AD-A035 145 TEXAS UNIV AT AUSTIN ELECTRONICS RESEARCH CENTER F/G 12/1
NONPARAMETRIC ESTIMATION WITH LOCAL RULES.(U)
OCT 76 C S PENROD, T J WAGNER F44620-76-C-0089
UNCLASSIFIED TR-182 AFOSR-TR-77-0019 NL

1 OF 2
AD-A
035145

NONPARAMETRIC ESTIMATION WITH LOCAL RULES

TEXAS UNIVERSITY AT AUSTIN

11 OCTOBER 1976

NONPARAMETRIC ESTIMATION WITH LOCAL RULES

by

C. S. Penrod and T. J. Wagner

Department of Electrical Engineering

Technical Report No. 182

October 11, 1976

INFORMATION SYSTEMS RESEARCH LABORATORY

ELECTRONICS RESEARCH CENTER
THE UNIVERSITY OF TEXAS AT AUSTIN
Austin, Texas 78712

The Electronics Research Center at The University of Texas at Austin constitutes interdisciplinary laboratories in which graduate faculty members and graduate candidates from numerous academic disciplines conduct research.

Research conducted for this technical report was supported in part by the Department of Defense's JOINT SERVICES ELECTRONICS PROGRAM (U.S. Army, U.S. Navy, and the U.S. Air Force) through the Research Contract AFOSR F44620-76-C-0089. This program is monitored by the Department of Defense's JSEP Technical Advisory Committee consisting of representatives from the U.S. Army Electronics Command, U.S. Army Research Office, Office of Naval Research, and the U.S. Air Force Office of Scientific Research.

Additional support of specific projects by other Federal Agencies, Foundations, and The University of Texas at Austin is acknowledged in footnotes to the appropriate sections.

UNCLASSIFIED

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

REPORT DOCUMENTATION PAGE		READ INSTRUCTIONS BEFORE COMPLETING FORM
1. REPORT NUMBER AFOSR - TR - 77 - 0019	2. GOVT ACCESSION NO.	3. RECIPIENT'S CATALOG NUMBER
4. TITLE (and Subtitle) Nonparametric Estimation with Local Rules		5. TYPE OF REPORT & PERIOD COVERED TECHNICAL REPORT INTERIM 6. PERFORMING ORG. REPORT NUMBER
7. AUTHOR(s) Penrod, C. S. Wagner, T. J.		8. CONTRACT OR GRANT NUMBER(s) F44620-76-C-0089
9. PERFORMING ORGANIZATION NAME AND ADDRESS University of Texas at Austin Austin TX 78712		10. PROGRAM ELEMENT, PROJECT, TASK AREA & WORK UNIT NUMBERS 61102F
11. CONTROLLING OFFICE NAME AND ADDRESS AFOSR/NE Bolling AFB, Washington DC		12. REPORT DATE 1976 13. NUMBER OF PAGES 104
14. MONITORING AGENCY NAME & ADDRESS(if different from Controlling Office)		15. SECURITY CLASS. (of this report) UNCLASSIFIED 15a. DECLASSIFICATION/DOWNGRADING SCHEDULE

16. DISTRIBUTION STATEMENT (of this Report)

Approved for public release; distribution unlimited

17. DISTRIBUTION STATEMENT (of the abstract entered in Block 20, if different from Report)

18. SUPPLEMENTARY NOTES

19. KEY WORDS (Continue on reverse side if necessary and identify by block number)

20. ABSTRACT (Continue on reverse side if necessary and identify by block number)

The application of nearest neighbor rules and other local rules to the problem of estima ting a parameter is investigated. It is assumed that aloss function L, and observed random vector X, and data consisting of a sequence of independent random vectors $(X_1, \Theta_1) \ldots, (X_n \theta_n)$ with the same distribution as (X, θ) are given. Condttions are shown for which, if R* denotes the Bayes risk (the minimum expected loss possible), then the conditional expected loss of the k-nearest neighbor rule, conditioned on the data, converges to (1+1/k)R* for

DD FORM 1 JAN 73 1473 EDITION OF 1 NOV 65 IS OBSOLETE

UNCLASSIFIED

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

squared-error loss function. For k_n-nearest neighbor rules where $k_n \to \infty$ and $k_n/n \to 0$, conditions are given under which the rules are asymptotically optimal.

In addition, methods of estimating the conditional risk of a rule with a particular data set are investigated. For a class of rules called local rules, the performance of two different estimates of the risk is bounded independently of the underlying distribution of (X, θ). This enables the statistician to construct confidence intervals for the risk of the rule and data he is using, without knowledge of the distribution of (X, θ).

NONPARAMETRIC ESTIMATION WITH LOCAL RULES

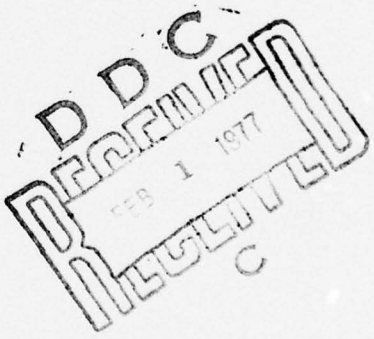
by

C. S. Penrod and T. J. Wagner
Department of Electrical Engineering

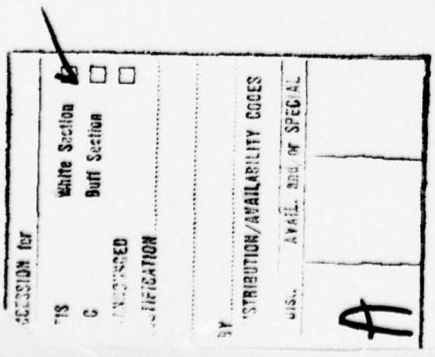
Technical Report No. 182
October 11, 1976

INFORMATION SYSTEMS RESEARCH LABORATORY

ELECTRONICS RESEARCH CENTER
THE UNIVERSITY OF TEXAS AT AUSTIN
Austin, Texas 78712



*Research Sponsored by the Joint Services Electronics Program under Research Contract F44620-76-C-0089



ABSTRACT

The application of nearest neighbor rules and other local rules to the problem of estimating a parameter θ is investigated. It is assumed that a loss function L, an observed random vector X, and data consisting of a sequence of independent random vectors $(X_1, \theta_1), \ldots, (X_n, \theta_n)$ with the same distribution as (X, θ) are given. Conditions are shown for which, if R^* denotes the Bayes risk (the minimum expected loss possible), then the conditional expected loss of the k-nearest neighbor rule, conditioned on the data, converges to $(1 + 1/k)R^*$ for squared-error loss functions. For k_n-nearest neighbor rules where $k_n \to \infty$ and $k_n/n \to 0$, conditions are given under which the rules are asymptotically optimal.

In addition, methods of estimating the conditional risk of a rule with a particular data set are investigated. For a class of rules called local rules, the performance of two different estimates of the risk is bounded independently of the underlying distribution of (X, θ). This enables the statistician to construct confidence intervals for the risk of the rule and data he is using, without knowledge of the distribution of (X, θ).

TABLE OF CONTENTS

LIST OF FIGURES

I. INTRODUCTION

I.1 A Description of the Nonparametric Estimation Problem

The estimation problem to be considered can be loosely described as the problem of determining how to guess the value of an unknown parameter θ, when the only available information concerning the value of θ is contained in i) an observation X which is related to θ in some probabilistic sense, and ii) some form of information concerning the probability structure underlying the relationship between X and θ. A simple example of this type of problem is the question of determining how to estimate the weight of an individual selected at random from some population, when the individual's height and certain statistics concerning the heights and weights of members of the population are known. A more interesting and realistic example could be the problem of determining how to estimate the production to be expected from an oil well when the available information is in the form of measurements such as pressure and temperature within the well, and knowledge of past experience concerning the relationship between such measurements and production for preceding wells.

In order to complete the description of the estimation problem, the relationship between the observation X and the parameter θ must be specified. In addition, some method of comparing the performance of various estimators must be determined. In the analysis to follow, it will be assumed that (X, θ) is a random vector with joint distribution

function $F(x, \theta)$. (This formulation is known as the Bayesian estimation problem.) We will assume that X takes values in $\mathbb{R}^d$ and θ takes values in $\mathbb{R}^p$. We will also assume the knowledge of a loss function L defined on $\mathbb{R}^p \times \mathbb{R}^p$ so that $L(\theta, \hat{\theta})$ is the loss incurred when θ is the true value of the parameter and $\hat{\theta}$ is the estimate. If we define an estimation rule as a function $\hat{\theta}: \mathbb{R}^d \to \mathbb{R}^p$, and if we intend to use $\hat{\theta}$ to estimate the parameters for a number of observations, then it is natural to consider $E\{L(\theta, \hat{\theta}(X))\}$ as a performance criterion for $\hat{\theta}$. ($E\{L(\theta, \hat{\theta}(X))\}$ is known as the risk associated with the estimation rule $\hat{\theta}$.) The statistician will be interested in choosing a function $\hat{\theta}$ to minimize $E\{L(\theta, \hat{\theta}(X))\}$ since the average loss incurred when $\hat{\theta}$ is used on a large number of observations will be near this value.

From this discussion it is clear that the amount and type of information available to the statistician concerning the distribution $F(x, \theta)$ must be a crucial factor in the determination of the estimator $\hat{\theta}$. We will briefly discuss two possible degrees of knowledge of $F(x, \theta)$ which the statistician may possess.

The first case, in which $F(x, \theta)$ is completely known, is interesting because it yields the optimal estimator. If, for each $x \in \mathbb{R}^d$, we define $\theta^* = \theta^*(x)$ such that for all $\theta' \in \mathbb{R}^p$

$$E\{L(\theta, \theta^*)/X\} \leq E\{L(\theta, \theta')/X\} \tag{1.1}$$

then

$$R^* = E[E\{L(\theta, \theta^*) \,|X\}] = E\{L(\theta, \theta^*)\} \tag{1.2}$$

is known as the Bayes risk and is clearly less than or equal to the risk, $E\{L(\theta, \hat{\theta})\}$, for any other estimation rule. R^* is the smallest possible risk and any estimation rule which has risk R^* is called an optimal rule. In general, in the absence of fairly specific information concerning $F(x, \theta)$, an optimal rule cannot be constructed.

There are any number of problems in which partial knowledge of $F(x, \theta)$ is available. The one to be discussed throughout the remainder of this report is known as the nonparametric estimation problem, in which the only information concerning $F(x, \theta)$ is contained in a data sequence $(X_1, \theta_1), \ldots, (X_n, \theta_n)$ of independent, identically distributed random vectors, with distribution $F(x, \theta)$. A nonparametric estimation rule $\hat{\theta}$ will be defined as a mapping

$$\hat{\theta}: \mathbb{R}^d \times \left(\mathbb{R}^d \times \mathbb{R}^p\right)^n \to \mathbb{R}^p . \tag{1.3}$$

(We do not exclude the possibility that $\hat{\theta}$ may incorporate randomization, as will frequently be the case.) An example of such an estimation rule is the nearest neighbor rule. This rule was first discussed in the context of estimation by Cover [1], but it was originally presented as a discrimination procedure in a pair of reports by Fix and Hodges [2, 3]. The nearest neighbor rule is quite obvious in concept. If X is an observation for which θ is to be estimated, the rule chooses as its estimate the value θ_j associated with the observation X_j from the data set which is closest to X.

A simple generalization of the nearest neighbor rule is the k-nearest neighbor rule. This rule uses as its estimate of θ the average of the parameters of the k-closest observations to X from the data set. Another example of a nonparametric estimation rule is one in which the data set is used to estimate $F(x, \theta)$. If $\hat{F}(x, \theta)$ is the estimate of $F(x, \theta)$, then $\hat{\theta}(x)$ is chosen to minimize $E\{L(\theta, \hat{\theta}(x))/X = x\}$ where the expectation is taken with respect to the distribution $\hat{F}$. The rule $\hat{\theta}$ thus constructed would actually be optimal if $\hat{F}(x, \theta)$ were equal to $F(x, \theta)$.

I.2 Evaluation of Estimation Rules

The process of selecting and evaluating an estimation rule should consist basically of examining the following three factors:

i) cost of implementation and use

ii) asymptotic performance

iii) finite sample performance.

The question of implementation cost will not be considered in any detail here. Both of the remaining factors have strong significance for the statistician.

Asymptotic performance is often called "large sample" performance, but the question of how large n (the size of the data set) should be so that a rule approaches its large sample performance is a difficult one. The adequacy of a data set depends heavily on the distribution $F(x, \theta)$, and it requires but a modicum of effort to imagine problems where fifty

samples is either a small or a large data set. In spite of this, if the statistician has any hope at all of obtaining a data set which is adequate for his problem, he will be interested in knowing how the large sample performance of his rule compares with R*, the performance of an optimal rule. In cases where there is reason to believe that an adequate data set is available, the statistician's knowledge of asymptotic behavior alone may enable him to choose one rule over another. Hence the importance of knowledge concerning the asymptotic behavior of the risk of a rule.

The importance of knowledge concerning the finite sample performance of a rule with the particular data set which is currently available is even more readily apparent. For example, suppose that a statistician is using a rule for which he knows the risk converges in some sense to R*. Without knowledge of $F(x,\theta)$, however, the statistician does not know the value of R*, so that even if his data set is quite large, he does not know how well his rule will perform. In fact, even though the rule should do as well as possible in the large sample case, its performance with the data set available may be unacceptably bad if either R* is large or the data set is not adequate. As another example of the need for a good estimate of finite sample performance, consider the case where additional data may be acquired at some significant cost. In this case the statistician's knowledge of current performance may indicate that there is no need to gather additional data, or it may enable him to

measure the performance improvement achieved by expanding the data set, so that he can determine if the improvement is worth the cost. In order to be of benefit in solving these problems, the estimate of finite sample performance should be a function only of the rule and the data set to be used.

I.3 Discussion of Results

Most of the results presented are for the nearest neighbor rule and variations. These rules remain among the most interesting solutions to the estimation problem because of their simplicity and because of the strength of the results which may be obtained for them. The major criticisms of the nearest neighbor rules are that the entire data set must be stored, and, for large data sets in a many dimensional observation space, the computation required to find the nearest neighbors can be significant. But the fact that most of the information contained in the data about an observation X is inherently contained in the observations near X implies that any rule which would take full advantage of the data set must be generally subject to the same criticisms. The nearest neighbor rules continue to serve as a benchmark against which other estimation rules should be compared.

In order to discuss the results, some notation to be used in the following pages will be described. The conditional n sample risk for an estimation rule $\hat{\theta}$: $\mathbb{R}^d \times (\mathbb{R}^d \times \mathbb{R}^p)^n \to \mathbb{R}^p$ conditioned on the data

$(X_1, \theta_1), \ldots, (X_n, \theta_n)$ will be denoted

$$L_n = E\left\{L(\theta, \hat{\theta})/(X_1, \theta_1), \ldots, (X_n, \theta_n)\right\} . \tag{1.4}$$

Note that the dependence of $\hat{\theta} = \hat{\theta}(X, (X_1, \theta_1), \ldots, (X_n, \theta_n))$ on X and the data is not explicit in this notation. L_n is clearly a random variable since it is readily seen to be a measurable function of the data.

In Cover's [1] original paper on estimation with nearest neighbor rules it was shown that

$$R^* \leq \limsup E L_n \leq 2R^* \tag{1.5}$$

for bounded metric loss functions with certain continuity assumptions on $E(L(\theta, \theta_0)/X, \theta_0)$. A metric loss function is a function $L: \mathbb{R}^p \times \mathbb{R}^p \to \mathbb{R}$ which is a metric on $\mathbb{R}^p$. (Cover actually claims that $EL_n \to R$, but his proof fails to demonstrate this.) If $\|\cdot\|$ denotes a norm on $\mathbb{R}^p$, then the squared-error loss function is defined as

$$L(\theta_1, \theta_2) = \|\theta_1 - \theta_2\|^2 . \tag{1.6}$$

For the case of the squared-error loss function with the k-nearest neighbor rule Cover gives conditions for which

$$EL_n \to R \tag{1.7}$$

where $R = \left(1 + \frac{1}{k}\right)R^*$. (1.8)

This result provides bounds on R, which can be interpreted loosely as the <u>average</u> large sample risk, where the averaging is performed over

all large data sets. This result, although of interest to the statistician, does not really address itself to the questions concerning asymptotic performance which were pointed out as being most interesting in section I.2. The reason is that the statistician does not have a large number of large data sets over which he will average the performance of his rule. The question concerning asymptotic performance which is really of interest is what happens when a single data set is made large. This question is answered in Chapter II in which it is shown that for the k nearest neighbor rule with the squared-error loss function

$$L_n \to R \text{ in probability} \tag{1.9}$$

where R satisfies (1.8). This result is analogous to results obtained by Wagner [4] for the discrimination problem. In addition, for metric loss functions conditions are given under which, for all $\varepsilon > 0$

$$P\{L_n - 2R^* \geq \varepsilon\} \to 0 \ . \tag{1.10}$$

It is also shown that if $k_n/n \to 0$ and $k_n \to \infty$, then

$$L_n \to R^* \text{ in probability} \tag{1.11}$$

for the k_n-nearest neighbor rule with the squared-error loss function. A rule which satisfies (1.11) is called asymptotically optimal.

Chapter III is primarily concerned with the evaluation of finite

sample performance for estimation rules. (Reference is made to Toussaint's [5] survey of techniques employed for this purpose in discrimination problems. Most of the same methods can be applied to estimation problems.) This problem can be restated in terms of using the given data set to estimate L_n. Two estimates of L_n are shown to have the property that for certain classes of estimation rules

$$P\{|L_n - \hat{L}_n| \geq \epsilon\}$$

can be bounded independently of the distribution $F(x, \theta)$. The bound obtained for the deleted estimate decreases at rate $1/n$, while the bound for the holdout estimate decreases at rate $1/\sqrt{n}$. These bounds provide a solution to the need for methods of evaluating finite sample performance by allowing the statistician to construct confidence intervals for L_n which are independent of the underlying distribution $F(x, \theta)$.

In Chapters II and III, the main emphasis is on providing solutions to the problems discussed in I.2 concerning the evaluation of estimation rules. Chapter IV consists of a discussion of the results of a simulation study performed on the deleted estimate and the holdout estimate of L_n. This study was performed in order to gain some experimental verification of the theoretical results of Chapter III.

II. ASYMPTOTIC PERFORMANCE OF NEAREST NEIGHBOR RULES IN ESTIMATION

II.1 Preliminary Remarks

This chapter will present results concerning the convergence of the conditional n sample risk for k-nearest neighbor rules and k_n-nearest neighbor rules. In order to obtain these results it has been necessary to restrict the class of loss functions to squared-error loss functions and metric loss functions. A squared error loss function satisfies

$$L(\theta_1, \theta_2) = \|\theta_1 - \theta_2\|^2 \tag{2.1}$$

where $\|\cdot\|$ is a norm on the space $\mathbb{R}^p$. A metric loss function is one in which $L(\theta_1, \theta_2)$ is a metric on the space $\mathbb{R}^p$. These two types of loss functions cover a broad range of practical applications.

The types of nearest neighbor rules to be discussed will be chiefly those which weight the k-nearest neighbors equally in forming the estimate. Such rules can be described as follows. Let $D_n = (X_1, \theta_1), \ldots, (X_n, \theta_n)$ be the data sequence consisting of n independent identically distributed random vectors with distribution $F(x, \theta)$, where for each i, X_i takes values in $\mathbb{R}^d$ and θ_i takes values in the parameter space $\mathbb{R}^p$. Then, if (X, θ) has distribution $F(x, \theta)$, the k-nearest neighbor rule estimate of θ is

$$\hat{\theta} = \frac{1}{k}\sum_{i=1}^{k} \theta_{(i)} \tag{2.2}$$

where $\theta_{(i)}$ is the parameter associated with the i^{th} closest observation to X from D_n, distance being measured in the usual Euclidean metric on $\mathbb{R}^d$, or any other metric on $\mathbb{R}^d$. The k_n-nearest neighbor rule is obtained by simply allowing k to be a function of n in (2.2).

Section II.2 will contain theorems concerning the convergence of the conditional risk for the single (k=1) nearest neighbor rule. In section II.3 these results will be generalized to $k > 1$, and in section II.4 results will be shown for k_n-nearest neighbor rules. Section II.5 will contain a brief discussion of nearest neighbor rules which do not weight the k-nearest neighbors equally in forming the estimate of θ.

II.2 The Single-Nearest Neighbor Rule

The following lemma will be used extensively in this chapter. We first define the support of a probability density function as the smallest closed set E such that

$$\int_E f(x)dx = 1 \ . \tag{2.3}$$

Lemma 1: Let $X_1, X_2, \ldots, X_n$ be a sequence of independent identically distributed random vectors taking values in $\mathbb{R}^d$, and let f be a probability density function on $\mathbb{R}^d$ corresponding to the distribution of X_1. Let $\rho(\cdot,\cdot)$ be a metric on $\mathbb{R}^d$, let $E \subset \mathbb{R}^d$ be the support of f, and let $K \subset E$ be compact with the metric ρ.

Let $$V_{j,n} = \left\{x \in \mathbb{R}^d: \ \rho(X_j - x) < \rho(X_i - x) \ , \ 1 \le i \le n\right\}. \tag{2.4}$$

Then

$$r_n = \max_{1\le j\le n}\left\{ \sup_{x,y\in KV_{j,n}} \rho(x,y)\right\} \to 0 \quad \text{w.p.1} \tag{2.5}$$

and

$$\lambda_n = \max_{1\le j\le n}\left\{\mu KV_{j,n}\right\} \to 0 \quad \text{w.p.1} \tag{2.6}$$

where μ denotes Lebesgue measure on $\mathbb{R}^d$.

