estimate converge more quickly than a more general estimate would whenever one of the more restrictive hypotheses is true. First, however, I will explain the way in which the data were generated so that I may motivate the parametric hypotheses that are of interest in this situation.

The data set I will be using in this chapter was generated by a nuclear fuel rod scanner. This machine is used to measure the fissile content of rods of uranium in order to measure their U-238 content and the uniformity of its distribution along the rod. The way this is done is to move the rod past detectors which measure the density of the rod and the radioactivity emanating from it. The scanner then accumulates the radioactive counts over fixed time periods, and rejects the rod on the basis of nonuniformity if the count is surprisingly high or low in any one of the time intervals. It is reasonable to assume that the process that generates the counts is a Poisson process with constant rate when the fuel rod is truly uniform. However, the radioactive counts accumulate at such a high rate that the count in any given time interval is very nearly normally distributed. In the data set I will be using, there were 876 observations taken, each of which had a mean number of counts on the order of 100,000 counts per interval, so the data are approximately Normal with mean and variance equal to 100,000. In addition, this data set was not generated by scanning an actual rod, but rather by placing a stationary radiation source inside the scanner, and scanning it for approximately the length of time it would have taken a fuel rod to pass through the scanner. Thus, for this data set, there is no reason to question the assumption that the fuel rod is uniform, and consequently, the observations may be assumed to be identically distributed.

This is a well-behaved situation, and a straightforward application of Gnedenko’s extreme value results ought to provide the desired approximate rejection probabilities. However, the question arose whether the interaction between the scanner’s digital counter and its computer were leading to erroneous counts. In particular, it was hypothesized that the computer might read the counter incorrectly if the counter was in the process of incrementing itself. The recorded count would then represent the contents of the counter’s register in an intermediate stage of the addition of one to the existing count. For example, if the counter were in the process of incrementing itself from 19,999 to 20,000, the computer might with small probability read 19,999, or 19,900, or 19,000, or 10,000. If this hypothesis were true, then the distribution wouldn’t be Poisson, but rather a contaminated Poisson, where the contamination would be a mixture of a Poisson and a discrete distribution. Thus, the consulting problem was to test the hypothesis that the data were Poisson distributed, against the alternative that they were distributed according to a contaminated Poisson. Due to the large count rate, the Poisson distributions were indistinguishable from normal distributions for the sample sizes that were of interest,
so the problem reduced to testing whether the data came from a normal distribution or a contaminated normal distribution. Since the variance of a Poisson variable is equal to its mean, if the data is in fact approximately normally distributed, then the distribution should be well approximated by the $\mathcal{N}(\mu, \mu)$ distribution rather than a general $\mathcal{N}(\mu, \sigma^2)$ distribution.

Just as it is inappropriate to use a screwdriver to drive a nail, it is generally inappropriate to use density estimation methods to solve hypothesis testing problems. Consequently, I don’t advocate the use of $\hat{q}(\alpha)$ as the solution to this testing problem. However, this problem does give an example of a nested series of three hypotheses, namely:

(i). The $X$’s are absolutely continuously distributed.
(ii). The $X$’s are normally distributed, and
(iii). The $X$’s are distributed $\mathcal{N}(\mu, \mu)$.

Although it is clear from the nature of the data as counts that hypothesis (i) isn’t satisfied, the data are measured to many significant figures, so that, from a practical standpoint they are indistinguishable from data which are inherently continuous and have been discretized in the measurement process. Thus the question arises of how to estimate the underlying density of these data when the density is suspected to be either $\mathcal{N}(\mu, \mu)$ or a normal density of general form.

Before estimating the density, note that the estimate $\hat{q}(\alpha)$ isn’t scale invariant, and fails to be translation invariant unless the norm $||\cdot||$ is translation invariant. This failure is in general undesirable, if only because data typically don’t have natural units of measurement, and the estimation method shouldn’t depend on the unit of measurement chosen. In fact, the data set of this example is quite peculiar in that its units aren’t subject to an arbitrary choice of a measurement system.