Proof: Note that if $r_n \to 0$, then $\lambda_n \to 0$. Since r_n is monotonically decreasing in n, it suffices to show that for all $\epsilon > 0$

$$P\{r_n \ge \epsilon\} \to 0 \ . \tag{2.7}$$

Let $S_{\epsilon/4}(y)$ denote the open sphere in $(\mathbb{R}^d, \rho)$ with center y and radius $\epsilon/4$. Since K is compact, there exists a finite collection of points $\{y_1, \ldots, y_m\} \subset K$ such that

$$K \subset \bigcup_{i=1}^{m} S_{\epsilon/4}(y_i). \tag{2.8}$$

Now, suppose that for each i, the sphere $S_{\epsilon/4}(y_i)$ contains at least one of the random vectors X_j. Let $x \in K$ so that $x \in S_{\epsilon/4}(y_i)$ for some i, $1 \le i \le m$. Since $S_{\epsilon/4}(y_i)$ contains one of the X_j, the distance from x to its nearest neighbor from $X_1, \ldots, X_n$ must be less than $\epsilon/2$. Since this is clearly true for all $x \in K$ we have $KV_{j,n} \subset S_{\epsilon/2}(X_j)$, $1 \le j \le n$, implying that $r_n < \epsilon$. We can conclude that the event $\{r_n \ge \epsilon\}$ occurs only if one or more of the spheres $S_{\epsilon/4}(y_i)$ contains none of the random vectors $X_1, \ldots, X_n$. Then, if $\{X_j \notin S_{\epsilon/4}(y_i)\}$ denotes the event that X_j

is not an element of $S_{\epsilon/4}(y_i)$,

$$P\{r_n \geq \epsilon\} \leq P\left[\bigcup_{i=1}^{m} \bigcap_{j=1}^{n} \{X_j \notin S_{\epsilon/4}(y_i)\}\right]$$

$$\leq \sum_{i=1}^{m} \left\{1 - \int_{S_{\epsilon/4}(y_i)} f(x)dx\right\}^n \qquad (2.9)$$

Since $y_i \in E$, $\int_{S_{\epsilon/4}(y_i)} f(x)dx > 0$, so that

$$P\{r_n \geq \epsilon\} \to 0$$

which proves the lemma.

In the theorems to follow, it will be necessary to have the sets $V_{j,n}$, $1 \leq j \leq n$, form a partition of $\mathbb{R}^d$. As defined in (2.4), they are disjoint, but their union does not cover $\mathbb{R}^d$. Let $x \in \mathbb{R}^d$ such that $x \notin \bigcup_{j=1}^{n} V_{j,n}$. Then there must exist $i \leq n$ and $j \leq n$, $i \neq j$, such that

$$\rho(X_i, x) = \rho(X_j, x) .$$

Clearly x is a boundary point of the sets $V_{i,n}$ and $V_{j,n}$. In order to modify the sets $V_{j,n}$, $1 \leq j \leq n$, so that they form a partition of $\mathbb{R}^d$, we arbitrarily assign each $x \notin \bigcup_{j=1}^{n} V_{j,n}$ to one of the sets for which it is a boundary point. The proof of Lemma 1 is unchanged when the $V_{j,n}$ are modified in this way, and throughout the rest of the chapter we will assume that the sets $V_{j,n}$ are a partition of $\mathbb{R}^d$.

Lemma 2: Let $X_1, \ldots, X_n$ be a sequence of independent identically distributed random vectors taking values in $\mathbb{R}^d$, and let F be the distribution of X_1. Let f be a density corresponding to F, and let $V_{j,n}$, $1 \le j \le n$, be as defined previously. Then

$$\max_{1 \le j \le n} P\{V_{j,n} / X_1, \ldots, X_n\} \to 0 \quad \text{w.p.1} .$$

Proof: Let E be the support of f, and let $\epsilon > 0$ be given. Then, there exists a set $K \subset E$, compact with the metric ρ, such that

$$P\{K^c\} < \epsilon .$$

Then,

$$\max_{1 \le j \le n} P\{V_{j,n} / X_1, \ldots, X_n\} \le \max_{1 \le j \le n} P\{KV_{j,n} / X_1, \ldots, X_n\} + P\{K^c\}$$

$$\le \max_{1 \le j \le n} P\{KV_{j,n} / X_1, \ldots, X_n\} + \epsilon .$$

Hence it suffices to show that

$$\max_{1 \le j \le n} P\{KV_{j,n} / X_1, \ldots, X_n\} \to 0 \quad \text{w.p.1} .$$

Let S_n denote the set $KV_{j,n}$ for which $P\{KV_{j,n} / X_1, \ldots, X_n\}$ is maximized over $1 \le j \le n$. Now

$$\mu S_n \le \max_{1 \le j \le n} \mu KV_{j,n} ,$$

so that, by Lemma 1,

$$\mu S_n \to 0 \quad \text{w.p.1} .$$

Hence,

$$\max_{1\le j\le n} P\{KV_{j,n}/X_1,\ldots,X_n\} \to 0 \quad \text{w.p.1}$$

since F is absolutely continuous with respect to Lebesgue measure on $\mathbb{R}^d$. This concludes the proof of Lemma 2.

Lemma 3: Let $X_1,\ldots,X_n$ and F be as in Lemma 2. Let $X'_n(x)$ be the nearest neighbor to $x \in \mathbb{R}^d$ from $X_1,\ldots,X_n$. Then

$$P\{x \in \mathbb{R}^d:\ X'_n(x) \to x\} = 1 .$$

Proof: Let E be the support of F, and let $\delta > 0$ be given. Then, there exists $K \subset E$ such that K is compact with the metric ρ and

$$P\{K\} > 1 - \delta .$$

By Lemma 1,

$$\sup_{x\in K} \{\rho(X'_n(x),x)\} \to 0 \quad \text{w.p.1} .$$

Hence $$P\{x \in \mathbb{R}^d:\ X'_n(x) \to x\} > 1 - \delta ,$$

which proves Lemma 3.

Before proving Theorem 1, it will be necessary to briefly discuss some well-known facts concerning optimal estimation rules for squared-error loss functions. Let $\hat{\theta}$: $\mathbb{R}^d \to \mathbb{R}^p$ be an estimation rule, and let $\theta^*(X)$ be any version of $E(\theta/X)$. Then θ^* also defines an estimation rule. With the squared-error loss criterion, the risk associated with $\hat{\theta}$ is given by

$$E\{\|\theta-\hat{\theta}(X)\|^2\} = E\{\|\theta-\theta^*(X)+\theta^*(X)-\hat{\theta}(X)\|^2\}$$

$$= E\{\|\theta-\theta^*(X)\|^2 + 2\left(\theta-\theta^*(X), \theta^*(X)-\hat{\theta}(X)\right)$$

$$+ \|\hat{\theta}(X)-\theta^*(X)\|^2\},$$

where $(\cdot,\cdot)$ denotes the usual inner product on $\mathbb{R}^p$. Now,

$$2E\left\{\left(\theta-\theta^*(X), \theta^*(X)-\hat{\theta}(X)\right)\right\}$$

$$= 2E\left[E\left\{\left(\theta-\theta^*(X), \theta^*(X)-\hat{\theta}(X)\right)/X\right\}\right].$$

Examining the conditional expectation inside the brackets, we have

$$E\left\{\left(\theta-\theta^*(X), \theta^*(X)-\hat{\theta}(X)\right)/X\right\} = E\left\{\left(\theta-\theta^*(X), \theta^*(X)\right)/X\right\}$$
$$- E\left\{\left(\theta-\theta^*(X), \hat{\theta}(X)\right)/X\right\}.$$

Letting θ_i, $\theta_i^*(X)$, and $\hat{\theta}_i(X)$ denote the ith components of θ, $\theta^*(X)$, and $\hat{\theta}(X)$ respectively, we see that

$$E\left\{\left(\theta_i-\theta_i^*(X)\right)\left(\theta_i^*(X)\right)/X\right\} = \theta_i^*(X)[E(\theta_i/X) - \theta_i^*(X)]$$

$$= 0 \quad \text{w.p.1},$$

since $\theta_i^*(X)$ is a version of $E(\theta_i/X)$. We can conclude that

$$E\left\{\left(\theta-\theta^*(X), \theta^*(X)\right)/X\right\} = 0$$

and, similarly,

$$E\left\{\left(\theta-\theta^*(X), \hat{\theta}(X)\right)/X\right\} = 0.$$

This implies that

$$E\left\{\left(\theta-\theta^*(X), \theta^*(X) - \hat{\theta}(X)\right)\right\} = 0 .$$

Then,

$$E\left\{\|\theta-\hat{\theta}(X)\|^2\right\} = E\left\{\|\theta-\theta^*(X)\|^2\right\} + E\left\{\|\hat{\theta}(X) - \theta^*(X)\|^2\right\}$$

implying

$$E\left\{\|\theta-\theta^*(X)\|^2\right\} \leq E\left\{\|\theta-\hat{\theta}(X)\|^2\right\} \tag{2.10}$$

with equality if and only if

$$E\left\{\|\hat{\theta}(X) - \theta^*(X)\|^2\right\} = 0 .$$

The conclusion is that $\theta^*(X)$ is an optimal estimation rule, also known as a Bayes rule, if, and only if it is a version of $E(\theta/X)$.

The risk associated with an optimal rule, known as the Bayes risk and denoted R^*, is given by

$$\begin{aligned} R^* &= E\left\{L\left(\theta, \theta^*(X)\right)\right\} \\ &= E\left\{\|\theta-E(\theta/X)\|^2\right\} \\ &= E\left[E\left\{\|\theta-E(\theta/X)\|^2/X\right\}\right] \\ &= E\left[E(\|\theta\|^2/X) - 2E\left\{\left(\theta, E(\theta/X)\right)/X\right\} + \|E(\theta/X)\|^2\right] . \end{aligned} \tag{2.11}$$

Using the fact that

$$E\{\theta_i E(\theta_i/X)/X\} = E^2(\theta_i/X)$$

we have

$$E\{\left(\theta, E(\theta/X)\right)/X\} = \|E(\theta/X)\|^2 .$$

Then,

$$R^* = E\left[E\left(\|\theta\|^2/X\right) - \|E(\theta/X)\|^2\right] . \tag{2.12}$$

Theorem 1: Let (X, θ), $(X_1, \theta_1), \ldots, (X_n, \theta_n)$ be a sequence of independent identically distributed random vectors, each with distribution $F(x, \theta)$, with X_i taking values in $\mathbb{R}^d$ and θ_i taking values in a compact subset of $\mathbb{R}^p$, $1 \leq i \leq n$. Let $L(\theta_1, \theta_2)$ be the squared-error loss function. If the distribution of X has a density and versions of $E(\theta/X)$ and $E(\|\theta\|^2/X)$ exist which are continuous on $\mathbb{R}^d$ with probability one, then

$$L_n \to 2R^* \text{ in probability}$$

for the single-nearest neighbor rule.

Proof: In order to simplify the notational difficulties, the proof will be carried out for $p = 1$. The modifications necessary to carry out the proof for $p > 1$ will be discussed in the remarks at the end of this chapter. Let $\epsilon > 0$ be given. Then,

$$P\left\{|L_n - 2R^*| \geq \epsilon\right\} \leq P\left\{|L_n - E(L_n/X_1, X_2, \ldots, X_n)| \geq \epsilon/2\right\}$$
$$+ P\left\{|E(L_n/X_1, X_2, \ldots, X_n) - 2R^*| \geq \epsilon/2\right\} . \tag{2.13}$$

We will show that each term on the right-hand side of (2.13) tends to zero as n becomes large. The first term can be bounded by Chebychev's inequality so that it will be sufficient to show that

$$E[L_n - E(L_n/X_1, \ldots, X_n)]^2 \to 0$$

or, equivalently,

$$E\left\{E\left[\left(L_n - E(L_n/X_1,\ldots,X_n)\right)^2/X_1,\ldots,X_n\right]\right\} \to 0 \ . \tag{2.14}$$

For the squared-error loss function, the hypothesis that θ takes values in a compact set guarantees that there is a constant $M < \infty$ such that

$$E\left[\left(L_n - E(L_n/X_1,\ldots,X_n)\right)^2/X_1,\ldots,X_n\right] \le M \tag{2.15}$$

with probability one for all n. Then, by the Lebesgue dominated convergence theorem, (2.14) will be established if it is shown that

$$E\left[\left(L_n - E(L_n/X_1,\ldots,X_n)\right)^2/X_1,\ldots,X_n\right] \to 0 \text{ in probability.} \tag{2.16}$$

Let $D_n = \{(X_1,\theta_1),\ldots,(X_n,\theta_n)\}$. Then

$$E\left[\left(L_n - E(L_n/X_1,\ldots,X_n)\right)^2/X_1,\ldots,X_n\right]$$
$$= E\left[\left\{E\left(L(\theta,\hat{\theta})/D_n\right) - E\left(L(\theta,\hat{\theta})/X_1,\ldots,X_n\right)\right\}^2/X_1,\ldots,X_n\right] \tag{2.17}$$

The right-hand side of (2.17) can be written as

$$E\left[\left\{\sum_{j=1}^{n}\left[E\left(L(\theta,\theta_j)I_{[V_{j,n}]}(X)/D_n\right)\right.\right.\right.$$
$$\left.\left.\left. - E\left(L(\theta,\theta_j)I_{[V_{j,n}]}(X)/X_1,\ldots,X_n\right)\right]\right\}^2/X_1,\ldots,X_n\right] \tag{2.18}$$

where $V_{j,n}$ was defined previously, and

$$I_{[V_{j,n}]}(X) = \begin{cases} 1 \ , & \text{if } X \in V_{j,n} \\ 0 \ , & \text{otherwise .} \end{cases} \tag{2.19}$$

Now, the summation in (2.18) will be squared, giving:

$$\begin{aligned} & E\Bigg[\sum_{j=1}^{n} \Big\{E\big(L(\theta,\theta_j)I_{[V_{j,n}]}(X)/D_n\big) \\ & -E\big(L(\theta,\theta_j)I_{[V_{j,n}]}(X)/X_1,\ldots,X_n\big)\Big\}^2/X_1,\ldots,X_n\Bigg] \\ & + E\Bigg[\sum_{i\neq j} \Big\{\Big[E\big(L(\theta,\theta_j)I_{[V_{j,n}]}(X)/D_n\big) \\ & -E\big(L(\theta,\theta_j)I_{[V_{j,n}]}(X)/X_1,\ldots,X_n\big)\Big] \\ & \times \Big[E\big(L(\theta,\theta_i)I_{[V_{i,n}]}(X)/D_n\big) \\ & - E\big(L(\theta,\theta_i)I_{[V_{i,n}]}(X)/X_1,\ldots,X_n\big)\Big]\Big\}/X_1,\ldots,X_n\Bigg] \end{aligned} \tag{2.20}$$

The second term of (2.20), which is the expectation of the sum of the cross-product terms, can be written as follows by taking advantage of the fact that θ_j, conditioned on X_j, is independent of θ_i for all $i \neq j$:

$$\begin{aligned} & \sum_{i\neq j} E\Big[\Big\{E\big(L(\theta,\theta_j)I_{[V_{j,n}]}(X)/X_1,\ldots,X_n,\theta_j\big) \\ & - E\big(L(\theta,\theta_j)I_{[V_{j,n}]}(X)/X_1,\ldots,X_n\big)\Big\} \\ & \times \Big\{E\big(L(\theta,\theta_i)I_{[V_{i,n}]}(X)/X_1,\ldots,X_n,\theta_{i'}\big) \\ & - E\big(L(\theta,\theta_i)I_{[V_{i,n}]}(X)/X_1,\ldots,X_n\big)\Big\}\Big]. \end{aligned} \tag{2.21}$$

Let $g_j(X_1,\ldots,X_n,\theta_j)$ be any version of

$$E\left(L(\theta,\theta_j)I_{[V_{j,n}]}(X)/X_1,\ldots,X_n,\theta_j\right) - E\left(L(\theta,\theta_j)I_{[V_{j,n}]}(X)/X_1,\ldots,X_n\right).$$

Then (2.21) is equal to

$$\sum_{i\neq j} E[g_j(X_1,\ldots,X_n,\theta_j)g_i(X_1,\ldots,X_n,\theta_i)/X_1,\ldots,X_n] \quad . \tag{2.22}$$

Since θ_i and θ_j, conditioned on X_i and X_j, are independent, (2.22) is equal to

$$\sum_{i\neq j} E\left(g_j(X_1,\ldots,X_n,\theta_j)/X_1,\ldots,X_n\right)E\left(g_i(X_1,\ldots,X_n,\theta_i)/X_1,\ldots,X_n\right)$$

$$= 0 \quad . \tag{2.23}$$

The first term of (2.20) is rewritten as follows, again employing the conditional independence of $\theta_1,\ldots,\theta_n$.

$$\sum_{j=1}^{n} E\left[\left\{E\left(L(\theta,\theta_j)I_{[V_{j,n}]}(X)/X_1,\ldots,X_n,\theta_j\right) - E\left(L(\theta,\theta_j)I_{[V_{j,n}]}(X)/X_1,\ldots,X_n\right)\right\}^2\right.$$

$$\left./X_1,\ldots,X_n\right] \tag{2.24}$$

$$= \sum_{j=1}^{n} E\left[\left\{E\left[\left\{E\left(L(\theta,\theta_j)/X,X_j,\theta_j\right)-E\left(L(\theta,\theta_j)/X,X_j\right)\right\}I_{[V_{j,n}]}(X)/X_1,\ldots,X_n\right]\right\}^2\right.$$

$$\left./X_1,\ldots,X_n\right] \quad . \tag{2.25}$$

Taking advantage of the conditional independence of θ and θ_j given X and X_j

$$E[L(\theta,\theta_j)/X,X_j,\theta_j] = E(\theta^2/X) - 2\theta_j E(\theta/X) + \theta_j^2 \tag{2.26}$$

and $$E[L(\theta,\theta_j)/X,X_j] = E(\theta^2/X) - 2E(\theta_j/X_j)E(\theta/X) + E(\theta_j^2/X_j) \ . \tag{2.27}$$

Since the parameters take values in a compact set, (2.26) and (2.27) are bounded, so that (2.24) is bounded with probability one by

$$\sum_{j=1}^{n} ME^2\left(I_{[V_{j,n}]}(X)/X_1,\ldots,X_n\right) \tag{2.28}$$

for some $M < \infty$. Since $\sum_{j=1}^{n} EI_{[V_{j,n}]}(X) = 1$, (2.28) is bounded by

$$M\left[\max_{1\le j\le n}\{EI_{[V_{j,n}]}(X)/X_1,\ldots,X_n\}\right]. \tag{2.29}$$

By Lemma 2, (2.29) converges to zero with probability one, establishing (2.14).

In order to complete the proof, it remains to show that the second term of (2.13) tends to zero, or

$$E(L_n/X_1,\ldots,X_n) \to 2R^* \text{ in probability} \ . \tag{2.30}$$

In order to prove (2.30) we first note that

$$E(L_n/X_1,\ldots,X_n) = E[E(L_n/X,X_1,\ldots,X_n)/X_1,\ldots,X_n] \ . \tag{2.31}$$

The right-hand side of (2.31) is equal to

$$E\left[\sum_{j=1}^{n} E\{E[(\theta-\theta_j)^2 I_{[V_{j,n}]}(X)/X,D_n]/X,X_1,\ldots,X_n\}/X_1,\ldots,X_n\right] \tag{2.32}$$

$$= E\left[\sum_{j=1}^{n} \{E(\theta^2/X) - 2E(\theta/X)E(\theta_j/X_j) + E(\theta_j^2/X_j)\} I_{[V_{j,n}]}(X)/X_1, \ldots, X_n\right]. \tag{2.33}$$

By Lemma 3, there exists a set $C \subset \mathbb{R}^d$ such that $P\{C\} = 1$, and for all $x \in C$, $X_n'(x) \to x$ with probability one. Let $x \in C$. Then

$$\sum_{j=1}^{n} \{E(\theta^2/X=x) - 2E(\theta/X=x)E(\theta_j/X_j) + E(\theta_j^2/X_j)\} I_{[V_{j,n}]}(x)$$

$$\to 2[E(\theta^2/X=x) + E^2(\theta/X=x)] \quad \text{w.p.1} \tag{2.34}$$

where we have used the continuity with probability one of $E(\theta/X)$ and $E(\theta^2/X)$. Since θ takes values in a compact set, and since the convergence in (2.34) holds for a set in $\mathbb{R}^d$ which has probability one, the Lebesgue dominated convergence theorem implies

$$E(L_n/X_1, \ldots, X_n) \to 2R^* \quad \text{w.p.1} .$$

This concludes the proof of Theorem 1.

The convergence of L_n for the case where $L(\theta, \hat{\theta})$ is a metric loss function is difficult to prove. However, in the following theorem it is shown that L_n is dominated by a random variable which converges in probability to $2R^*$.

Theorem 2: Let $(X, \theta), (X_1, \theta_1), \ldots, (X_n, \theta_n)$ be a sequence of independent identically distributed random vectors, each with distribution $F(x, \theta)$, with X taking values in $\mathbb{R}^d$ and θ taking values in $\mathbb{R}^p$. Let $L(\theta, \hat{\theta})$ be a

bounded metric on $\mathbb{R}^p$. Then, for all $\epsilon > 0$

$$P\{L_n - 2R^* \geq \epsilon\} \to 0$$

if $E[L(\theta^*(x), \theta_j)/X=x, X_j]$ is continuous with probability one for each $x \in \mathbb{R}^d$ and if a marginal density for X exists.