There are a number ways around this problem. The one that I have chosen is to center my attention on the estimation of the density of the variable $Y$, where $Y$ is the variable $X$, normalized by a measure of location and a measure of scale. The choice of the location and scale measures may be made somewhat arbitrarily, but it is desirable that they be estimates of theoretically defined quantities under a wide variety of situations. In order to facilitate this, the measures I have chosen are the sample median and the sample MAD. The latter quantity is defined to be the sample median of the quantities $\{|X_i - \text{median } \{X_i\}'s\}|$. This choice is a good one, since these quantities are consistent estimates of the population median and MAD, which are in turn well defined whenever the distribution of the $X$’s has convex support. When the data are multivariate, somewhat greater care should be taken in choosing the scale and location measures, since in that case it is desirable that the density estimates also be invariant under rotations.

The density estimate that will be chosen in the example of the radioactive count data will be a convex combination of estimates which are chosen to cater to each of the nested hypotheses. An estimate that is appropriate when nothing is known
other than that the \( X \)'s are absolutely continuous is a kernel or spline estimate. A kernel estimate with normal window was chosen, due largely to its prevalence both in the literature and in practice. Since the resulting density estimate was found by first normalising the data, and then rescaling the resulting density estimate, this kernel estimate is scale invariant and no effort was made to optimize the window width. However, the window width for the transformed data was chosen as \( N^{-\frac{1}{2}} \), which will yield a convergence rate of \( N^{-\frac{1}{4}} \) for the mean integrated squared error.

A density estimate that is appropriate when the \( X \)'s are normal with unknown mean and variance is either the maximum likelihood estimate or the UMVU estimate, both of which have MISE that converges at the rate \( N^{-\frac{1}{4}} \). For this example, the UMVU estimate was chosen although with such large sample sizes, the difference between the two estimates is negligible.

It is a somewhat trickier problem to select an appropriate estimate when hypothesis (iii) is satisfied. One course to follow is to use the maximum likelihood estimate for this normal distribution. In this case, the maximum likelihood estimate of \( \mu \) is

\[
\hat{\mu} = -\frac{1}{2} + \sqrt{\frac{\sum X_i^2}{2}} + \frac{1}{4}.
\]

and the corresponding density estimate is the \( N(\hat{\mu}, \hat{\mu}) \) density.

However, hypothesis (iii) was rationalized by supposing that the data were Poisson distributed with large enough mean to make the Normal approximation useful. If the \( X \)'s are Poisson distributed with mean \( \mu \), then the maximum likelihood estimate of \( \mu \) is simply the sample mean. This points out a potential danger in blindly using methods based on the normal distribution when the data are only asymptotically normal. If hypothesis (iii) was false, then the two methods could give quite different estimates for \( \mu \). However, if this were the case, then no density of the form \( N(\mu, \mu) \) would come very close to the true density, and neither of the \( N(\mu, \mu) \) density estimates would be close to a well chosen nonparametric estimate.

A nontrivial and ever present numerical problem in the evaluation of the estimate \( \hat{\varphi}(\alpha) \) is the evaluation of the constant \( \alpha \), and in particular, the squared norm \( \| \hat{p} - \hat{\lambda} \|^2 \). When the underlying Hilbert space is \( L_2 \), then this evaluation involves a doubly infinite integral of the squared difference between two functions, one of which \( (\hat{p}) \) is numerically difficult to evaluate. One way around this problem is to evaluate \( \hat{p} - \hat{\lambda} \) at relatively few points on a wide but finite grid and trust that a Riemann integral approximation on this grid represents the actual integral well. Somewhat more sophisticated numerical techniques, such as the trapezoidal method or Simpson's method might have been used to improve on the numerical integration, but they were not used in the calculations that follow.

The method used is to choose first an estimate that is a linear combination of the estimates that are appropriate for hypotheses (ii) and (iii) and then choose a
further linear combination of that first linear combination and the nonparametric estimate that is appropriate when only hypothesis (i) is satisfied. Since both the maximum likelihood estimate of the \( \mathcal{N}(\mu, \mu) \) distribution (called \( \hat{\lambda}_3 \)) and of the \( \mathcal{N}(\mu, \sigma^2) \) distribution (called \( \hat{\lambda}_2 \)) converge in mean squared error at the rate \( \frac{1}{\sqrt{N}} \), the constant \( \bar{\alpha}_1 \) of the first linear combination which is equal to