Proof: Since $L(\theta, \hat{\theta})$ is a metric on $\mathbb{R}^p$,

$$L(\theta, \hat{\theta}) \leq L(\theta, \theta^*) + L(\theta^*, \hat{\theta}) \tag{2.35}$$

where $\theta^* = \theta^*(X)$ is an optimal (Bayes) estimate of θ. Then

$$L_n = E\{L(\theta, \hat{\theta})/(X_1, \theta_1), \ldots, (X_n, \theta_n)\} \tag{2.36}$$

$$\leq E\{L(\theta, \theta^*)/(X_1, \theta_1), \ldots, (X_n, \theta_n)\} + E\{L(\theta^*, \hat{\theta})/(X_1, \theta_1), \ldots, (X_n, \theta_n)\}$$

$$= R^* + E\{L(\theta^*, \hat{\theta})/(X_1, \theta_1), \ldots, (X_n, \theta_n)\} . \tag{2.37}$$

The theorem will be proved if we show that

$$E\{L(\theta^*, \hat{\theta})/(X_1, \theta_1), \ldots, (X_n, \theta_n)\} \to R^* \text{ in probability} . \tag{2.38}$$

Let $L'_n = E\{L(\theta^*, \hat{\theta})/(X_1, \theta_1), \ldots, (X_n, \theta_n)\}$. The proof of (2.38) will be performed in two steps, by showing that, for each $\epsilon > 0$,

$$P\{|L'_n - E(L'_n/X_1, \ldots, X_n)| \geq \epsilon\} \to 0 \tag{2.39}$$

and $$P\{|E(L'_n/X_1, \ldots, X_n) - R^*| \geq \epsilon\} \to 0 . \tag{2.40}$$

The proof of (2.39) will be shown first. By Chebychev's inequality it suffices to show that

$$E[L'_n - E(L'_n/X_1, \ldots, X_n)]^2 \to 0 \tag{2.41}$$

or, equivalently $$E\{E[L'_n - E(L'_n/X_1,\ldots,X_n))^2/X_1,\ldots,X_n]\} \to 0. \quad (2.42)$$

Since L is bounded, the Lebesgue dominated convergence theorem can be used to show that (2.42) is true if

$$E\{[L'_n - E(L'_n/X_1,\ldots,X_n)]^2/X_1,\ldots,X_n\} \to 0 \text{ in probability.} \quad (2.43)$$

Let $V_{j,n}$ and $I_{[V_{j,n}]}(X)$ be as defined previously. Then,

$$E\{[L'_n - E(L'_n/X_1,\ldots,X_n)]^2/X_1,\ldots,X_n\}$$

$$= E\left\{\left[\sum_{j=1}^{n} \{E(L(\theta^*,\theta_j)I_{[V_{j,n}]}(X)/D_n)\right.\right.$$

$$\left.\left. - E(L(\theta^*,\theta_j)I_{[V_{j,n}]}(X)/X_1,\ldots,X_n)\}\right]^2/X_1,\ldots,X_n\right\} \quad (2.44)$$

where $D_n = \{(X_1,\theta_1),\ldots,(X_n,\theta_n)\}$. Squaring the summation in (2.44) yields

$$= E\left\{\left[\sum_{j=1}^{n} \{E(L(\theta^*,\theta_j)I_{[V_{j,n}]}(X)/D_n)\right.\right.$$

$$\left.\left. - E(L(\theta^*,\theta_j)I_{[V_{j,n}]}(X)/X_1,\ldots,X_n)\}^2\right]/X_1,\ldots,X_n\right\}$$

$$+ E\left\{\left[\sum_{i\neq j} \left\{E(L(\theta^*,\theta_j)I_{[V_{j,n}]}(X)/D_n)\right.\right.\right.$$

$$\left.\left. - E\left(L(\theta^*,\theta_j)I_{[V_{j,n}]}(X)/X_1,\ldots,X_n\right)\right\}\right.$$

$$\left. \times \left\{E\left(L(\theta^*,\theta_i)I_{[V_{i,n}]}(X)/D_n\right)\right.\right.$$

$$\left.\left. - E\left(L(\theta^*,\theta_i)I_{[V_{i,n}]}(X)/X_1,\ldots,X_n\right)\right\}\right]/X_1,\ldots,X_n\right\} . \quad (2.45)$$

The second term of (2.45), the expectation of the sum of the cross-product terms, is easily seen to be identically zero by using the conditional independence of θ_i and θ_j to bring the expectation over θ_i and θ_j inside the summation as a product of expectations. The techniques are similar to those used in the proof of Theorem 1.

The first term of (2.45) is equal to

$$\sum_{j=1}^{n} E\left\{\left[E\left\{E\left(L(\theta^*, \theta_j)/X, X_j, \theta_j\right) - E\left(L(\theta^*, \theta_j)/X, X_j\right)\right\}/X_j\right)$$

$$\times I_{[V_{j,n}]}(X)/X_1, \ldots, X_n\Big]^2/X_1, \ldots, X_n\Big\}. \tag{2.46}$$

By the hypothesis concerning the boundedness of L, there exists $M < \infty$ such that (2.46) is bounded by

$$\sum_{j=1}^{n} ME^2\left\{I_{[V_{j,n}]}(X)/X_1, \ldots, X_n\right\} \text{ w.p.1 .} \tag{2.47}$$

Since $\sum_{j=1}^{n} EI_{[V_{j,n}]}(X) = 1$, we can bound (2.47) by

$$M\left[\max_{1 \le j \le n} \{E(I_{[V_{j,n}]}(X)/X_1, \ldots, X_n)\}\right]$$

which converges to zero with probability one by Lemma 2. This establishes (2.39).

In order to complete the proof, we must show

$$E(L_n'/X_1, \ldots, X_n) \to R^* \text{ in probability .} \tag{2.48}$$

The proof of (2.48) is done by noting

$$E(L_n'/X_1,\ldots,X_n) = E\{E[L(\theta^*,\hat{\theta})/D_n]/X_1,\ldots,X_n\}$$

$$= \sum_{j=1}^{n}\left\{E\left[E\left\{E\left(L(\theta^*,\theta_j)/X,X_j,\theta_j\right)I_{[V_{j,n}]}(X)/X,X_1,\ldots,X_n\right\}/X_1,\ldots,X_n\right\}\right.$$

$$= E\left[\sum_{j=1}^{n} E\{L(\theta^*,\theta_j)/X,X_j\}I_{[V_{j,n}]}(X)/X_1,\ldots,X_n\right]. \qquad (2.49)$$

Now, by Lemma 3, there exists a set $C \subset \mathbb{R}^d$ such that $P\{C\} = 1$, and for all $x \in C$, $X_n'(x) \to x$ with probability one. Then, for all $x \in C$,

$$\sum_{j=1}^{n} E\{L(\theta^*(x),\theta_j)/X=x,X_j\}I_{V_{j,n}}(x) \to E\{L(\theta^*(x),\theta)/X=x\} \quad \text{w.p.1}, \qquad (2.50)$$

since $E\{L(\theta_0,\theta)/X\}$ is continuous with probability one by hypothesis. Then, since L is bounded, (2.50) and the Lebesgue dominated convergence theorem imply (2.48). This concludes the proof of Theorem 2.

II.3 The k-Nearest Neighbor Rule

In this section, the results of section II.2 will be generalized to the k-nearest neighbor rules where more than one neighbor of X enters into the estimate of θ. The theorems in this section will assume k is fixed, while in section II.4, k will be allowed to increase with n.

Intuitively, the advantage of using more than one nearest neighbor in the estimation process is simply the hope that the larger sample will smooth out some of the statistical variation and produce a better

estimate. The pitfall associated with increasing k stems from the fact that as more samples are used in the estimate, the k^{th}-nearest neighbor to X gets farther away from X, on the average. Hence, the strength of the statistical relationship between X and its k^{th}-nearest neighbor declines as k increases. In general then, the above arguments would indicate that if n is fixed, increasing k should result in improved performance until k reaches the point where too many samples are being used which have little or no statistical relationship to (X, θ). The difficulty lies in determining where this value of k is reached, a non-trivial problem which is quite dependent on the form of $F(x, \theta)$. In fact, Cover and Hart [6] give an example of a distribution for the discrimination problem in which the single-nearest neighbor rule performs better than any k-nearest neighbor rule where $k > 1$. The same ideas can be used to produce a similar example for the estimation problem. Basically, a distribution $F(x, \theta)$ is chosen on $\mathbb{R}^2$ which is uniform on ℓ unit discs centered at $\{(10j, 10j), j=1, \ldots, \ell\}$. In this case, with the squared-error loss function, the expected loss for a k-nearest neighbor rule is bounded by 1 if at least k data points lie on each disc. If one of the discs contains fewer than k data points, the expected loss increases dramatically. For a fixed n, as k increases, the likelihood of k surpassing the smallest number of samples on a disc increases, so that the risk increases. This discussion, while not precise, shows how a rigorous example can be constructed.

The purpose of the preceding paragraph was to point out the fact that when n is fixed, it is always possible to find distributions for which the single nearest neighbor rule can be expected to perform better than any $k > 1$ nearest neighbor rule. Generally speaking, larger values of k require larger data sets to insure adequate performance. However, the theorems in this section will show that when sufficiently large data sets are available, the use of larger values of k can cut the expected loss significantly. The added computational cost entailed by increasing k is associated with the cost of creating and maintaining a memory stack of the current k-nearest neighbors as the distance from X to each observation in the data is computed. The additional cost does not appear to be prohibitive.

The proofs of Lemmas 4, 5 and 6 and Theorem 3 will make use of the proofs in section II.2. Lemma 4 is the k-nearest neighbor rule analog of Lemma 1. The following notation will be necessary. Let $\ell_n = \binom{n}{k} = \frac{n!}{k!(n-k)!}$, so that there are ℓ_n possible different subsets of $\{X_1, \ldots, X_n\}$, where each subset contains k observations. These subsets will be denoted $S_{1,n}, \ldots, S_{\ell_n,n}$, and $T_{j,n}$ will be defined as the set of points which have $S_{j,n}$ as their set of k-nearest neighbors from the data. (Note that $\{T_{j,n}\}_{j=1}^{\ell_n}$ may be assumed to form a partition of R^d as was discussed previously.)

Lemma 4: Let $X_1, \ldots, X_n$ be a sequence of independent identically distributed random vectors taking values in $\mathbb{R}^d$, and let f be a probability density function on $\mathbb{R}^d$ corresponding to the distribution of X_1. Let $\rho(\cdot,\cdot)$ be a metric on $\mathbb{R}^d$, let $E \subset \mathbb{R}^d$ be the support of f, and let $K \subset E$ be compact with the metric ρ. Then

$$r_n = \max_{1 \le j \le \ell_n} \left\{ \sup_{x,y \in KT_{j,n}} \rho(x,y) \right\} \to 0 \quad \text{w.p.1} \tag{2.51}$$

and

$$\lambda_n = \max_{1 \le j \le \ell_n} \left\{ \mu KT_{j,n} \right\} \to 0 \quad \text{w.p.1} . \tag{2.52}$$

Proof: As in the proof of Lemma 1, it is necessary only to show that for each $\epsilon > 0$,

$$P\{r_n \ge \epsilon\} \to 0 . \tag{2.53}$$

By the compactness of K there exists a finite set of spheres $S_{\epsilon/4k}(y_i)$, $1 \le i \le m$, such that

$$K \subset \bigcup_{i=1}^{m} S_{\epsilon/4k}(y_i) . \tag{2.54}$$

As in the proof of Lemma 1, it is easy to see that if each of the spheres contains at least one of the observations $X_1, \ldots, X_n$, then each $x \in K$ has its k^{th} nearest neighbor no further away than $\epsilon/2$. Conclude that if each sphere contains at least one of the $X_1, \ldots, X_n$, then $T_{j,n}$, $1 \le j \le \ell_n$, can be contained in a sphere of radius less than $\epsilon/2$. The proof is concluded in the same manner as the proof of Lemma 1, by

showing that the probability that any of the spheres contain none of the X_i, $1 \le i \le n$, tends to zero with n tending to infinity.

Lemma 5: Let $X_1, \ldots, X_n$ and F be as in Lemma 2, and let $T_{j,n}$, $1 \le j \le \ell_n$, be as defined previously. Then

$$\max_{1 \le j \le \ell_n} P\{T_{j,n}/X_1, \ldots, X_n\} \to 0 \text{ w.p.1.} \tag{2.55}$$

Proof: The proof of this lemma is identical to the proof of Lemma 2, except that Lemma 4 must be used instead of Lemma 1.

Lemma 6: Let $X_1, \ldots, X_n$ and F be as in Lemma 2, and let $X_{(k),n}(x)$ be the k^{th} nearest neighbor to $x \in \mathbb{R}^d$ from $X_1, \ldots, X_n$. Then

$$P\{x \in \mathbb{R}^d: X_{(k),n}(x) \to x\} = 1 . \tag{2.56}$$

Proof: This proof is essentially the same as the proof of Lemma 3, except that Lemma 4 is used in place of Lemma 1.

The following theorem relates the asymptotic performance of the k-nearest neighbor rule to the performance of the optimal Bayesian estimator for squared-error loss functions.

Theorem 3: Let $(X, \theta), (X_1, \theta_1), \ldots, (X_n, \theta_n)$ be a sequence of independent identically distributed random vectors, each with distribution $F(x, \theta)$, with each X_i taking values in $\mathbb{R}^d$, and each θ_i taking values in a compact subset of $\mathbb{R}^p$. Let $L(\theta_1, \theta_2)$ be the squared-error loss function. If the distribution of X has a density and versions of $E(\theta/X)$ and $E(\|\theta\|^2/X)$

exist which are continuous on $\mathbb{R}^d$ with probability one, then

$$L_n \to (1 + \frac{1}{k})R^* \quad \text{in probability}$$

for the k-nearest neighbor rule.

Proof: We again assume $p = 1$, the case of $p > 1$ being deferred to the remarks. First, note that

$$P\{|L_n - \frac{k+1}{k} R^*| \geq \epsilon\} \leq P\{|L_n - E(L_n/X_1, \ldots, X_n)| \geq \epsilon/2\}$$

$$+ P\{|E(L_n/X_1, \ldots, X_n) - \frac{k+1}{k} R^*| \geq \epsilon/2\} . \qquad (2.57)$$

As in the proof of Theorem 1, each term of (2.57) will be shown to go to zero. The proof that

$$P\{|L_n - E(L_n/X_1, \ldots, X_n)| \geq \epsilon/2\} \to 0$$

proceeds exactly as before except that $\{T_{j,n}\}_{j=1}^{\ell_n}$ are used to partition the space into estimation regions instead of $\{V_{j,n}\}_{j=1}^{n}$ as in the proof of Theorem 1. Also, Lemma 5 must be used instead of Lemma 2, but no other serious difficulties are encountered.

The proof that

$$E(L_n/X_1, \ldots, X_n) \to \frac{k+1}{k} R^* \quad \text{in probability}$$

proceeds as follows. As defined previously, the set $S_{j,n} = \{X_{j,n}^1, \ldots, X_{j,n}^k\}$ consists of the k observations from the data which are the k-nearest neighbors to each $x \in T_{j,n}$, $1 \leq j \leq \ell_n$. Let

$\theta_{j,n}^1, \ldots, \theta_{j,n}^k$ be the parameters associated with the observations in $S_{j,n}$, and let

$$\bar{\theta}_{j,n} = \frac{1}{k} \sum_{i=1}^{k} \theta_{j,n}^i \,. \tag{2.58}$$

Note that (2.58) agrees with the definition given by (2.2) for the k-nearest neighbor rule estimate of θ for each $x \in T_{j,n}$.

Then,

$$E(L_n/X_1, \ldots, X_n) = E[E(L_n/X, X_1, \ldots, X_n)/X_1, \ldots, X_n]$$

$$= E\left[\sum_{j=1}^{\ell_n} E\{E[(\theta - \bar{\theta}_{j,n})^2 I_{[T_{j,n}]}(X)/X, D_n]/X, X_1, \ldots, X_n\}/X_1, \ldots, X_n\right]$$

$$= E\left[\sum_{j=1}^{\ell_n} \{E(\theta^2/X) - 2E(\theta/X)E(\bar{\theta}_{j,n}/S_{j,n}) + E(\bar{\theta}_{j,n}^2/S_{j,n})\} I_{[T_{j,n}]}(X)/X_1, \ldots, X_n\right]. \tag{2.59}$$

Notice that, because of the conditional independence of the parameters given the observations,

$$E(\bar{\theta}_{j,n}/S_{j,n}) = \frac{1}{k} \sum_{i=1}^{k} E(\theta_{j,n}^i / X_{j,n}^i) \tag{2.60}$$

and

$$E(\bar{\theta}_{j,n}^2/S_{j,n}) = \frac{1}{k^2} \sum_{i=1}^{k} E(\theta_{j,n}^{i\,2}/X_{j,n}) + \frac{1}{k^2} \sum_{i \neq h} E(\theta_{j,n}^i / X_{j,n}^i) E(\theta_{j,n}^h / X_{j,n}^h). \tag{2.61}$$

Now, if $X_{j,n}^i \to x$ for $1 \leq i \leq k$, then, since $E(\theta/X)$ and $E(\theta^2/X)$ are continuous with probability one, from (2.60) and (2.61) it can be seen that

$$\sum_{j=1}^{n} E(\bar{\theta}_{j,n}/S_{j,n})I_{[T_{j,n}]}(x) \to E(\theta/X=x) \quad \text{w.p.1,} \tag{2.62}$$

and

$$\sum_{j=1}^{n} E(\bar{\theta}^2_{j,n}/S_{j,n})I_{[T_{j,n}]}(x) \to \frac{1}{k} E(\theta^2/X=x) + \frac{k-1}{k} E^2(\theta/X=x) \quad \text{w.p.1.} \tag{2.63}$$

By Lemma 6, there exists a set $C \subset \mathbb{R}^d$ such that $P\{C\} = 1$ and, for all $x \in C$, the k^{th}-nearest neighbor to x converges to x with probability one. Then, for each $x \in C$

$$\sum_{j=1}^{\ell_n} \{E(\theta^2/X=x) - 2E(\theta/X=x)E(\bar{\theta}_{j,n}/S_{j,n}) + E(\bar{\theta}^2_{j,n}/S_{j,n})\}I_{[T_{j,n}]}(x)$$

$$\to (1 + \frac{1}{k})[E(\theta^2/X=x) - E^2(\theta/X=x)] \quad \text{w.p.1.} \tag{2.64}$$

Since θ takes values in a compact set, and (2.64) holds on a set which has probability one, the Lebesgue dominated convergence theorem implies that (2.59) converges to $(1 + 1/k)R^*$ with probability one. This completes the proof.

The advantage of using more than one nearest neighbor in the estimation process when large data sets are available is made apparent by Theorem 3. In fact, for large values of k the risk approaches being half that of the single-nearest neighbor rule.

II.4 The k_n-Nearest Neighbor Rule

In many applications, the data set available to the statistician is continually increasing in size. In cases like this, a fixed value of

k will eventually become too small to take fullest advantage of the size of the data set. An obvious solution to this problem is to allow k to be an increasing function of n. In this section conditions will be given for which the performance of the k_n-nearest neighbor rule is asymptotically optimal for squared-error loss functions.

The following lemma is another generalization of Lemma 1. We now let $\ell_n = \binom{n}{k_n}$ so that ℓ_n is the number of different subsets that can be obtained from $\{X_1,\ldots,X_n\}$ where each subset contains k_n elements. These subsets will be denoted $S_{1,n},\ldots,S_{\ell_n,n}$. We will define $T_{j,n}$ for $1 \le j \le \ell_n$ as the set of $x \in \mathbb{R}^d$ such that $S_{j,n}$ is the set which contains the k_n-nearest neighbors to x from $X_1,\ldots,X_n$. Once again we note that the sets $T_{j,n}$, $1 \le j \le \ell_n$, can be modified to form a partition of $\mathbb{R}^d$, as discussed previously.

Lemma 7: Let $X_1,\ldots,X_n$ be independent identically distributed random vectors taking values in $\mathbb{R}^d$, and let f be a probability density function on $\mathbb{R}^d$ corresponding to the distribution of X_1. Let $\rho(\cdot,\cdot)$ be a metric on $\mathbb{R}^d$, let E be the support of f, and let $K \subset E$ be compact with the metric ρ. Then, if $k_n/n \to 0$,

$$r_n = \max_{1 \le j \le \ell_n} \left\{ \sup_{x,y \in KT_{j,n}} \rho(x,y) \right\} \to 0 \quad \text{w.p.1}$$

and

and $$\lambda_n = \max_{1 \le j \le \ell_n} \{\mu KT_{j,n}\} \to 0 \text{ w.p.1} . \tag{2.65}$$

Proof: Let $\epsilon > 0$ be given. As in the proof of Lemma 1, since r_n is monotonically decreasing, it suffices to show

$$P\{r_n \ge \epsilon\} \to 0 .$$

Since K is compact there exists a finite set of points $\{y_1, \ldots, y_m\}$ such that

$$K \subset \bigcup_{i=1}^{m} S_{\epsilon/4}(y_i) .$$

Assume that each of the spheres $S_{\epsilon/4}(y_i)$, $1 \le i \le n$, contains at least k_n of the observations $X_1, \ldots, X_n$. Then the distance from any $x \in K$ to its k_n^{th} nearest neighbor is less than $\epsilon/4$. This implies that each set $KT_{j,n}$ can be contained in a sphere with radius less than or equal to $\epsilon/2$. It remains to show that the probability that each sphere contains at least k_n observations converges to one, or that the converse converges to zero.

$$P\{r_n \ge \epsilon\} \le P\left[\bigcup_{i=1}^{m} \{S_{\epsilon/4}(y_i) \text{ contains fewer than } k_n \text{ observations}\}\right]$$

$$\le \sum_{i=1}^{m}\left(\sum_{j=0}^{k_n-1} \binom{n}{j} p_i^{\,j}(1-p_i)^{n-j}\right) \tag{2.66}$$

where $p_i = \int_{S_{\epsilon/4}(y_i)} f(x)dx$, $1 \le i \le m$. Feller [13, p.51] shows that if $k_n < np_i$, then

$$\sum_{j=0}^{k_n-1} \binom{n}{j} p_i^j (1-p_i)^{n-j} \le \frac{(n-k_n+1)p_i}{(np_i-k_n+1)^2} \tag{2.67}$$

$$= \frac{\frac{n-k_n+1}{n^2} p_i}{p_i^2 - \frac{2(k_n+1)}{n} p_i + \frac{(k_n+1)^2}{n^2}} \quad . \tag{2.68}$$

Since $k_n/n \to 0$, $k_n < np_i$ for n sufficiently large, and (2.68) converges to zero as $n \to \infty$ almost surely. The convergence of λ_n follows immediately, which proves the lemma.

The proofs of the following two lemmas follow from Lemma 7 in the same way that Lemmas 2 and 3 follow from Lemma 1.

Lemma 8: Let $X_1, \ldots, X_n$ and F be as in Lemma 2, and let $T_{j,n}$, $1 \le j \le \ell_n$, be as defined for k_n-nearest neighbor rules. Then

$$\max_{1 \le j \le \ell_n} P\{T_{j,n} / X_1, \ldots, X_n\} \to 0 \quad \text{w.p.1} \; . \tag{2.69}$$

Lemma 9: Let $X_1, \ldots, X_n$ and F be as in Lemma 2, and let $X_{(k_n),n}(x)$ be the k_n^{th} nearest neighbor to $x \in \mathbb{R}^d$ from $X_1, \ldots, X_n$. Then

$$P\{x \in \mathbb{R}^d : X_{(k_n),n}(x) \to x\} = 1 \; . \tag{2.70}$$

Theorem 4 proves the asymptotic optimality of k_n-nearest neighbor rules in estimation for squared-error loss functions.