\[
\min \left( \frac{N^{-c_1}}{||\hat{\lambda}_1 - \hat{\lambda}_2||^2}, 1 \right)
\]

should be chosen so that the constant \( c_1 \) lies between .5 and 1 in order to guarantee that the hypotheses of the theorems of chapter IV will be satisfied. The resulting estimate will have mean squared error that is of order \( O\left(\frac{1}{\sqrt{N}}\right) \) regardless of the truth or falsehood of hypothesis (iii). However, the estimate \( \hat{\lambda}_3 \) is considered preferable to the estimate \( \hat{\lambda}_2 \) when hypothesis (iii) is true and it will typically have an asymptotically smaller MISE whenever hypothesis (iii) is satisfied.

If this first estimate is called

\[
\hat{q}_1(\alpha_1) = \alpha_1 \hat{\lambda}_2 + \bar{\alpha}_1 \hat{\lambda}_3,
\]

then the second stage of the estimation will take a convex combination of the nonparametric estimate \( \hat{p} \) and \( \hat{q}_1(\alpha_1) \) of the form

\[
\alpha_2 \hat{p} + \bar{\alpha}_2 \hat{q}_1(\alpha_1) =
\alpha_2 \hat{p} + \bar{\alpha}_2 \alpha_1 \hat{\lambda}_2 + \bar{\alpha}_2 \bar{\alpha}_1 \hat{\lambda}_3,
\]

where

\[
\bar{\alpha}_2 = \min \left( \frac{N^{-c_2}}{||\hat{p} - \hat{q}_1(\alpha_1)||^2}, 1 \right).
\]

This estimate will be called \( \hat{q}_2(\alpha_2) \), and will converge in MISE at the rate \( \frac{1}{\sqrt{N}} \) whenever hypothesis (ii) is satisfied and will converge at the rate \( N^{-\frac{1}{8}} \) whenever it fails to be satisfied.

The following calculations are intended to demonstrate the behavior of \( \hat{q}(\alpha) \) under various conditions. Naturally, no finite sample example can demonstrate asymptotic properties, but these calculations do give a feel for the behavior of the estimate. The actual calculations came out as follows:

\[
||\hat{p} - \hat{\lambda}_2||^2 = .00703,
\]

\[
||\hat{p} - \hat{\lambda}_3||^2 = .00599,
\]

59
and

$$||\hat{\lambda}_3 - \lambda_2||^2 = .00154.$$  

A plot of the estimate \( \hat{p} \) against the estimate \( \hat{\lambda}_3 \) is shown in Figure 1. The scale of the abscissa is in terms of the normalized variable

$$Y = (X - \text{median}(X))/\text{MAD}(X).$$

The scale of the ordinate axis is chosen so that the maximum value that either density takes is a constant.

Since \( \hat{p} \) converges at the rate \( N^{-\frac{3}{4}} \) and in this case, \( N = 876 \), a value of .007 or .006 isn’t particularly surprising when compared to \( N^{-\frac{3}{4}} = .00443 \). Similarly, the value .00154 isn’t surprising when compared to \( \frac{1}{N} = .001 \). In fact, for choices of \( c_1 \) greater less than .9558, the estimate \( \hat{q}_1(\alpha_1) \) will be equal to \( \hat{\lambda}_3 \), which is the estimate that makes the most assumptions about \( p \), and for choices of \( c_2 \) less than .7553, the estimate \( \hat{q}_2(\alpha_2) \) will be equal to \( \hat{\lambda}_3 \) as well. Since \( c_1 \) was to be chosen in the interval \([.5,1]\), it would be only a very conservative choice of \( c_1 \) (that is, a choice close to 1) that would fail to estimate \( \lambda \) by \( \hat{\lambda}_3 \). Similarly, since \( c_2 \) was to be chosen from the interval \([.4,8]\), only a very conservative choice of \( c_2 \) would fail to estimate \( p \) by \( \hat{\lambda}_3 \).

In conclusion, for this sample size, there is no reason to believe that the parametric estimate behaves any worse than a nonparametric estimate, provided it is borne in mind that the truth or falsehood of the parametric assumptions hasn’t been established. It is plausible that, if the sample size were increased, statistically significant differences between the parametric and nonparametric methods would be detected and the estimate would place greater weight on the nonparametric estimate. However, for this sample size, the slower convergence of the nonparametric estimate is offset by the risk of falsehood of the parametric assumptions, and the consequent possible systematic bias of the parametric estimate.