Theorem 4: Let (X, θ), $(X_1, \theta_1), \ldots, (X_n, \theta_n)$ be a sequence of independent identically distributed random vectors, each with distribution $F(x, \theta)$, with each observation X_i taking values in $\mathbb{R}^d$ and each θ_i taking values in a compact subset of $\mathbb{R}^p$. Let $L(\theta_1, \theta_2)$ be the squared-error loss function. If the distribution of X has a density and there exist versions of $E(\|\theta\|^2/X)$ and $E(\theta/X)$ which are continuous with probability one, $k_n \to \infty$ and $k_n/n \to 0$, then

$$L_n \to R^* \text{ in probability}$$

for the k_n-nearest neighbor rule.

Proof: The proof is quite similar to the proof of Theorem 3, except that Lemmas 8 and 9 are used instead of Lemmas 5 and 6, and we must show that

$$E(L_n/X_1, \ldots, X_n) \to R^* \text{ in probability.} \tag{2.71}$$

As in the proof of Theorem 3, let $\theta_{j,n}^1, \ldots, \theta_{j,n}^{k_n}$ be the parameters associated with the k_n observations in $S_{j,n}$. Then, as before, if we define

$$\bar{\theta}_{j,n} = \frac{1}{k_n} \sum_{i=1}^{k_n} \theta_{j,n}^i \tag{2.72}$$

then $\bar{\theta}_{j,n}$ is the k_n-nearest neighbor rule estimate of θ for each $x \in T_{j,n}$. As before,

$$E(L_n/X_1, \ldots, X_n)$$

$$= E\Big[\sum_{j=1}^{\ell_n} \{E(\theta^2/X) - 2E(\theta/X)E(\bar{\theta}_{j,n}/S_{j,n}) + E(\bar{\theta}^2_{j,n}/S_{j,n})\} I_{[T_{j,n}]}(X)$$

$$/X_1, \ldots, X_n\Big]. \tag{2.73}$$

The same techniques that were used in the proof of Theorem 3 can be used to show that

$$\sum_{j=1}^{\ell_n} \{E(\theta^2/X=x) - 2E(\theta/X=x)E(\bar{\theta}_{j,n}/S_{j,n}) + E(\bar{\theta}^2_{j,n}/S_{j,n})\} I_{[T_{j,n}]}(x)$$

$$\rightarrow \Big[E(\theta^2/X=x) - E^2(\theta/X=x)\Big] \text{ w.p.1} \tag{2.74}$$

for each x in a set which has probability one. Then, since θ takes values in a compact set, the Lebesgue dominated convergence theorem implies that

$$E(L_n/X_1, \ldots, X_n) \rightarrow R^* \text{ w.p.1}, \tag{2.75}$$

which completes the proof of Theorem 4.

II.5 Nearest Neighbor Rules With Unequal Weighting

The theorems in the preceding sections of this chapter are concerned with the asymptotic performance of k-nearest neighbor rules which assign equal weight to each of the k-nearest neighbors used in forming the estimate. In some cases the statistician may want to make use of an unequal weighting scheme so that observations which are closer to X contribute more heavily to the estimate of θ than observations which are farther away. In such a scheme, the estimate $\hat{\theta}$ would be given by

$$\hat{\theta} = \sum_{i=1}^{k} a_i \theta_{(i)} \tag{2.76}$$

where $\theta_{(i)}$ is the parameter associated with the observation $X_{(i)}$ which is the i^{th} closest observation to X in the data set. Substituting the version given by (2.76) for $\bar{\theta}_{j,n}$ in equations (2.62), (2.63) and (2.73) and completing the remaining steps in those proofs yields

$$L_n \to (1 + \sum_{1}^{k} a_i^2) R^* \text{ in probability} \tag{2.77}$$

where we have assumed $\sum_{1}^{k} a_i = 1$ which, of course, we can do without loss of generality. Lagrangian techniques can be employed to show that $\sum_{1}^{k} a_i^2$ is minimized, subject to the constraint that $\sum_{1}^{k} a_i = 1$, when the a_i are all equal to $1/k$. The conclusion is that large sample performance suffers when unequal weighting schemes are used. This suggests that rather than decrease the weighting of the nearest neighbors which are farther from X, the statistician may get better results by using equal weighting with a smaller value for k.

II.6 Remarks

The theorems in this chapter have all required continuity with probability one of $E(\|\theta\|^2/X)$ and $E(\theta/X)$. A function g of a random variable X is continuous with probability one if

$$P\{X \in D\} = 0$$

where D is the set of discontinuity points of g. We note that if a joint density $f(x, \theta)$ exists for $F(x, \theta)$, then one version of $E(\theta/X=x)$ is given by

$$E(\theta/X = x) = \int \theta \frac{f(x, \theta)}{f(x)} d\theta \tag{2.78}$$

if $$f(x) = \int f(x, \theta) d\theta > 0 \ . \tag{2.79}$$

(See, for example, Breiman [7], Ch. 4.)

Similarly, when (2.79) holds,

$$E(\|\theta\|^2/X=x) = \int \|\theta\|^2 \frac{f(x, \theta)}{f(x)} d\theta \ . \tag{2.80}$$

From (2.78) and (2.80) it can be seen that if $f(x, \theta)$ is μ- almost everywhere continuous in x, μ denoting Lebesgue measure on $\mathbb{R}^d$, then the versions of $E(\theta/X)$ and $E(\|\theta\|^2/X)$ given by (2.78) and (2.80) are continuous with probability one.

The proof of Theorem 1 for p dimensional parameter spaces is the same in all major respects, the only minor difficulty being the question of the convergence of

$$\sum_{j=1}^{n} E\{\|\theta-\theta_j\|^2/X=x, X_j\} I_{[V_{j,n}]}(x) \ .$$

To establish Theorem 1 it will be sufficient to show that if $X'_n(x) \to x$ with probability one, then

$$\sum_{j=1}^{n} E\{\|\theta-\theta_j\|^2/X=x, X_j\} I_{[V_{j,n}]}(x)$$

$$\rightarrow 2\{(\|\theta\|^2/X=x) - \|E(\theta/X=x)\|^2\} \text{ w.p.1.} \tag{2.81}$$

Expanding the expectation on the left side of (2.81),

$$E\{\|\theta-\theta_j\|^2/X=x,X_j\} = E\{\|\theta\|^2/X=x\} - 2E\{(\theta,\theta_j)/X=x,X_j\}$$
$$+ E\{\|\theta_j\|^2/X_j\} . \tag{2.82}$$

In order to prove (2.81), we first show

$$\sum_{j=1}^{n} E\{(\theta,\theta_j)/X=x,X_j\} I_{[V_{j,n}]}(x) \rightarrow \|E(\theta/X=x)\|^2 . \tag{2.83}$$

The left side of (2.83) is equal to

$$\sum_{j=1}^{n}\sum_{i=1}^{p} E(\theta^i/X=x)\ E(\theta_j^i/X_j)\ I_{[V_{j,n}]}(x)$$

$$= \sum_{i=1}^{p}\sum_{j=1}^{n} E(\theta^i/X=x)\ E(\theta_j^i/X_j)\ I_{[V_{j,n}]}(x) , \tag{2.84}$$

where θ^i and θ_j^i are the i^{th} components of θ and θ_j respectively. Since $E(\theta/X)$ is assumed continuous with probability one, and $X_n'(x) \rightarrow x$ with probability one, (2.84) can be seen to converge to

$$\sum_{i=1}^{p} E^2(\theta^i/X=x) \tag{2.85}$$

with probability one. But (2.85) is equal to the right-hand side of (2.83), as we intended to prove.

Noting that $E\{\|\theta\|^2/X\}$ was assumed continuous with probability one, the convergence of $X_n'(x) \rightarrow x$ with probability one yields

$$\sum_{j=1}^{n} E\{\|\theta_j\|^2/X_j\} I_{[V_{j,n}]}(x) \to E\{\|\theta\|^2/X=x\} \text{ w.p.1 .} \qquad (2.86)$$

The desired convergence in (2.81) follows immediately from (2.82), (2.83), and (2.86). The remainder of the proof of Theorem 1 for $p > 1$ uses predominantly the same techniques as the proof for $p = 1$. For k- and k_n-nearest neighbor rules, the same remarks will still hold, but the notation and algebra become more involved since we must now deal with the average of the parameters of the nearest neighbors instead of just θ'_n.

The proofs in this chapter have employed the condition that the distribution of X have a density. This condition can easily be weakened to the requirement that the distribution for X be continuous. This permits the distribution of X to have a singular continuous part.

The proof of Theorem 2 contains the tacit assumption that a Bayes rule exists, which may not always be the case. However, for any $\epsilon > 0$ we can find a rule which has risk less than $R^* + \epsilon$. Such rules are known as ϵ-Bayes rules (see, for example, Ferguson [15]). Let $\epsilon > 0$ be given, and let $\theta'(x)$ be a rule with risk $R' \leq R^* + \epsilon$. Then, the techniques used in the proof of Theorem 2 can be used to show that

$$P\{L_n - 2R' \geq \epsilon\} \to 0 \qquad (2.87)$$

But $\{L_n - 2R' \geq \epsilon\} \supseteq \{L_n - 2R^* \geq 3\epsilon\}$,

so that (2.87) implies

$$P\{L_n - 2R^* \geq 3\epsilon\} \to 0 .$$

Hence Theorem 2 is true whether or not a Bayes rule exists.

As a final remark concerning the convergence of L_n, we will give an example concerning the rate at which $P\{|L_n - R| \geq \epsilon\}$ goes to zero, where R is $(1 + 1/k)R^*$. Consider the case where $\theta \in \{1,2,\ldots,\ell\}$ and $P\{\theta = i\} = 1/\ell$, $1 \leq i \leq \ell$. Let $f(x/\theta)$ be defined as

$$f(x/\theta=i) = \begin{cases} 1, & i - \frac{1}{2} < x \leq i + \frac{1}{2} \\ 0, & \text{otherwise.} \end{cases}$$

Then $R^* = 0$ since given X, the value of θ is determined. Hence $R = (1 + 1/k)R^* = 0$. But for any fixed n, we can increase ℓ to spread out the distribution so that when $\ell > n$, $L_n > \ell - n/\ell > \epsilon$ for ℓ sufficiently large. We can conclude that the rate of convergence of L_n can be arbitrarily slow in the sense that for any n and $\epsilon > 0$, there exist distributions $F(x,\theta)$ for which $P\{|L_n - R| > \epsilon\} = 1$. This example can easily be modified so that θ is a continuous parameter by allowing θ to take values in small intervals surrounding $1,2,\ldots,\ell$. R^* is then no longer zero, but can be made arbitrarily small by appropriately choosing the size of the intervals for θ. The rest of the example follows easily.

III. THE EVALUATION OF FINITE SAMPLE PERFORMANCE FOR AN ESTIMATION RULE

III.1 Motivation for Chapter III

The evaluation of the finite sample performance of an estimation rule is the process of estimating the value of L_n, the expected risk conditioned on a data set containing n observations. The problem is analogous to error estimation for discrimination rules. Toussaint [5] has compiled a bibliography of papers on error estimation in which he includes a brief discussion of the most commonly employed techniques. Most of these techniques also find application in estimating L_n for estimation rules.

In this chapter we will be concerned with finding an upper bound for

$$P\{|L_n - \hat{L}_n| \geq \epsilon\} \tag{3.1}$$

where $\hat{L}_n$ denotes the estimate of L_n. In the preceding chapter, we were concerned merely with the question of the convergence of L_n without bothering to investigate the rate of such convergence. Indeed, as was pointed out at the close of Chapter II, the rate of convergence of L_n can be arbitrarily slow, depending on the structure of $F(x, \theta)$. In this chapter, the rate at which (3.1) goes to zero will be of great importance, the reason being that if a minimal rate can be established which is

independent of $F(x, \theta)$, it will enable the statistician to determine for a given n and ϵ the value of $\delta(n, \epsilon)$ for which

$$P\{|L_n - \hat{L}_n| \geq \epsilon\} \leq \delta(n, \epsilon) . \tag{3.2}$$

Such a bound is termed distribution-free and it is quite useful to the statistician since it enables him to determine how much confidence he can place in his estimate of L_n.

As in the preceding chapter, we let $D_n = \left((X_1, \theta_1), \ldots, (X_n, \theta_n)\right)$ be a sequence of independent identically distributed random vectors, each with distribution $F(x, \theta)$, where each observation X_i takes values in $\mathbb{R}^d$ and each parameter θ_i takes values in a parameter space $\mathbb{R}^p$. Let (X, θ) be independent of the data D_n and distributed with $F(x, \theta)$. In this chapter, we will be concerned with a class of estimation rules somewhat broader than the k-nearest neighbor rules considered in Chapter II.

We will say that a randomized estimation rule is one which produces an estimate $\hat{\theta}$ which is a random vector chosen from a distribution which depends on X and D_n. We will assume that the rule is described by a jointly measurable function

$$\delta: \mathbb{R}^d \times (\mathbb{R}^d \times \mathbb{R}^p)^n \times \mathbb{R}^p \to [0,1] \tag{3.3}$$

where, for each X and D_n, δ_{X,D_n} is a distribution function on $\mathbb{R}^p$. We can define the joint distribution of $\hat{\theta}$, (X, θ), and D_n by

$$P\{D_n \in A, (X, \theta) \in B, \hat{\theta} \in C\}$$

$$= \int_A \int_B \int_C d\,\delta_{X,D_n}(\hat{\theta})\, dF(x,\theta)\, dF^n(x,\theta)$$

where A, B, and C are Borel sets in $(\mathbb{R}^d \times \mathbb{R}^p)^n$, $(\mathbb{R}^d \times \mathbb{R}^p)$, and $\mathbb{R}^p$ respectively.

Actually, a statistician may use a sequence $\{\delta_n\}$ of such randomized estimation rules whose properties will vary with n, the amount of data available. For the purposes of this chapter, however, we will only be concerned with the properties of the particular rule used with D_n, rather than with the whole sequence. It will be shown that if the rule is one of a certain class of rules to be described, then bounds of the type (3.2) can be given for two particular estimates of L_n.

The following terminology will be used to describe the rules in which we will be interested. A rule will be called <u>local with parameter r</u> if the distribution δ_{X,D_n} is, with probability one, a function only of X and $(X_{(1)}, \theta_{(1)}), \ldots, (X_{(r)}, \theta_{(r)})$, where $X_{(i)}$ is the i^{th} closest observation to X from D_n. For example, let

$$D_{n,i} = \left((X_1, \theta_1), \ldots, (X_{i-1}, \theta_{i-1}), (X_{i+1}, \theta_{i+1}), \ldots, (X_n, \theta_n)\right). \quad (3.4)$$

Then, if δ is local with parameter $r < n-1$,

$$\delta_{X,D_n}(\theta) = \delta_{X,D_{n,i}}(\theta)$$

for all $\theta \in \mathbb{R}^p$ whenever X_i is not one of the r-nearest neighbors to X from D_n. The intent here is simply to point out that, since δ depends

only on X and its r-nearest neighbors, any changes made in the data which do not affect the r-nearest neighbors, also do not affect the distribution of $\hat{\theta}$. Finally, we will call an estimation rule <u>symmetric</u> if, for any given X, permutations of the order of the elements of D_n do not affect δ_{X,D_n}.

An example of a local estimation rule which is symmetric is the k-nearest neighbor rule when $F(x,\theta)$ is nonatomic. Note however, that when $F(x,\theta)$ possesses atoms, certain difficulties may arise. In some cases there may be positive probability that several observations in the data are at the same distance from X as its k^{th}-nearest neighbor. In order to have a local rule then, it is necessary to devise some method to prevent some of these observations from contributing to the estimate. This problem will be discussed in more detail in the remarks at the close of this chapter.

The results to be presented are for two different estimates of L_n, which will be referred to as the deleted estimate and the holdout estimate. (Toussaint provides references for previous work on both of these estimates. In Toussaint's paper, the deleted estimate is called the U method, and the holdout estimate is called the H method.) The deleted estimate, $\hat{L}_n$, is defined as

$$\hat{L}_n = \frac{1}{n}\sum_{1}^{n} L(\theta_i, \hat{\theta}_i) \tag{3.5}$$

where $\hat{\theta}_i$ is the estimate produced by δ when X_i is the observation and $D_{n,i}$ as defined in (3.4) is the data. The deleted estimate, then, is just the average loss incurred when each (X_i, θ_i) in turn is deleted from the data set and δ is used with X_i and the remaining n-1 elements of D_n to estimate θ_i.

The holdout estimate, L'_n, is defined as

$$L'_n = \frac{1}{\ell} \sum_{n-\ell+1}^{n} L(\theta_i, \theta'_i) \tag{3.6}$$

where, for $n-\ell<i\leq n$, θ'_i is the estimate produced by δ with observation X_i and data $D_{n-\ell}$, where

$$D_{n-\ell} = \left((X_1, \theta_1), \ldots, (X_{n-\ell}, \theta_{n-\ell})\right). \tag{3.7}$$

The holdout estimate is the average loss when the first $n-\ell$ elements of D_n are used to estimate the parameters for the remaining ℓ observations. The first $n-\ell$ elements of D_n are usually called the training set, and the remaining elements are known as the test set. In the results presented in section III.2, the size of the test set, ℓ, will be allowed to increase with n.

III.2 Distribution-Free Bounds for Deleted and Holdout Estimates

Theorem 5 will establish a distribution-free bound on $P\{|L_n - \hat{L}_n| \geq \varepsilon\}$, while Theorem 6 establishes a similar bound on

$P\{|L_n - L_n'| \geq \epsilon\}$. We assume that (X, θ) and the data set D_n are as defined previously, and let (X_0, θ_0) be independent of (X, θ) and the data and distributed as $F(x, \theta)$. The result given by Theorem 5 is similar in nature to a result of Rogers and Wagner [9] for deleted estimates in discrimination problems.

Theorem 5: Let δ be a symmetric estimation rule which is local with parameter k. Let L be a loss function which satisfies

$$\sup_{\theta, \hat{\theta}} L(\theta, \hat{\theta}) = M < \infty \tag{3.8}$$

Then $$P\{|L_n - \hat{L}_n| \geq \epsilon\} \leq \frac{M^2}{\epsilon^2}\left[\frac{1+6k}{n} - \frac{k^2}{n^2} - \frac{k^2}{(n-1)^2}\right] \tag{3.9}$$

Proof: By Chebychev's inequality,

$$P\{|L_n - \hat{L}_n| \geq \epsilon\} \leq E(L_n - \hat{L}_n)^2/\epsilon^2$$

$$= (EL_n^2 - 2EL_n\hat{L}_n + E\hat{L}_n^2)/\epsilon^2 \tag{3.10}$$

Let $\hat{\theta}$, $\hat{\theta}_0$, and $\hat{\theta}_i$ be independent random vectors where $\hat{\theta}$ has distribution δ_{X, D_n}, $\hat{\theta}_0$ has distribution δ_{X_0, D_n}, and, for each i, $1 \leq i \leq n$, $\hat{\theta}_i$ has distribution $\delta_{X_i, D_{n,i}}$.

$$E(L_n^2) = E[E^2(L(\theta, \hat{\theta})/D_n)] \tag{3.11}$$

$$= E[E(L(\theta, \hat{\theta})/D_n)E(L(\theta_0, \hat{\theta}_0)/D_n)]$$

$$= E[L(\theta, \hat{\theta})L(\theta_0, \hat{\theta}_0)] \tag{3.12}$$

$$E(L_n \hat{L}_n) = E\left\{E[L(\theta, \hat{\theta})/D_n] \frac{1}{n} \sum_1^n L(\theta_i, \hat{\theta}_i)\right\} \tag{3.13}$$

$$= \frac{1}{n} \sum_1^n E\left\{L(\theta_i, \hat{\theta}_i) E[L(\theta, \hat{\theta})/D_n]\right\}$$

$$= \frac{1}{n} \sum_1^n E\left\{E[L(\theta_i, \hat{\theta}_i) L(\theta, \hat{\theta})/D_n]\right\}$$

$$= E[L(\theta_1, \hat{\theta}_1) L(\theta, \hat{\theta})] . \tag{3.14}$$

The last step above was made possible by the assumed symmetry of δ, which implies that

$$E\left\{E[L(\theta_i, \hat{\theta}_i) L(\theta, \hat{\theta})/D_n]\right\} = E\left\{E[L(\theta_j, \hat{\theta}_j) L(\theta, \hat{\theta})/D_n]\right\} \tag{3.15}$$

for all i,j. Finally,

$$E(\hat{L}_n)^2 = \frac{1}{n^2} E\left[\sum_1^n L^2(\theta_i, \hat{\theta}_i) + \sum_{i \neq j} L(\theta_i, \hat{\theta}_i) L(\theta_j, \hat{\theta}_j)\right]$$

$$= \frac{1}{n^2}\left[\sum_1^n E[L^2(\theta_i, \hat{\theta}_i)] + \sum_{i \neq j} E[L(\theta_i, \hat{\theta}_i) L(\theta_j, \hat{\theta}_j)]\right]$$

$$= \frac{1}{n}\left(E[L^2(\theta_1, \hat{\theta}_1)] + (n-1) E[L(\theta_1, \hat{\theta}_1) L(\theta_2, \hat{\theta}_2)]\right), \tag{3.16}$$

where the last step is due to symmetry. Combining (3.12), (3.14), and (3.16) above, we have

$$E(L_n - \hat{L}_n)^2 = E[L(\theta, \hat{\theta}) L(\theta_0, \hat{\theta}_0)]$$

$$-2 E[L(\theta_1, \hat{\theta}_1) L(\theta, \hat{\theta})]$$

$$+ E[L(\theta_1, \hat{\theta}_1) L(\theta_2, \hat{\theta}_2)]$$

$$+ \frac{1}{n}\left\{E[L^2(\theta_1, \hat{\theta}_1)] - E[L(\theta_1, \hat{\theta}_1) L(\theta_2, \hat{\theta}_2)]\right\} . \tag{3.17}$$