It is worthwhile to examine the estimate \( \hat{q}(\alpha) \) to see what would result if the parametric assumptions were violated. One interesting violation might occur if the true underlying density is a contamination of a member of the assumed family. In such a situation, it is likely that the nonparametric estimate will lie close to some member of the specified parametric family, but not close to the maximum likelihood estimate of the density, since maximum likelihood estimates are not robust against occasional extreme values in the data. In order to illustrate this, I repeated the previous calculation, but I replaced the smallest of the observations (119,800) by zero.

In this case, since the true underlying distribution is a contaminated normal (or Poisson) distribution, the choice of the constants \( c_1 \) and \( c_2 \) has much greater impact on the estimate \( \hat{q}(\alpha) \) than was the case when the underlying distribution was equal to the special distribution. A plot of the nonparametric estimate \( \hat{p} \) against the
maximum likelihood normal density is shown in Figure 2. Note that the maximum likelihood estimate is extremely sensitive to the single outlying value, whereas the nonparametric estimate remains virtually unchanged. Because of this fact, three sets of values for the c's were chosen, one of which was at the upper end of the intervals \([a,2g]\), one of which was in the middle of those intervals, and one of which was at the lower end of those intervals. The resulting \(\hat{\varphi}(\alpha)\)'s are conservative, somewhat less conservative, and relatively “devil–may–care”, respectively. The following table summarizes the calculations made:

<table>
<thead>
<tr>
<th></th>
<th>(c_1 = .75)</th>
<th>(c_1 = .90)</th>
<th>(c_1 = .60)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>(c_2 = .60)</td>
<td>(c_2 = .72)</td>
<td>(c_2 = .48)</td>
</tr>
<tr>
<td>(|\hat{\lambda}_2 - \hat{\lambda}_3|^2)</td>
<td>.3754</td>
<td>.3754</td>
<td>.3754</td>
</tr>
<tr>
<td>(\alpha_1)</td>
<td>.0165</td>
<td>.0060</td>
<td>.0457</td>
</tr>
<tr>
<td>(|\hat{\rho} - \hat{q}_1(\alpha_1)|^2)</td>
<td>.3769</td>
<td>.3754</td>
<td>.3555</td>
</tr>
<tr>
<td>(\alpha_2)</td>
<td>.0455</td>
<td>.0198</td>
<td>.1088</td>
</tr>
<tr>
<td>(|\hat{q}_2(\alpha_2) - \hat{\lambda}|^2)</td>
<td>.00536</td>
<td>.00523</td>
<td>.00758</td>
</tr>
</tbody>
</table>

\(\|\hat{\rho} - \hat{\lambda}\|^2 = .0597\), where \(\hat{\lambda}\) is the N(\(\mu, \varphi\)) estimate based on the original data and represents a best guess of the true underlying distribution.

From this table, it can be seen that the use of either the conservative or moderate choices of the constants \(c_1\) and \(c_2\), yields an estimate \(\hat{q}_2(\alpha_2)\) that is closer to \(\hat{\lambda}\) than \(\hat{\rho}\) itself. There are two reasons that this shouldn’t be surprising. First, although \(\hat{\lambda}\) represents the best guess of the underlying distribution, it is still only an estimate. Thus, if we knew that \(\hat{q}_2(\alpha_2)\) was a worse estimate of \(p\) than \(\hat{\rho}\), then all we could conclude would be that \(\|\hat{\rho} - \hat{\lambda}\|\) was less than or equal to \(\|\hat{q}_2(\alpha_2)\| + \|\hat{\lambda} - p\|\), and not that \(\|\hat{\rho} - \hat{\lambda}\| \leq \|\hat{q}_2(\alpha_2)\|\). Second, it’s not known that \(\hat{q}_2(\alpha_2)\) is a worse estimate of \(p\) than \(\hat{\rho}\), since this comparison will depend on the sign of \(E(\hat{\rho} - p, p - \lambda)\), which is unknown. Thus, it is as likely that \(\hat{\varphi}(\alpha)\) will produce a marginally improved estimate as a marginally worsened estimate.