Let $D_n' = (X_0, \theta_0), (X_2, \theta_2), \ldots, (X_n, \theta_n)$ and let θ' and θ_0' be independent random vectors which are also independent of $\hat{\theta}$, $\hat{\theta}_0$, and $\hat{\theta}_i$ conditioned on X, (X_0, θ_0), and D_n. Let θ' be distributed with $\delta_{X, D_n'}$ and θ_0' with $\delta_{X_0, D_{n,1}}$. Then it is not difficult to see that

$$E[L(\theta, \hat{\theta})L(\theta_0, \hat{\theta}_0)] - E[L(\theta, \hat{\theta})L(\theta_1, \hat{\theta}_1)]$$

$$= E[L(\theta, \hat{\theta})L(\theta_0, \hat{\theta}_0)] - E[L(\theta, \theta')L(\theta_0, \theta_0')] . \qquad (3.18)$$

Now, under the hypothesis that $X, (X_0, \theta_0)$, and D_n are such that $\delta_{X, D_n} = \delta_{X, D_n'}$ and $\delta_{X_0, D_n} = \delta_{X_0, D_{n,1}}$, (3.18) is equal to zero since it is the expectation of the difference of two identically distributed random variables. Hence (3.18) can be bounded from above by

$$M^2 E\{I_{[\delta_{X, D_n} \neq \delta_{X, D_n'}]} + I_{[\delta_{X_0, D_n} \neq \delta_{X_0, D_{n,1}}]}\}$$

$$= M^2[P\{\delta_{X, D_n} \neq \delta_{X, D_n'}\} + P\{\delta_{X_0, D_n} \neq \delta_{X_0, D_{n,1}}\}] . \qquad (3.19)$$

Next, we let $D_{n,1}' = (X, \theta), (X_3, \theta_3), \ldots, (X_n, \theta_n)$, and let θ'' and θ_1'' be conditionally independent random vectors given X and D_n, and let θ'' have distribution $\delta_{X, D_{n,2}}$ and θ_1'' have distribution $\delta_{X_1, D_n'}$. Then, using the same techniques, we can show that

$$E[L(\theta_1, \hat{\theta}_1)L(\theta_2, \hat{\theta}_2)] - E[L(\theta, \hat{\theta})L(\theta_1, \hat{\theta}_1)]$$

$$\leq M^2[P\{\delta_{X,D_n} \neq \delta_{X,D_{n,2}}\} + P\{\delta_{X_1,D_{n,1}} \neq \delta_{X_1,D'_{n,1}}\}] . \quad (3.20)$$

The remaining term of (3.17) can be bounded as follows:

$$\frac{1}{n}\{E[L^2(\theta_1,\hat{\theta}_1) - L(\theta_1,\hat{\theta}_1)L(\theta_2,\hat{\theta}_2)]\} \leq M^2/n . \quad (3.21)$$

Now, since δ is a local rule,

$$P\{\delta_{X,D_n} = \delta_{X,D'_n}\} \geq \left(\frac{n-k}{n}\right)^2 \quad (3.22)$$

since δ_{X,D_n} is not changed if X_1 is not one of the nearest neighbors to X from D_n, and X_0 is not closer to X than its k^{th}-nearest neighbor from D_n. From (3.22),

$$P\{\delta_{X,D_n} \neq \delta_{X,D'_n}\} \leq 1 - \left(\frac{n-k}{n}\right)^2 . \quad (3.23a)$$

Similarly,

$$P\{\delta_{X_0,L_n} \neq \delta_{X_0,D_{n,1}}\} \leq \frac{k}{n} \quad (3.23b)$$

$$P\{\delta_{X,D_n} \neq \delta_{X,D_{n,2}}\} \leq \frac{k}{n} \quad (3.23c)$$

$$P\{\delta_{X_1,D_{n,1}} \neq \delta_{X_1,D'_{n,1}}\} \leq 1 - \left(\frac{n-k-1}{n-1}\right)^2 . \quad (3.23d)$$

Combining (3.19), (3.20), (3.21), and (3.23), we arrive at the statement of the theorem.

For the holdout estimate, L'_n, we have the following theorem.

Theorem 6: Let δ be a symmetric estimation rule which is local with parameter k. Let L be a loss function which satisfies

$$\sup_{\theta, \hat{\theta}} L(\theta, \hat{\theta}) = M < \infty .$$

Then

$$P\{|L_n - L_n'| \geq \epsilon\} \leq 2e^{-\ell \epsilon^2/4M} + \frac{2M}{\epsilon}\left\{1 - \left(1 - \frac{k}{n}\right)^{\ell}\right\} \tag{3.24}$$

where ℓ is the number of samples held out.

Proof: Let $L_{n-\ell} = E[L(\theta, \theta')/D_n]$, where θ' is an independently chosen random vector with distribution $\delta_{X, D_{n-\ell}}$ and

$$D_{n-\ell} = \left((X_1, \theta_1), \ldots, (X_{n-\ell}, \theta_{n-\ell})\right) .$$

Then,

$$P\{|L_n - L_n'| \geq \epsilon\} \leq P\{|L_n - L_{n-\ell}| \geq \epsilon/2\} + P\{|L_{n-\ell} - L_n'| \geq \epsilon/2\} . \tag{3.25}$$

The first term on the right-hand side of (3.25) is bounded by Markov's inequality:

$$P\{|L_n - L_{n-\ell}| \geq \epsilon/2\} \leq \frac{2}{\epsilon} E|L_n - L_{n-\ell}|$$

$$= \frac{2}{\epsilon} E\left|E(L(\theta, \hat{\theta}) - L(\theta, \theta')/D_n)\right| . \tag{3.26}$$

By the same argument used in the proof of Theorem 5, (3.26) can be bounded by

$$\frac{2M}{\epsilon} E\{I_{[\delta_{X,D_n} \neq \delta_{X,D_{n-\ell}}]}\}$$

$$= \frac{2M}{\epsilon} P\{\delta_{X,D_n} \neq \delta_{X,D_{n-\ell}}\} . \tag{3.27}$$

For the second term,

$$P\left\{ |L_{n-\ell} - L'_n| \geq \frac{\epsilon}{2} \right\} = P\left\{ |L_{n-\ell} - \frac{1}{\ell} \sum_{i=0}^{\ell-1} L(\theta_{n-i}, \theta'_{n-i}) | \geq \frac{\epsilon}{2} \right\} .$$

Note that for each i,

$$E\left(L(\theta_{n-i}, \theta'_{n-i}) / D_n \right) = L_{n-\ell} .$$

Then, by Hoeffding's inequality [10],

$$P\left\{ |L_{n-\ell} - L'_n| \geq \frac{\epsilon}{2} \right\} \leq 2e^{-\ell\epsilon^2/4M} . \tag{3.28}$$

The inequality in (3.24) now follows from (3.25), (3.27), and (3.28).

For the k-local rules, $P\{\delta_{X,D_n} \neq \delta_{X,D_{n-\ell}}\}$ is bounded by the probability that not all of the k-nearest neighbors to X are in the first $n-\ell$ elements of D_n. Hence

$$P\{\hat{\theta}(n) \neq \hat{\theta}(n-\ell)\} \leq 1 - \frac{\binom{n-\ell}{k}}{\binom{n}{k}}$$

$$\leq 1 - \left(1 - \frac{k}{n}\right)^{\ell} .$$

This proves the theorem.

III.3 Remarks

Noting that both of the theorems in this chapter give bounds which depend on $M = \sup_{\theta, \hat{\theta}} L(\theta, \hat{\theta})$, some justification would appear to be in order. Consider the following example. Let the observations $X, X_1, \ldots, X_n$ take values in $\mathbb{R}$, and let $\Theta = [0,1]$, $\Theta' = [0,m]$ where $m > 1$. Suppose that $(X, \theta), (X_1, \theta_1), \ldots, (X_n, \theta_n)$ are independent and identically distributed on $\mathbb{R} \times \Theta$ with distribution $F(x, \theta)$. Also, suppose that $(X, \theta'), (X_1, \theta'_1), \ldots, (X_n, \theta'_n)$ are independent and identically distributed on $\mathbb{R} \times \Theta'$ with distribution $F(x, \theta/m)$. Let $L(\theta, \hat{\theta}) \triangleq |\theta - \hat{\theta}|$ be the loss function for both parameter spaces. By this construction, it is not difficult to show that

$$P\{L_n \leq x\} = P\{L'_n \leq mx\} \tag{3.27}$$

and

$$P\{\hat{L}_n \leq x\} = P\{\hat{L}'_n \leq mx\} \tag{3.28}$$

where L_n and $\hat{L}_n$ are associated with $\mathbb{R} \times \Theta$ and L'_n and $\hat{L}'_n$ are associated with $\mathbb{R} \times \Theta'$. To see this, note that $\mathbb{R} \times \Theta'$ is just $\mathbb{R} \times \Theta$ with a scale factor added to Θ. Let f be the obvious one-one mapping from $\mathbb{R} \times \Theta$ onto $\mathbb{R} \times \Theta'$. Now, if P is the probability measure on $\mathbb{R} \times \Theta$ corresponding to F, and P' on $\mathbb{R} \times \Theta'$ corresponds to F', then $P\{A\} = P'\{f(A)\}$ for all measurable $A \subset \mathbb{R} \times \Theta$. From this, and the fact that the same loss function is used on Θ and Θ', (3.27) and (3.28) follow easily. From (3.27) and (3.28)

$$E(L_n - \hat{L}_n)^2 = \frac{1}{m^2} E(L'_n - \hat{L}'_n)^2 \tag{3.29}$$

We can conclude that any distribution-free bound on $E(L_n - \hat{L}_n)^2$ must depend on M^2. Similarly, a bound on $E|L_n - \hat{L}_n|$ must depend on M, if it is to be distribution-free.

Finally, we note that if $F(x, \theta)$ possesses atoms, there may be positive probability that more than one observation is at the same distance from X as its k^{th}-nearest neighbor. In order to preserve the local nature of the rule, it will be necessary to break the tie in distance in order to use only k neighbors in the estimate of θ. This may be done by generating a sequence of random variables $Z_1, \ldots, Z_n$ which are independent and identically distributed uniformly on the interval (0,1). Ties in distance can then be broken by choosing from among those tied, the observation (or observations, if necessary) which has the smallest Z_i associated with it.

In cases where $F(x, \theta)$ is known to be atomic, a reasonable estimation rule would attempt to use only observations which lie on the same atom as X in forming its estimate of θ. This is because the conditional distribution of the parameters associated with one atom may be totally distinct from the conditional distribution of the parameters associated with any other atom. It is still necessary to break ties in distance since more than k observations from the data may lie on the same atom as X.

The bounds presented in Theorems 5 and 6 are limited in the sense that fairly large values of n are necessary to produce useful results, especially for the holdout estimate. The values of ℓ_n which minimize the bound in Theorem 6 is a complicated function of n, ϵ, m and k, however the optimum value is generally near $\sqrt{n}$. This implies that the best rate of decrease for the bound on the holdout estimate is on the order of $1/\sqrt{n}$. This is a factor of $\sqrt{n}$ slower than the bound on the deleted estimate. On the other hand, the amount of computation required to obtain the holdout estimate with $\ell_n = \sqrt{n}$ is a factor of $1/\sqrt{n}$ less than the computation required to obtain the deleted estimate. This suggests that the holdout estimate would only be practical in cases where the size of the data set is so large that the holdout bounds are useful and the deleted estimate is too costly.

IV. COMPUTER SIMULATION RESULTS

IV.1 Remarks Concerning the Experiments

The bounds established by Theorems 5 and 6 are, by their very nature, extremely general. While this generality is a virtue for reasons already mentioned, it also can be expected to result in rather loose bounds for a great many choices of the underlying distribution $F(x, \theta)$. In order to gain some feel for the performance of the deleted and holdout estimates in specific examples, some computer simulation experiments were performed.

The experiments are in the form of Monte Carlo studies, performed in the following fashion. For a particular distribution F, n independent pseudo-random vectors $((X_1, \theta_1), \ldots, (X_n, \theta_n))$ were generated having that distribution. Based on the generated data, L_n and its deleted and holdout estimates were computed and compared for k = 1,3,5,7 and 9 nearest neighbor rules. This process was repeated 1200 times for various values of n. For certain values of ε, the relative frequency of the event $\{|L_n - \hat{L}_n| > \varepsilon\}$ was returned as an estimate of $P\{|L_n - \hat{L}_n| > \varepsilon\}$. The number of runs, 1200, was selected to guarantee that the estimates produced would be within .03 of the correct value 95% of the time.

The experiments were performed for three specific distributions. The first is actually a discrimination problem where θ takes only

the values 0 or 1 with equal probability. The conditional densities $f(x/\theta=1)$ and $f(x/\theta=0)$ are the triangular densities for which Cover and Hart calculated EL_n and its limit. This example is used predominantly because Cover and Hart's results are quite well known. The second example is also a discrimination problem. In this case, θ again takes values 0 and 1 with equal probability, but $f(x/\theta)$ is now gaussian with mean θ and variance 1. This example is included mainly for comparison purposes with the third example, which is an estimation problem with θ uniformly distributed on the interval $[0,1]$, while $f(x/\theta)$ is again gaussian with mean θ and variance 1.

IV.2 Example 1: Discrimination with Triangular Densities

The triangular densities of Cover and Hart are shown in Fig. 1. For these densities and the discrimination version of the single-nearest neighbor rule, with the 0-1 loss function defined by

$$L(\theta,\hat{\theta}) = \begin{cases} 0, & \text{if } \theta = \hat{\theta} \\ 1, & \text{otherwise,} \end{cases}$$

they calculated

$$EL_n = \tfrac{1}{3} + \frac{1}{(n+1)(n+2)}$$

and $$R^* = \tfrac{1}{4} .$$

In Table 1 we have shown the average value of L_n obtained for $k = 1,3,5,$

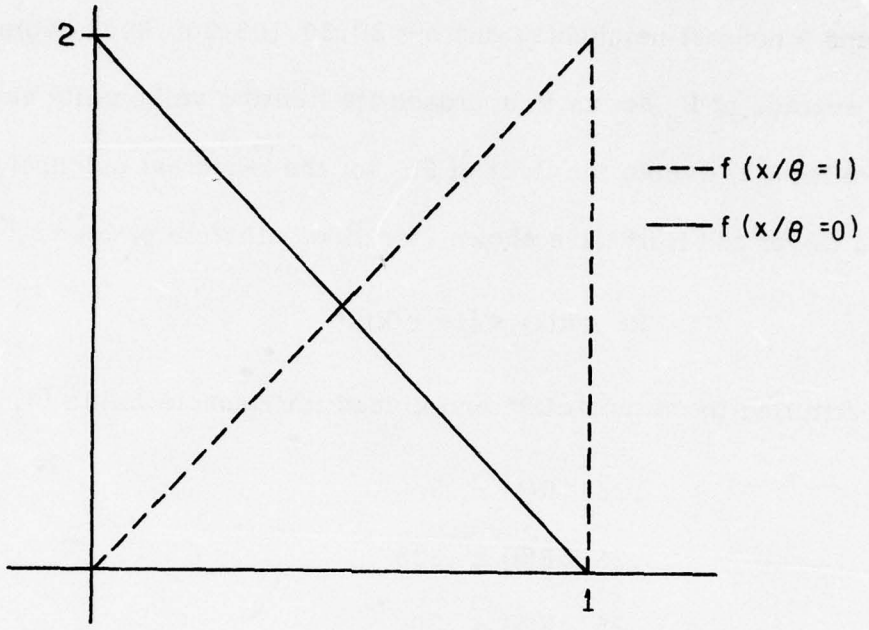


Figure 1. Conditional Densities for Example 1

Table 1. Experimentally Obtained Average Values of L_n for Example 1

	n=20	50	100	200	400
k=1	.3364	.3337	.3337	.3331	.3335
k=3	.3055	.3010	.3001	.3002	.3000
k=5	.2936	.2867	.2858	.2853	.2860
k=7	.2911	.2783	.2777	.2773	.2778
k=9	.2933	.2736	.2726	.2722	.2728

7, and 9 nearest neighbors, and n = 20,50,100,200,400. Note that the average of L_n seems to approach its limiting value quite early. If we let R(k) denote the limit of EL_n for the k-nearest neighbor rule, then Cover and Hart have shown, for discrimination problems,

$$R^* \leq R(k) \leq (1+ 1/k)R^* \ . \tag{4.1}$$

Substituting the values of R* and k used for Example 1 into (4.1) we have

$$.25 \leq R(1) \leq .5 \tag{4.2a}$$

$$.25 \leq R(3) \leq .333 \tag{4.2b}$$

$$.25 \leq R(5) \leq .300 \tag{4.2c}$$

$$.25 \leq R(7) \leq .286 \tag{4.2d}$$

$$.25 \leq R(9) \leq .278 \ . \tag{4.2e}$$

The values in Table 1 satisfy the bounds given by (4.2) quite easily.

In Figs. 2-6 we have shown, for the same values of k, graphs of $P\{|L_n - \hat{L}_n| > \epsilon\}$ versus n for ϵ = .025, .05, and .1. These graphs are for the deleted estimate. The curves for the various values of k exhibit somewhat similar behavior, decreasing quite slowly for the smaller values of ϵ, and much more quickly for larger values.

The theoretical results of Rogers and Wagner for discrimination, and the results shown here in Chapter III for estimation, present bounds which deteriorate as k increases. The deterioration of the performance of the deleted estimate as k increases is not in evidence in Figs. 2-6,

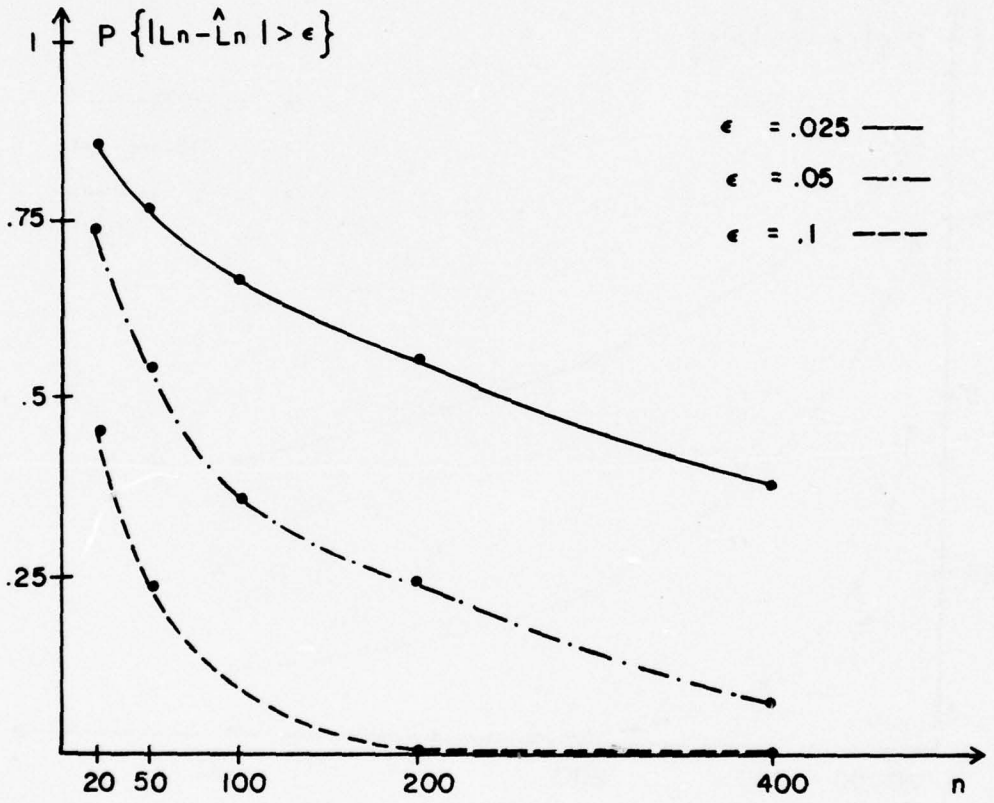


Figure 2. Performance of the Deleted Estimate for Example 1, k=1

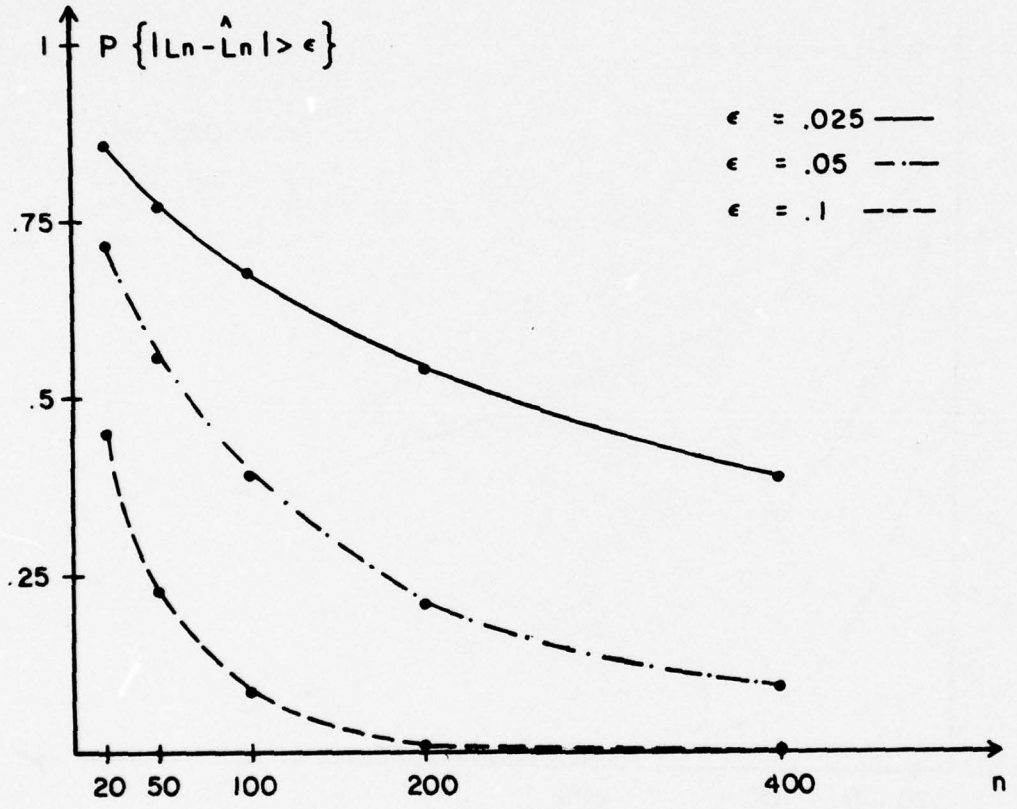


Figure 3. Performance of the Deleted Estimate for Example 1, k=3

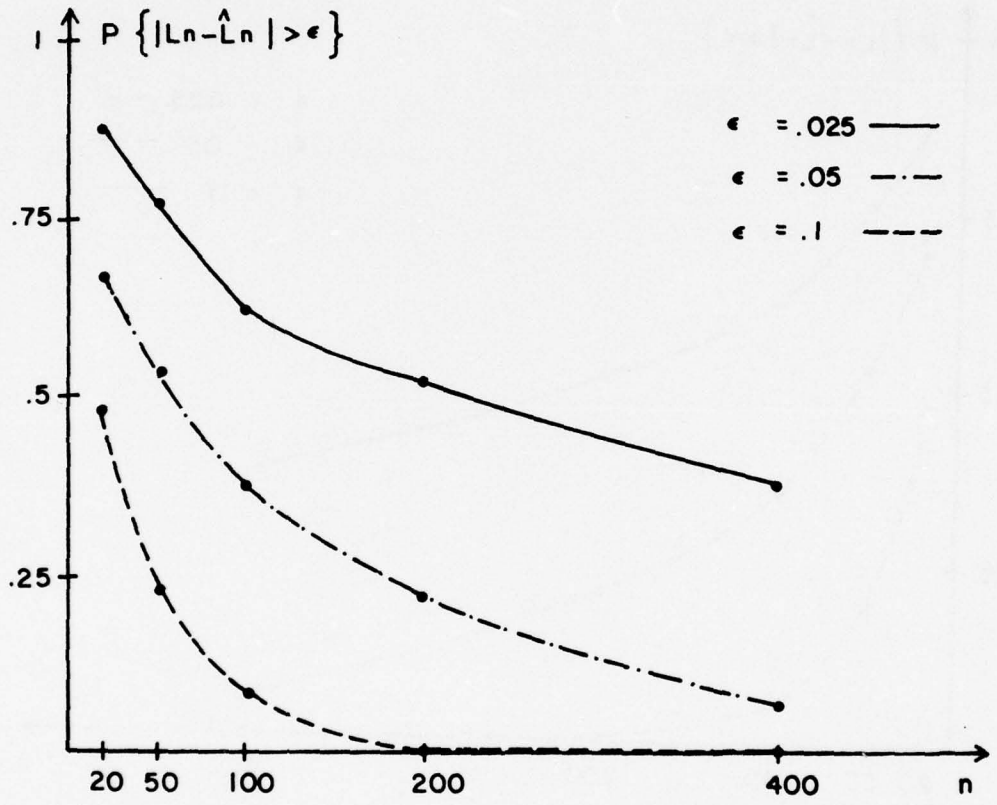


Figure 4. Performance of the Deleted Estimate for Example 1, k=5

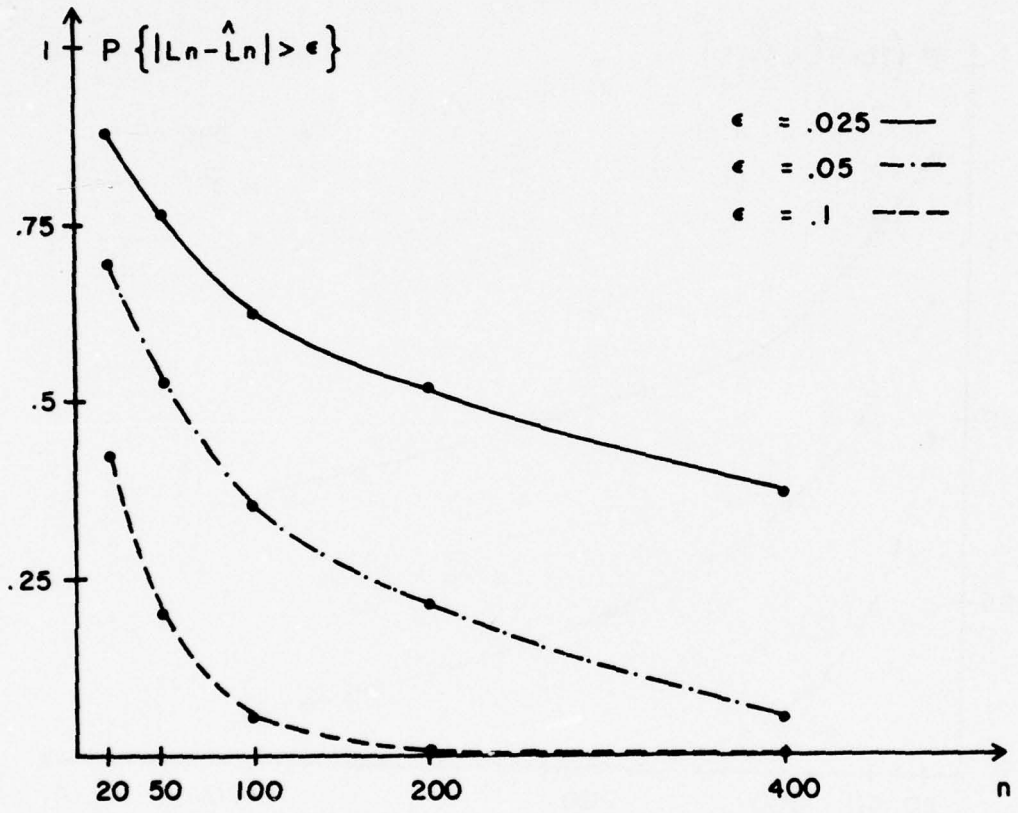


Figure 5. Performance of the Deleted Estimate for Example 1, k=7

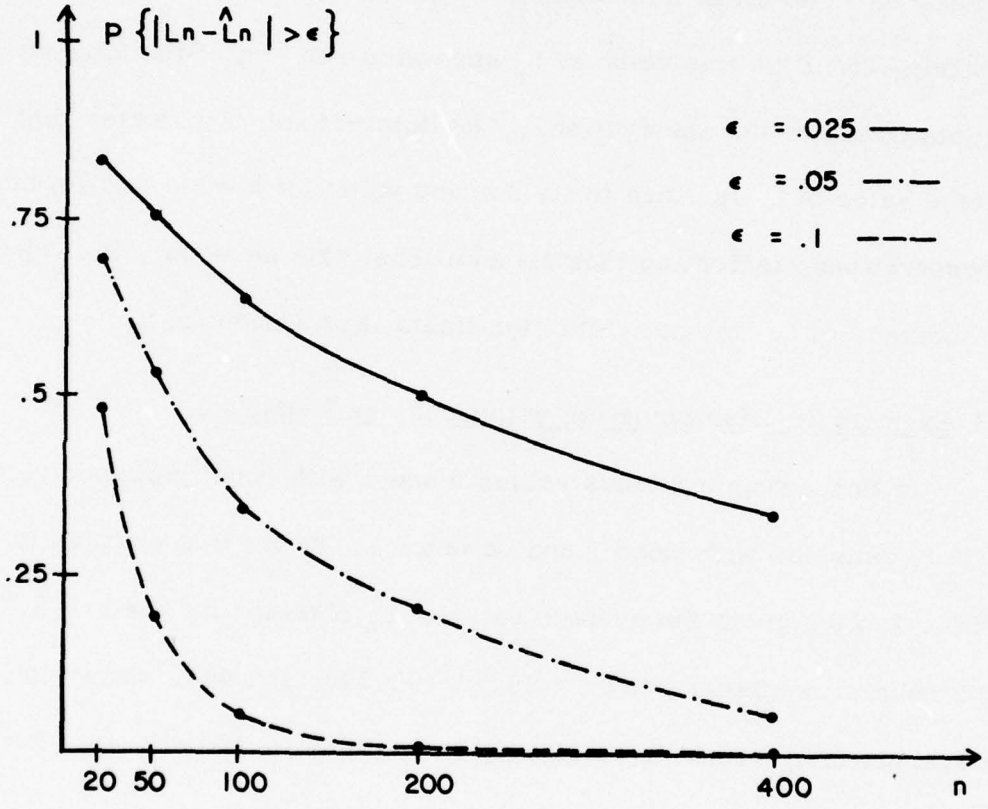


Figure 6. Performance of the Deleted Estimate for Example 1, k=9

even for the case where k = 9 and n is twenty. The theoretical bounds are quite pessimistic for the example, as expected, since they are equal to one for all the values of k,n, and ϵ used. The deleted estimate performs well enough in this example to guarantee that the estimate will be within .05 of the true value of L_n approximately 93% of the time when the data contains 400 observations. The data in Table I indicates that the average value of L_n is close to its limiting value for a much smaller number of observations, indicating that the additional data necessary to insure the accuracy of $\hat{L}_n$ may not yield significant improvement in L_n.

IV.3 Example 2: Discrimination with Gaussian Densities

In this example θ takes values 0 and 1 with equal probability, and $f(x/\theta)$ is gaussian with mean θ and variance 1. R^* for this example is .3096. Table 2 gives the average value of L_n obtained for k = 1,3,5,7, and 9 nearest neighbors, and n = 20,50,100,200, and 400. Once again, the average of L_n seems to be already quite close to its limit when n = 20. For this example, the bounds on R(k) are given by (4.3):

$$.3096 \leq R(1) \leq .5 \qquad (4.3a)$$

$$.3096 \leq R(3) \leq .402 \qquad (4.3b)$$

$$.3096 \leq R(5) \leq .361 \qquad (4.3c)$$

$$.3096 \leq R(7) \leq .344 \qquad (4.3d)$$

$$.3096 \leq R(9) \leq .334\ . \qquad (4.3e)$$

The data in Table 2 approaches these bounds closely.

Table 2. Experimentally Obtained Average Values of L_n for Example 2

	n=20	50	100	200	400
k=1	.4036	.4001	.3987	.3980	.3981
k=3	.3820	.3720	.3692	.3683	.3680
k=5	.3722	.3586	.3541	.3536	.3531
k=7	.3683	.3505	.3456	.3444	.3445
k=9	.3730	.3451	.3403	.3386	.3387

In Figs. 7-11 we have shown, for the same values of k, graphs of $P\{|L_n - \hat{L}_n| > \epsilon\}$ versus n for $\epsilon = .05, .1$, and $.2$, where $\hat{L}_n$ is the deleted estimate. Figs. 13-17 show the same data for the holdout estimate. The behavior of the deleted estimate in this example is quite similar to Example 1, although the values for ϵ have been adjusted upward slightly indicating the slightly more complex nature of this problem. In this example, increasing k shows some slight deterioration in the performance of the deleted estimate when n is small, but for larger values of n, increasing k actually seems to improve performance, if anything. A comparison of the curves for the holdout estimate with those for the deleted estimate reveals that much larger values of n are needed to achieve the same performance levels, as expected.

Figs. 12 and 18 show the average squared error of the deleted and holdout estimates, respectively, for the same values of n as before. The performance is so uniform in k that only one curve is shown in