Finally note that, when the “devil–may–care” choices of the constants \(c_1\) and \(c_2\) are made, the estimate \(\hat{q}_2(\alpha_2)\) is about 30 percent further away from \(\hat{\lambda}\) than is \(\hat{\rho}\). This is due to the fact that the sample size was small enough that the squared norm \(\|\hat{\rho} - \hat{\lambda}\|^2\) didn’t dwarf the constants \(N^{-c_1}\). This demonstrates both that the choice of the constants \(c_i\) is somewhat tricky and that the results of the last chapter are truly only asymptotic in nature. Still, the estimate \(\hat{q}_2(\alpha_2)\) has squared loss that is much preferable (by a factor of about 50) to that of \(\hat{\lambda}_1\), which one might be tempted to use, based on the general shape of the nonparametric estimate \(\hat{\rho}\).
These two examples give an indication of the use of the techniques of chapter IV, but they by no means exhaust their possibilities. Merely in the field of density estimation, there are many situations in which nested series of hypotheses exist which bring rise to density estimates that converge at different rates, based on the truth or falsehood of the various hypotheses. For example, the choice of an optimal order of magnitude for the window width depends on the smoothness of the underlying density, so, estimates can be calculated of the form of \( \hat{q} \) which are based on nonparametric estimates that make different smoothness assumptions. There are two problems with this approach, however. First, the global (MISE) properties of nonparametric estimates whose smoothness assumptions are violated must be scrutinized very carefully. Typically, the mean squared error at a point will converge at the optimal rate wherever the smoothness assumptions are locally satisfied, but will misbehave enough elsewhere that the MISE will converge at a less than optimal rate. However, the preceding statement is very much dependant on the precise form of the nonparametric density estimate and the nature of the violation of the smoothness assumptions and must be examined on a case by case basis. Second, any nonparametric density estimates converge in mean squared error or MISE at a very slow rate, and thus, the sample sizes would need to be very large indeed to distinguish between convergence at the rate \( N^{-d_1} \) and \( N^{-d_2} \), where \( d_1 \) and \( d_2 \) are close to one another.

A second area in which the methods of the last chapter might be applied is in multivariate density estimation. Recall from chapter III that in a \( d \)-dimensional space, typical kernel estimates converge at the rate \( N^{-\frac{d}{d+2}} \), which can be very slow if \( d \) is much greater than 1. If something is known about the dependence of the components of the random vector, then it may be possible to improve on this convergence. For example, if the components are known to be independent, then the estimate \( \hat{p} \) that is the product of the univariate nonparametric estimates will converge at the rate \( N^{-\frac{1}{d}} \). In general, if the covariance matrix of the components of the random vector is of less than full rank, then improvement can be made on the multivariate nonparametric estimate. Also, if the covariance matrix of any transform of the random vector is of less than full rank, then similar improvement can be achieved.
VI. Conclusions and Possible Refinements and Extensions

In chapter IV of this paper, the estimate \( \hat{q}(\alpha) \) was developed. It was defined to be

\[ \alpha \hat{p} + \bar{\alpha} \lambda \]

where

\[ \bar{\alpha} = \min \left( \frac{n^{-c}}{\| \hat{p} - \lambda \|^2} \right) \]

if \( \lambda \) is a fixed point of interest, and to be

\[ \alpha \hat{p} + \bar{\alpha} \hat{\lambda} \]

where

\[ \bar{\alpha} = \min \left( \frac{n^{-c}}{\| \hat{p} - \hat{\lambda} \|^2} \right) \]

if \( \hat{\lambda} \) is an estimate of the fixed point to be \( \lambda \). Several properties of this estimate were discussed, including its convergence in distribution, expected squared norm and probability, based only on assumptions about the convergence properties of the underlying estimates \( \hat{p} \) and \( \hat{\lambda} \). This estimate allows one to incorporate one’s best wishes or suspicions about the truth of various parametric hypotheses, asymptotically achieving the benefits if those suspicions were justified while not having to sacrifice the asymptotic convergence properties that would have resulted if a more conservative estimation approach had been taken. Moreover, the assumptions necessary for the proofs of the convergence results are extremely few, and consequently those results are apt to be applicable in a wide variety of circumstances. In some ways, however these results are unsatisfying. For example, theorem 3 indicates what the behavior of \( \hat{q}(\alpha) \) is like on the set where \( \| \hat{p} - p \|^2 \) is less than \( N^{-\frac{c}{2}} \).
Unfortunately, very little is known about how badly \( \hat{p} \) may behave on the complement of this set, or even about how large the probability measure of this set is. This problem will not be addressed in this paper, however, since to do so, one must become more specific about the choice of the estimate \( \hat{p} \).