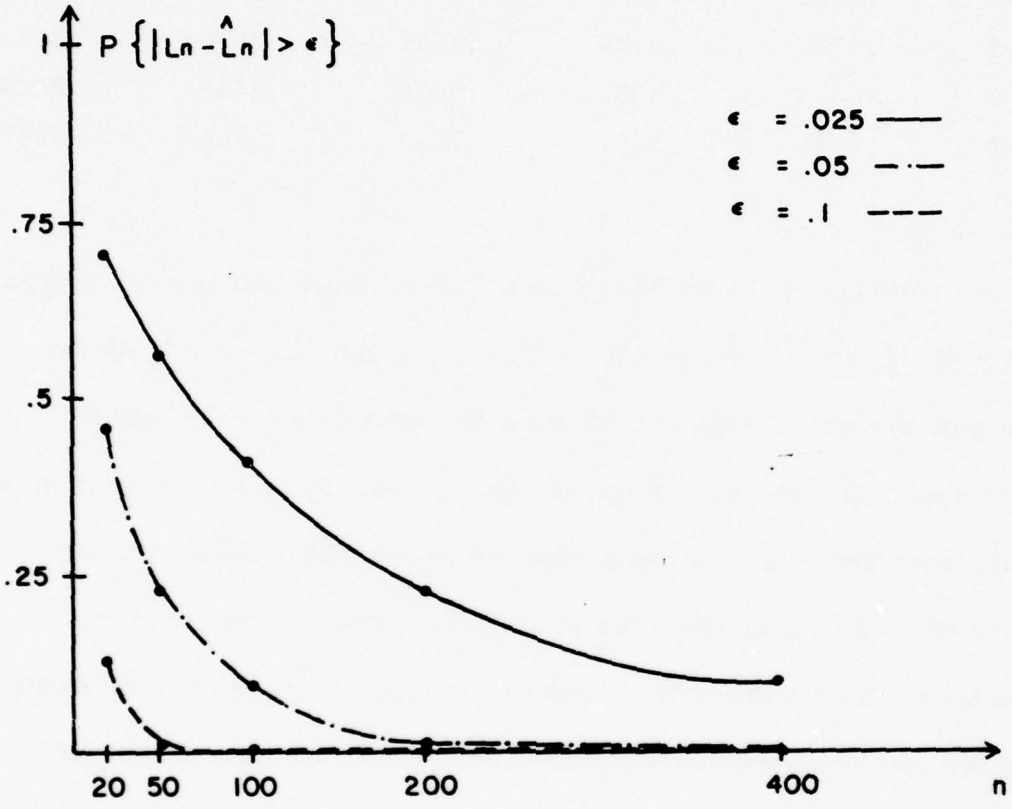


Figure 7. Performance of the Deleted Estimate for Example 2, k=1

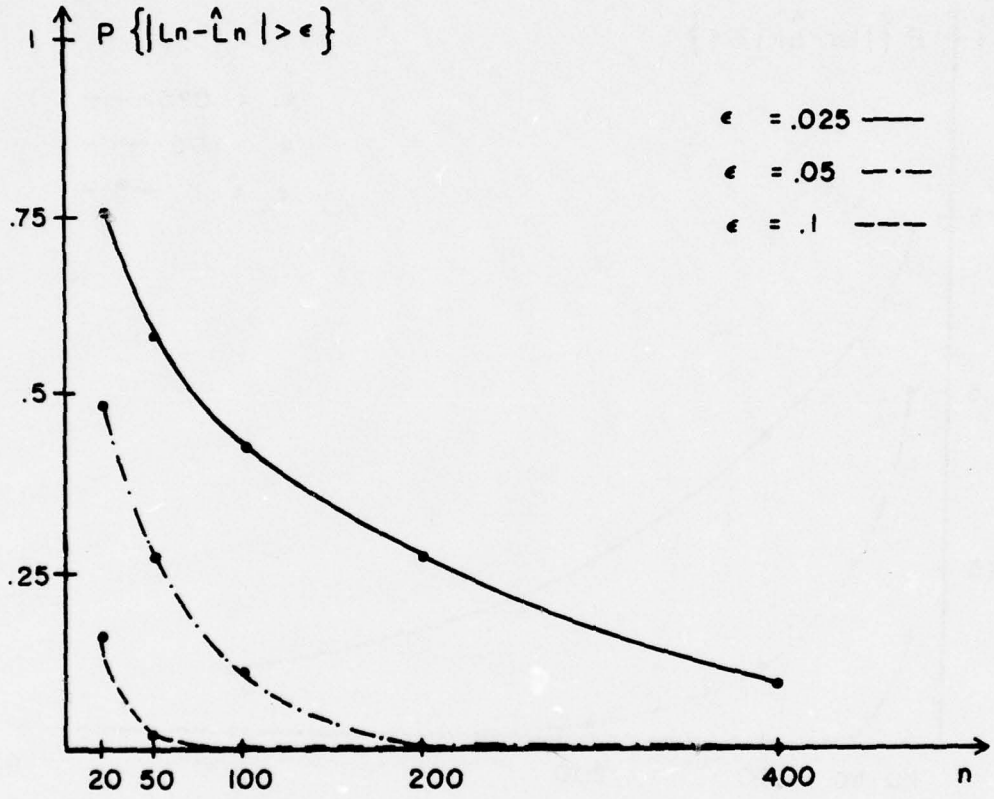


Figure 8. Performance of the Deleted Estimate for Example 2, k=3

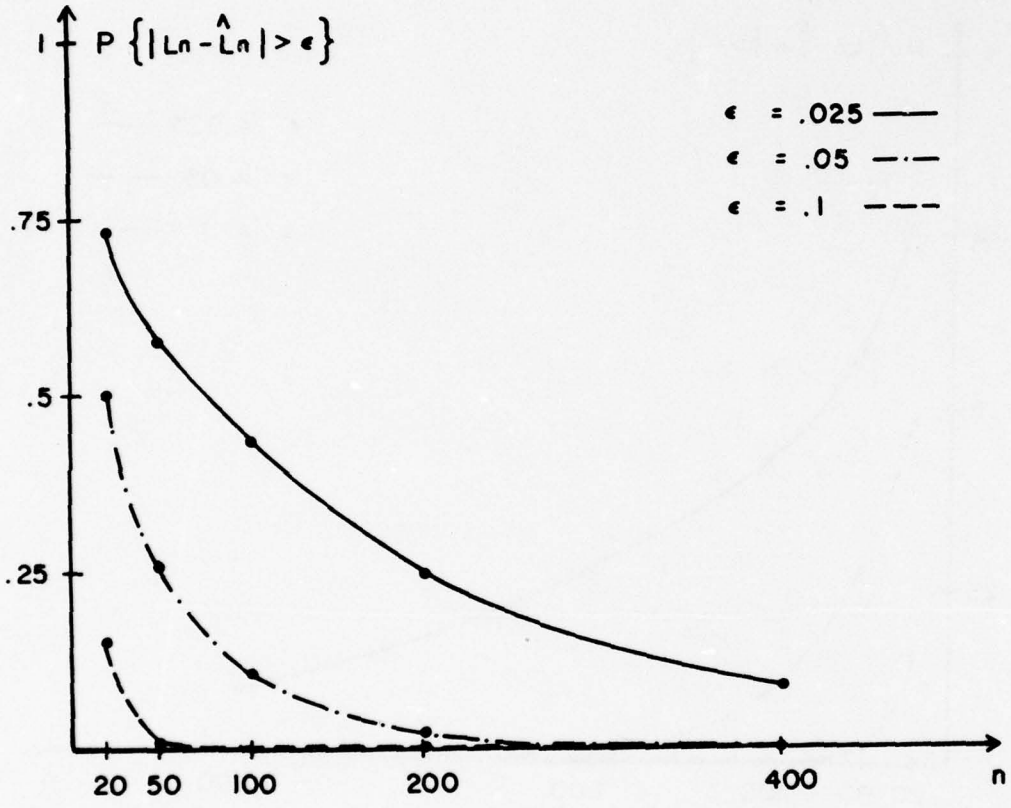


Figure 9. Performance of the Deleted Estimate for Example 2, k=5

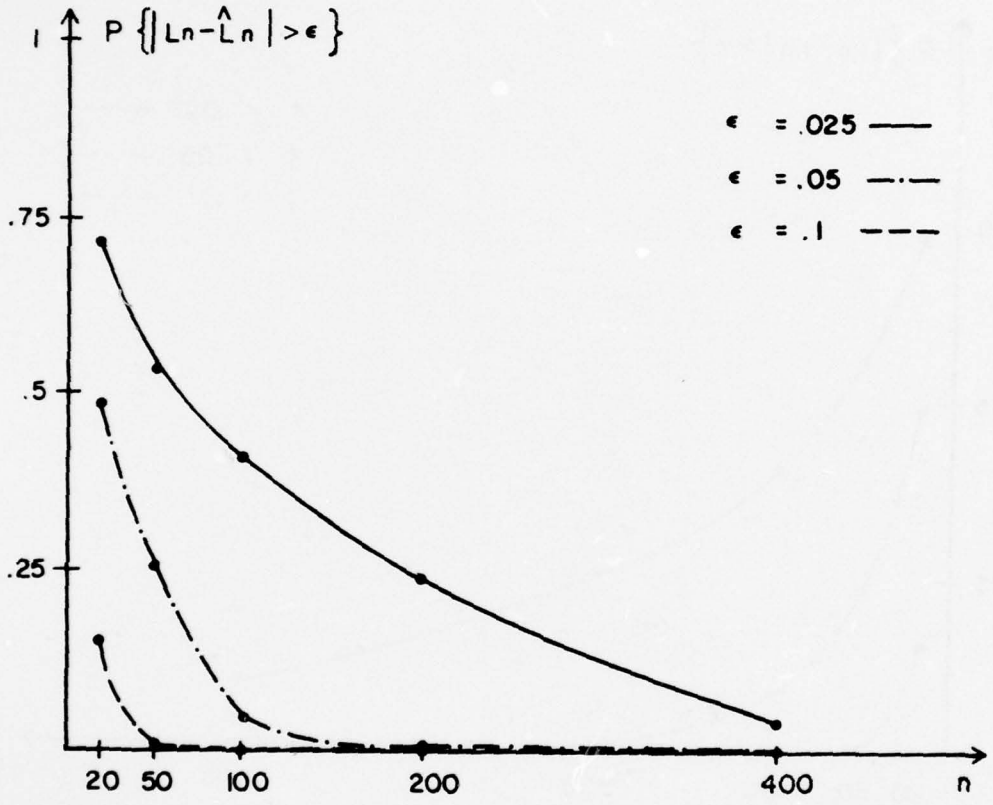


Figure 10. Performance of the Deleted Estimate for Example 2, k=7

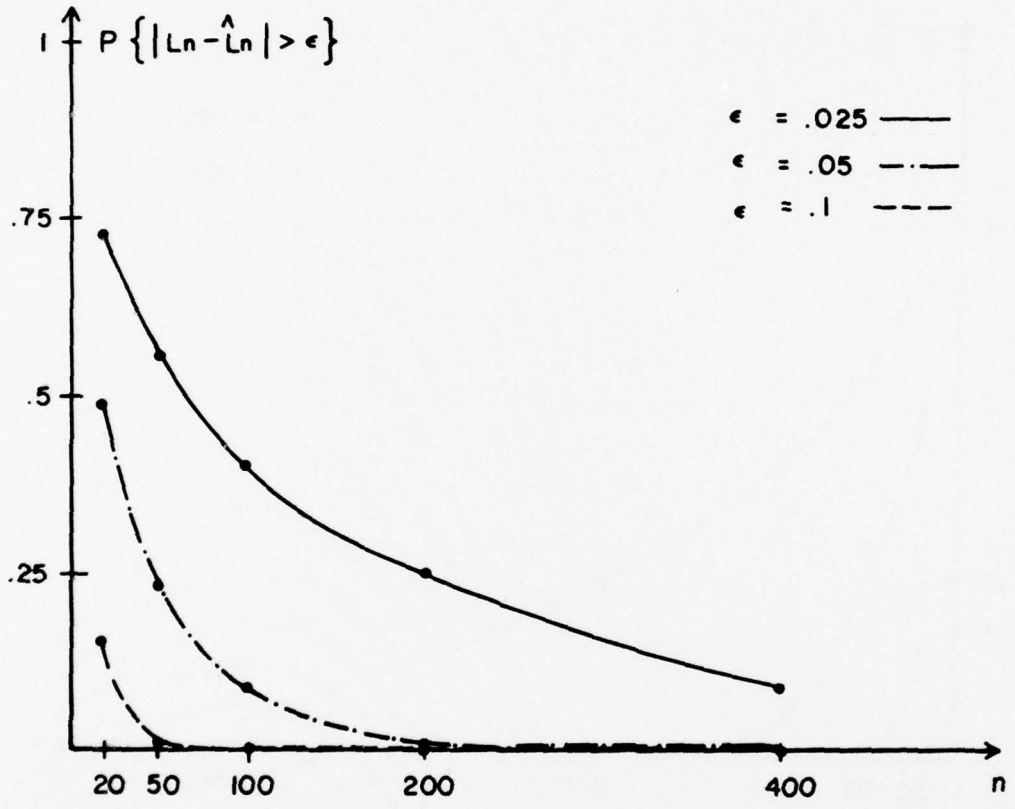


Figure 11. Performance of the Deleted Estimate for Example 2, k=9

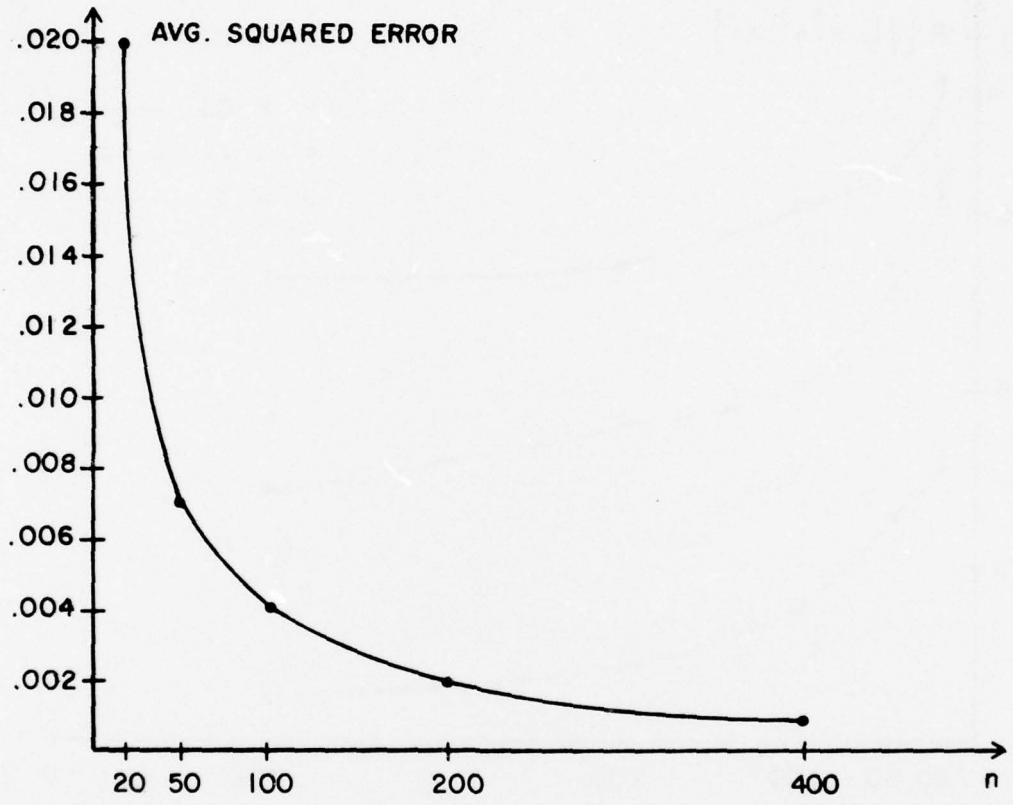


Figure 12. Average Squared Error of the Deleted Estimate for Example 2

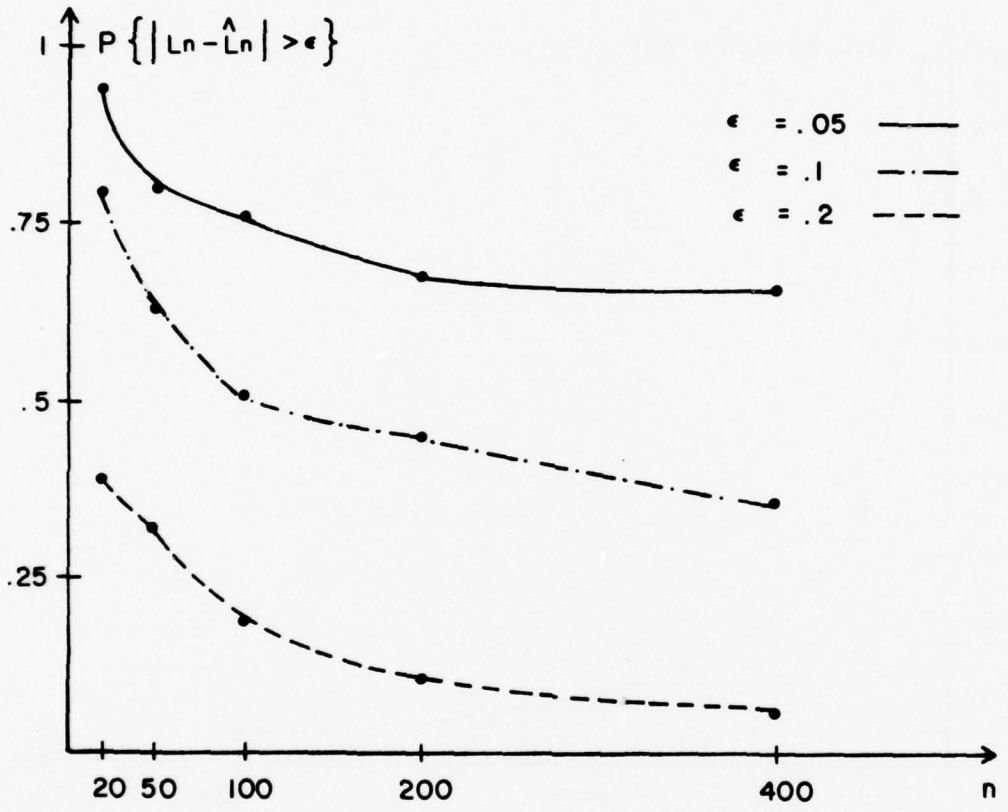


Figure 13. Performance of the Holdout Estimate for Example 2, k=1

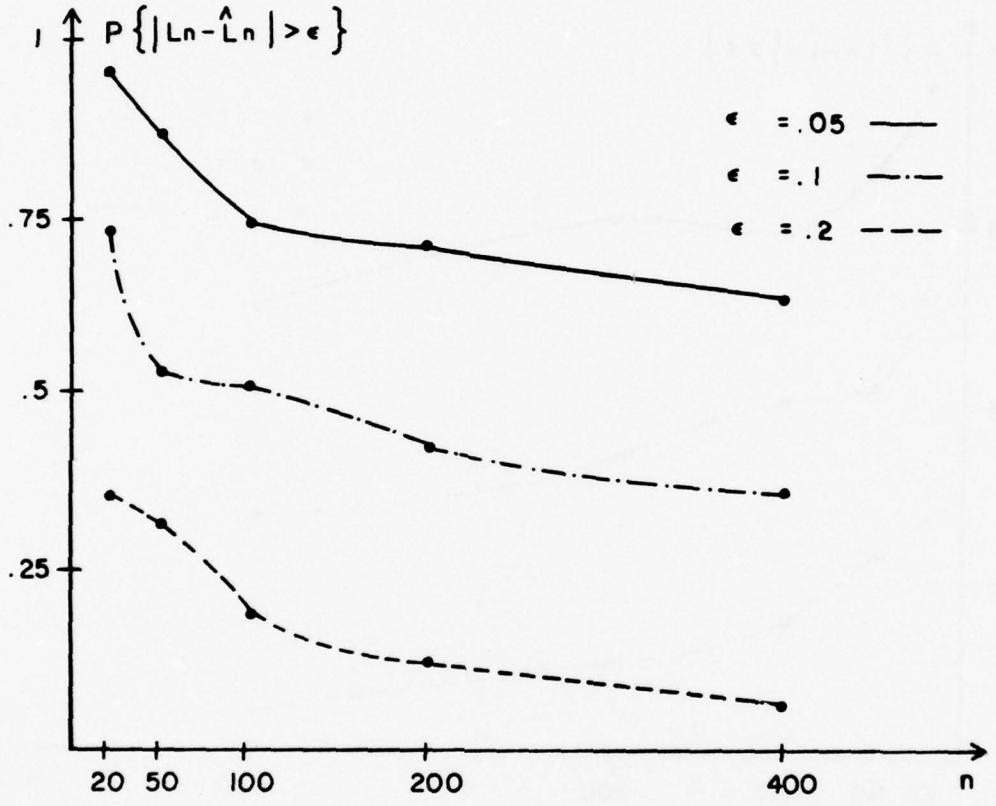


Figure 14. Performance of the Holdout Estimate for Example 2, k=3

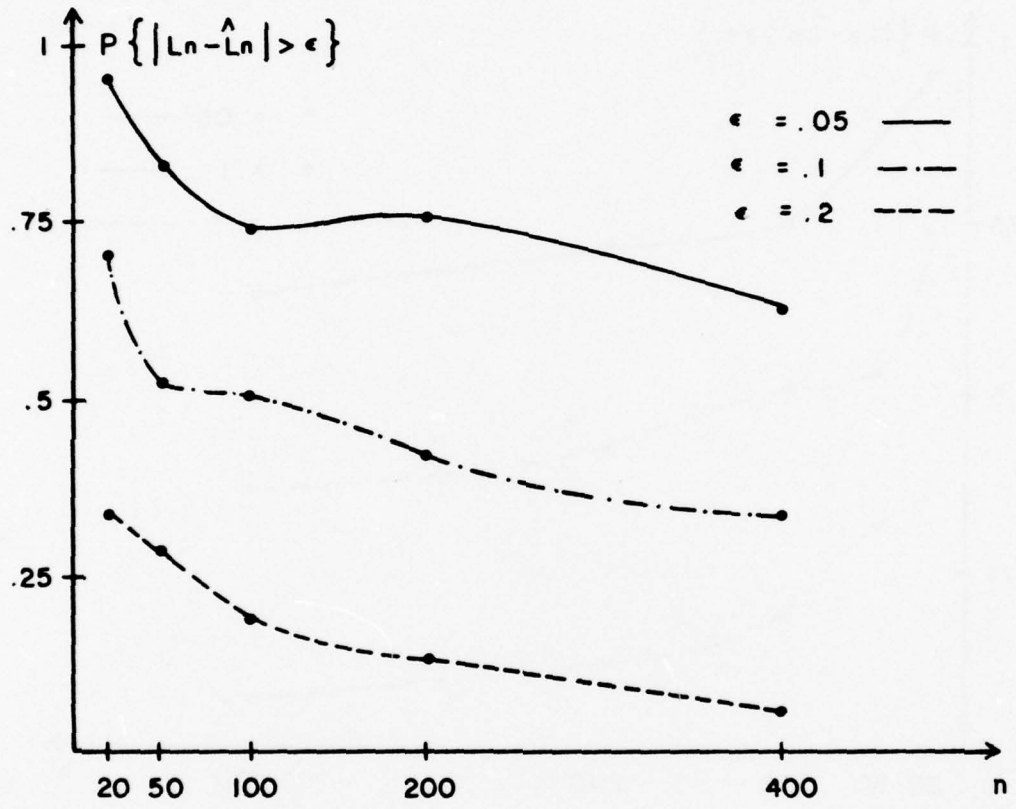


Figure 15. Performance of the Holdout Estimate for Example 2, k=5

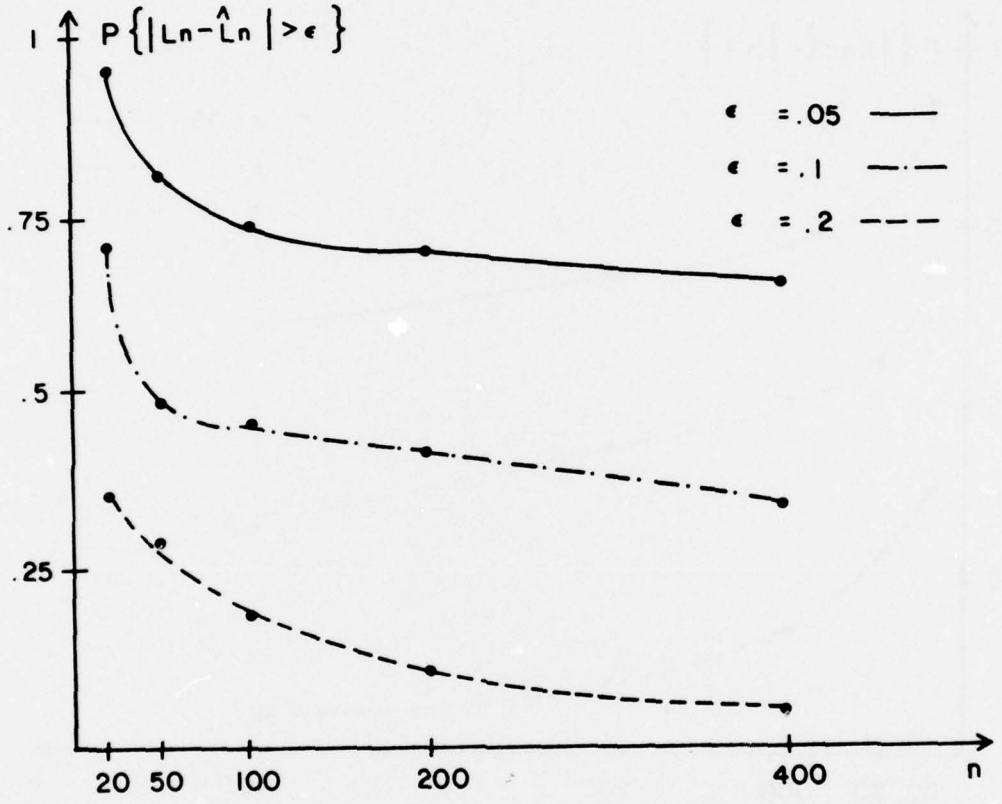


Figure 16. Performance of the Holdout Estimate for Example 2, k=7

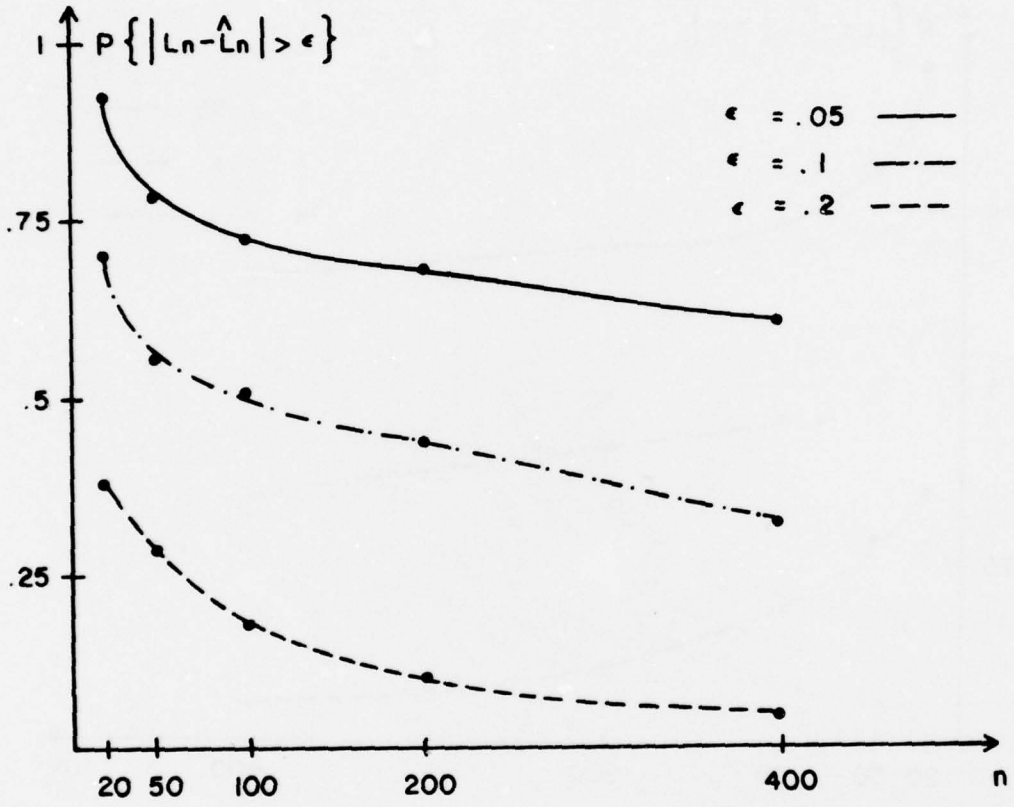


Figure 17. Performance of the Holdout Estimate for Example 2, k=9

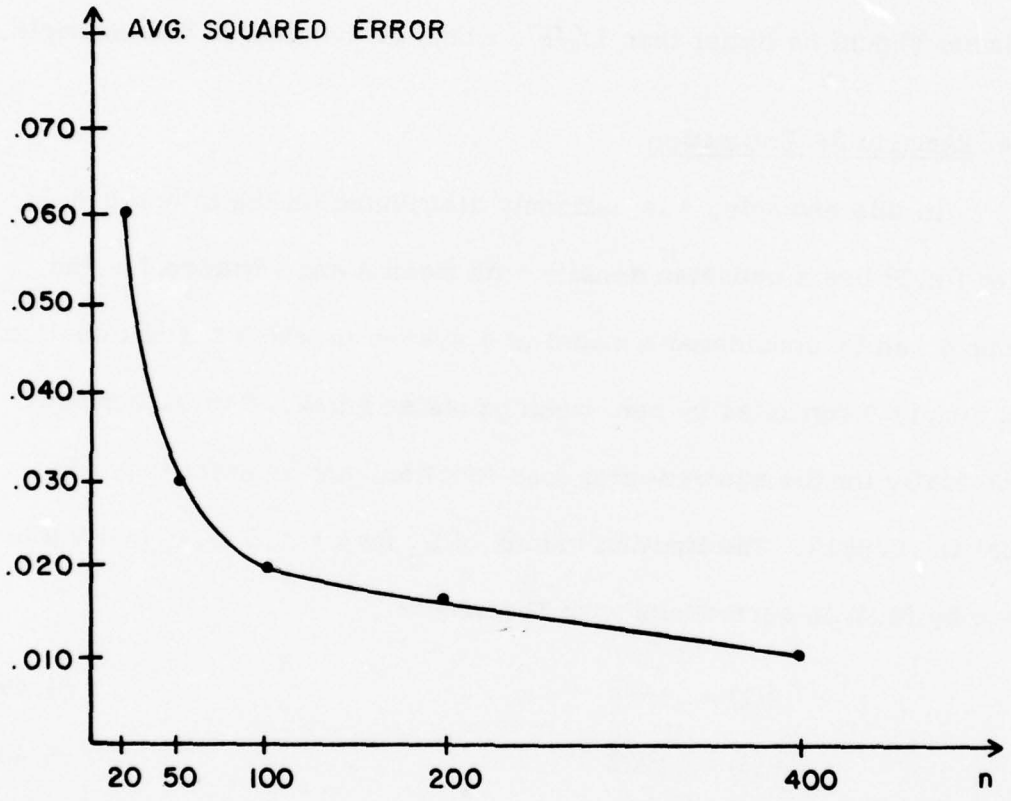


Figure 18. Average Squared Error of the Holdout Estimate for Example 2

Figs. 12 and 18. This curve represents all k = 1,3,5,7, and 9 nearest neighbor rules. The rate of decrease for the deleted estimate is slightly faster than 1/n as predicted by Rogers and Wagner. Theorem 6 in Chapter III indicates that the rate of decrease for the squared error of the holdout estimate should be better than $1/\sqrt{n}$, which is the case in this example.

IV.4 Example 3: Estimation

In this example, θ is uniformly distributed on the interval [0,1], while $f(x/\theta)$ has a gaussian density with mean θ and variance 1. The example can be considered a model of a system in which the observation X is simply θ corrupted by zero mean gaussian noise. R* was computed numerically for the squared-error loss function, and is approximately equal to .076913. The limiting values of L_n for k = 1,3,5,7, and 9 are given by (4.4) in accordance with Theorem 3:

$$R(1) = .1538 \tag{4.4a}$$

$$R(3) = .1025 \tag{4.4b}$$

$$R(5) = .0923 \tag{4.4c}$$

$$R(7) = .0879 \tag{4.4d}$$

$$R(9) = .0855\ . \tag{4.4e}$$

Table 3 shows the average values of L_n for k = 1,3,5,7 and 9, and n = 20,50,100,200,400. It is again apparent that EL_n approaches its limit fairly closely even for n = 20. The values for n = 400 are comparable

to the limits predicted by (4.4). Note that in this example the average risk for k = 1 is almost cut in half by the use of 9 neighbors for all the values of n shown.