The aim of this paper was to see what properties could be gleaned for an estimate of the form of \( \hat{q}(\alpha) \) without having to make very restrictive or unverifiable assumptions about the estimates \( \hat{p} \) and \( \hat{\lambda} \). The bonus gained from this approach is that the results are applicable in very general situations. The penalty that must be paid is that the few assumptions that were made about the estimates \( \hat{p} \) and \( \hat{\lambda} \) may not correspond to established results in the situation of interest. Even in the area of density estimation, the necessary global or pointwise convergence results aren’t established for certain estimates, in particular for nearest neighbor and penalized likelihood estimates. This is in part due to the fact that their appearance in the literature has been fairly recent, but it doesn’t alter the fact that the theorems of chapter IV need to be based on properties of the nonparametric estimate \( \hat{p} \).

In particular, if the results of chapter IV were to be applied in a different area of interest, a literature review analogous to that of chapter III would have to be undertaken before it could be asserted that the estimate \( \hat{q}(\alpha) \) that resulted from the choices of the estimates \( \hat{p} \) and \( \hat{\lambda} \) was in any sense a good one.

Additional work would also be worthwhile in the refinement of the theorems of chapter IV. In particular, it is thought that hypothesis (iv) of Theorem 4 may be unnecessary. This would make the result more attractive, since then the convergence in distribution of \( \hat{q}(\alpha) \) would depend only on the convergence in distribution of its components, \( \hat{p} \) and \( \hat{\lambda} \).

It is also the case that the estimates of chapter IV may be generalized to employ different constants \( \alpha \) without much difficulty, at the expense of making the proofs less transparent. This possibility was pointed out in chapter IV, though no such generalization have been attempted as of this writing.

Another way in which the estimates of chapter IV and the accompanying proofs should be generalized is to incorporate the modification that was used in the actual application of chapter V in order to make \( \hat{q}(\alpha) \) scale invariant. This may be accomplished by treating the estimate based on

\[
Y_i = \frac{X_i - \text{location estimate}}{\text{scale estimate}}
\]

as an approximation (that is, an estimate) of the estimate based on

\[
Z_i = \frac{X_i - \text{location parameter}}{\text{scale parameter}},
\]

and using the convergence properties of the location and scale estimates along with the continuity properties of the estimation procedures \( \hat{p} \) and \( \hat{\lambda} \). However, since this
will depend on these additional properties, it is not possible to treat this problem in as great generality as the problem of formulating \( \hat{q}(\alpha) \). Because of this, it will not be discussed in this paper.

Finally, it is important to note that theorems 1 through 3 of chapter IV rely on the use of the expected squared norm as a loss function by which to judge the goodness of an estimate. The selection of the particular norm to use wasn’t discussed, however, other than to say that the choice

\[
||p||^2 = \int p^2(u) \, d\mu(u)
\]

would yield the mean integrated squared error criterion with respect to the measure \( \mu \), and that the choice

\[
||p||^2 = p^2(u)
\]

would yield the mean squared error of \( \hat{p} \), evaluated at the point \( u \). In any applied situation, however, it is of utmost importance to assess the appropriateness of this loss function. Indeed, some of the early criticism of the James-Stein estimate of a multivariate normal mean was that it could make gross errors in individual components of the vector for the sake of making marginal gains in the other components. This is an invalid criticism of the James-Stein estimate itself, because it questions the appropriateness of the use of the expected \( d \)-dimensional Euclidean norm as a loss function, rather than the fact that the James-Stein estimate represents an improvement with respect to that loss function. However, in an applied context, such a question is quite valid, and consequently one must be very careful that the chosen loss function is the one with respect to which optimization is desirable. Obviously, this question can be addressed only in the context of the actual situation, but, if anything, its lack of generality makes its resolution all the more important.
Bibliography