Table 3. Experimentally Obtained Average Values of L_n for Example 3

	n=20	n=50	n=100	n=200	n=400
k=1	.1550	.1541	.1539	.1536	.1535
k=3	.1031	.1025	.1023	.1024	.1024
k=5	.0932	.0923	.0921	.0921	.0922
k=7	.0892	.0880	.0877	.0877	.0878
k=9	.0873	.0857	.0853	.0853	.0853

Figs. 19-23 show $P\{|L_n - \hat{L}_n| > \epsilon\}$ for $\epsilon = .01, .025$, and $.05$, where $\hat{L}_n$ is the deleted estimate. Figs. 25-29 show the same data for the holdout estimate. The curves are similar to those of the preceding examples, and the same comments apply, with the following exception. The performance of the estimates improves markedly as k increases for all values of n and ϵ, for both the deleted and holdout estimates. This is in contrast to the deterioration in performance predicted by Theorems 5 and 6 for increasing k. A partial explanantion for this appears to be in the uniform distribution chosen for θ. For example, if we had chosen

$$f(\theta) = \frac{1}{2}\left(\frac{1}{\sigma\sqrt{2\pi}} e^{-\theta^2/2\sigma^2} + \frac{1}{\sigma\sqrt{2\pi}} e^{-(\theta-1)^2/2\sigma^2}\right),$$

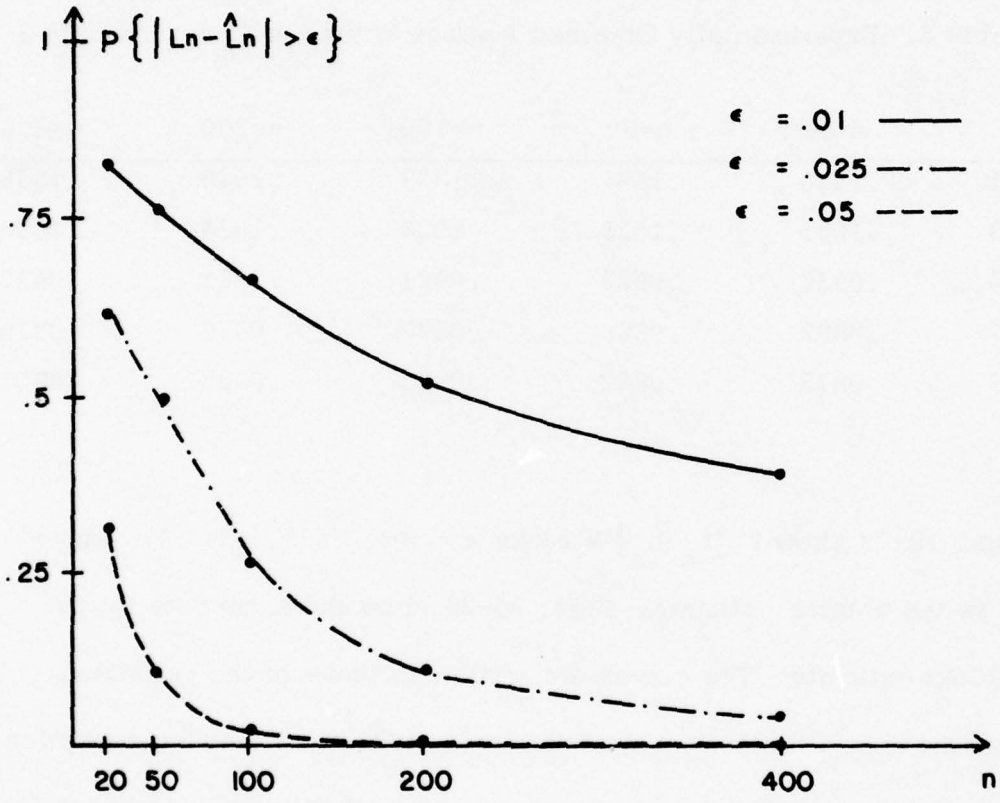


Figure 19. Performance of the Deleted Estimate for Example 3, k=1

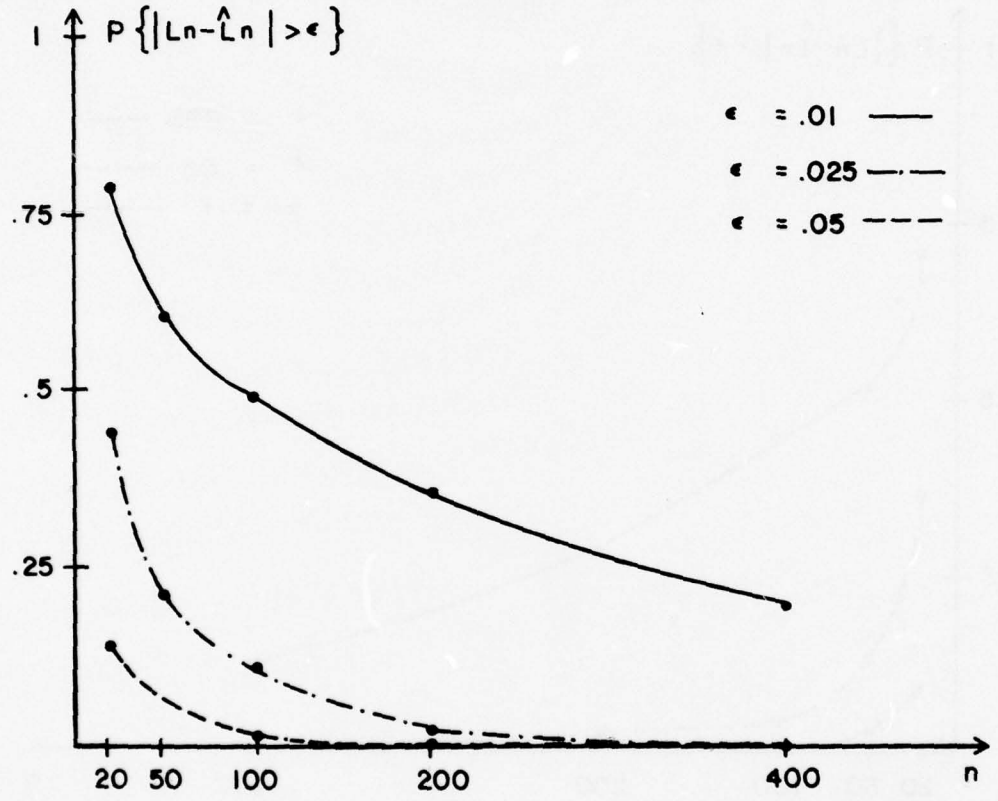


Figure 20. Performance of the Deleted Estimate for Example 3, k=3

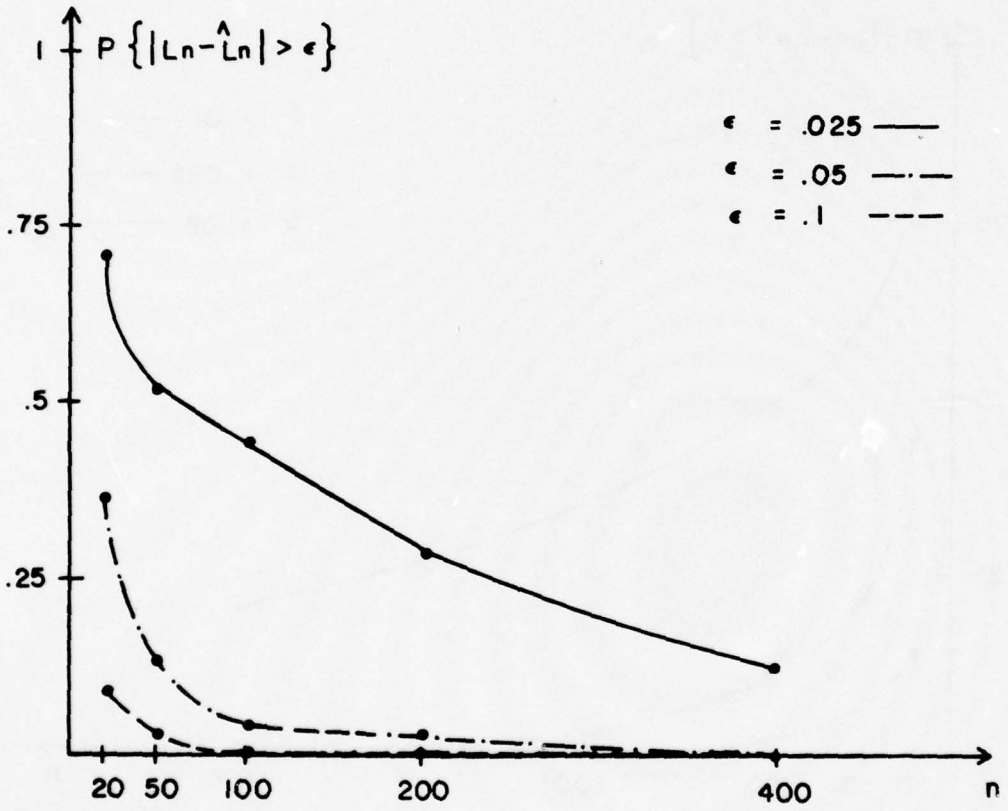
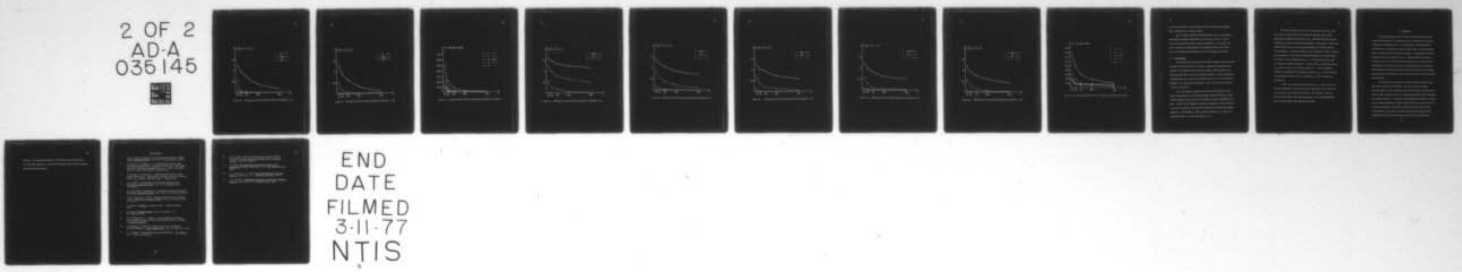


Figure 21. Performance of the Deleted Estimate for Example 3, k=5

AD-A035 145 TEXAS UNIV AT AUSTIN ELECTRONICS RESEARCH CENTER F/G 12/1
NONPARAMETRIC ESTIMATION WITH LOCAL RULES.(U)
OCT 76 C S PENROD, T J WAGNER F44620-76-C-0089
UNCLASSIFIED TR-182 AFOSR-TR-77-0019 NL

2 OF 2
AD-A
035145



END
DATE
FILMED
3-11-77
NTIS

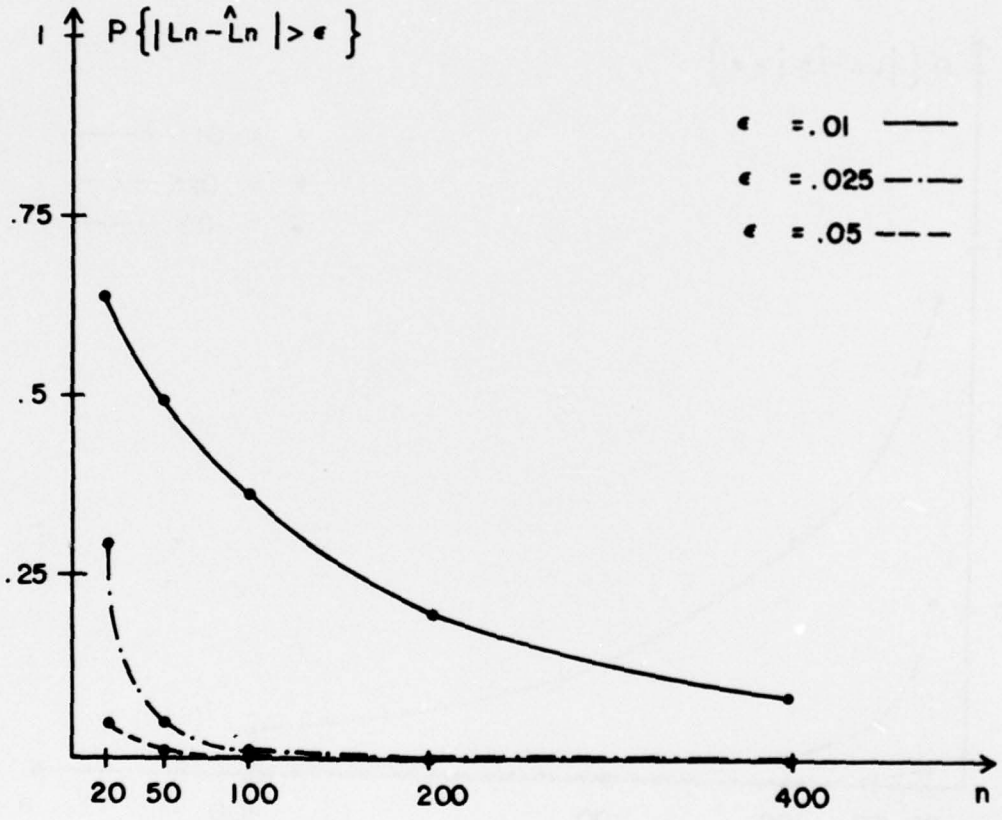


Figure 22. Performance of the Deleted Estimate for Example 3, k=7

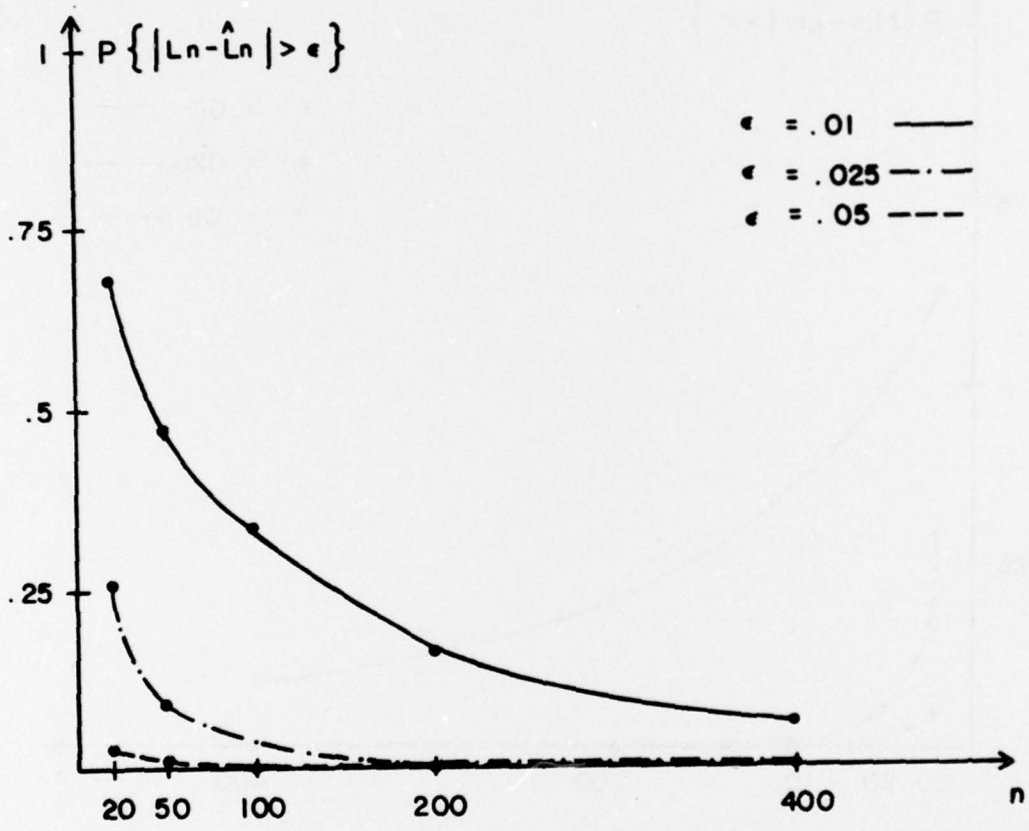


Figure 23. Performance of the Deleted Estimate for Example 3, k=9

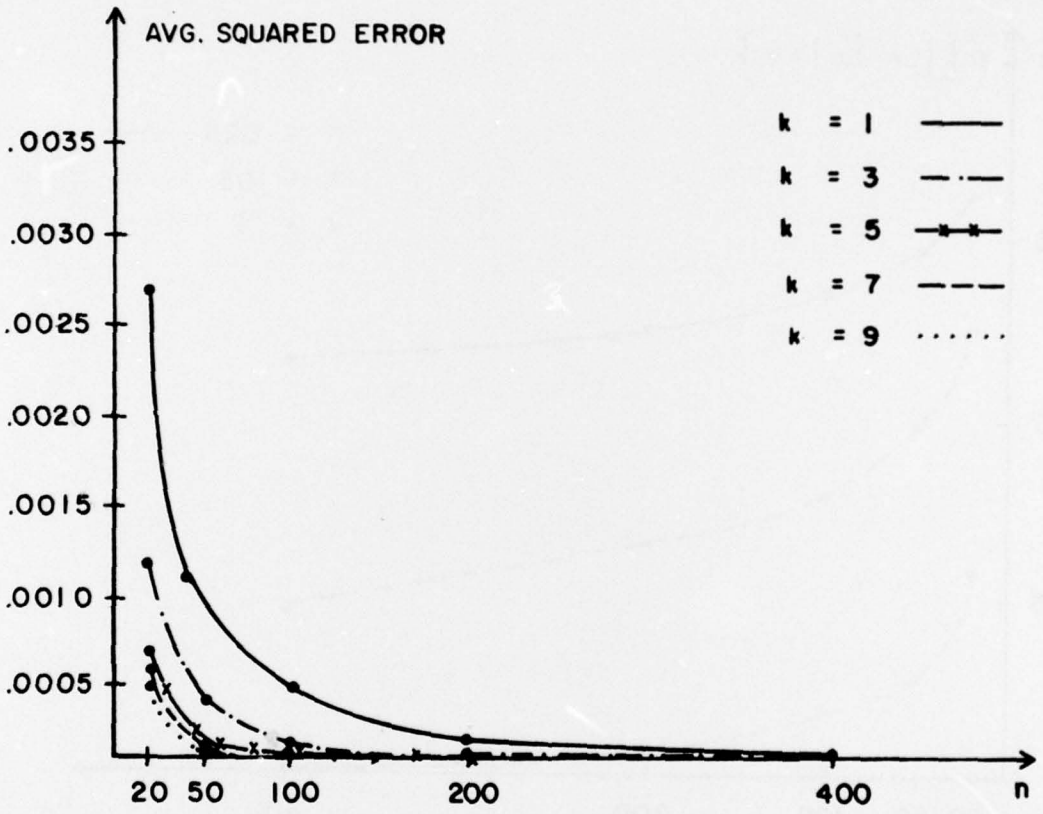


Figure 24. Average Squared Error of the Deleted Estimate for Example 3

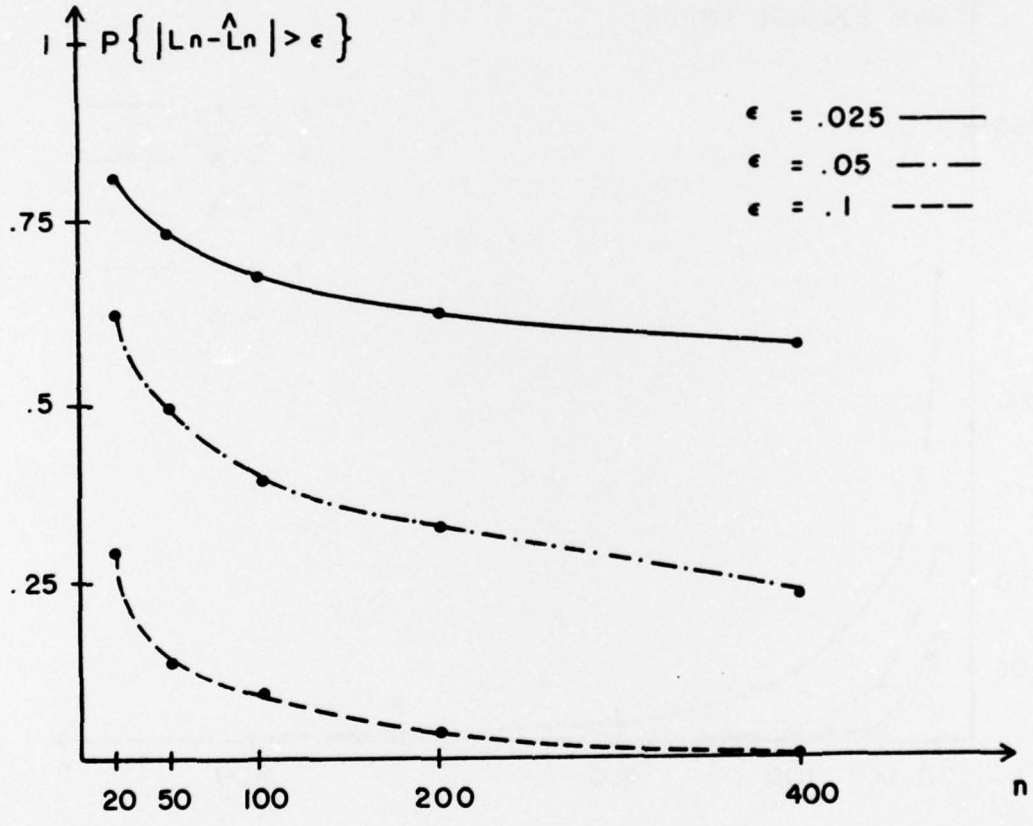


Figure 25. Performance of the Holdout Estimate for Example 3, k=1

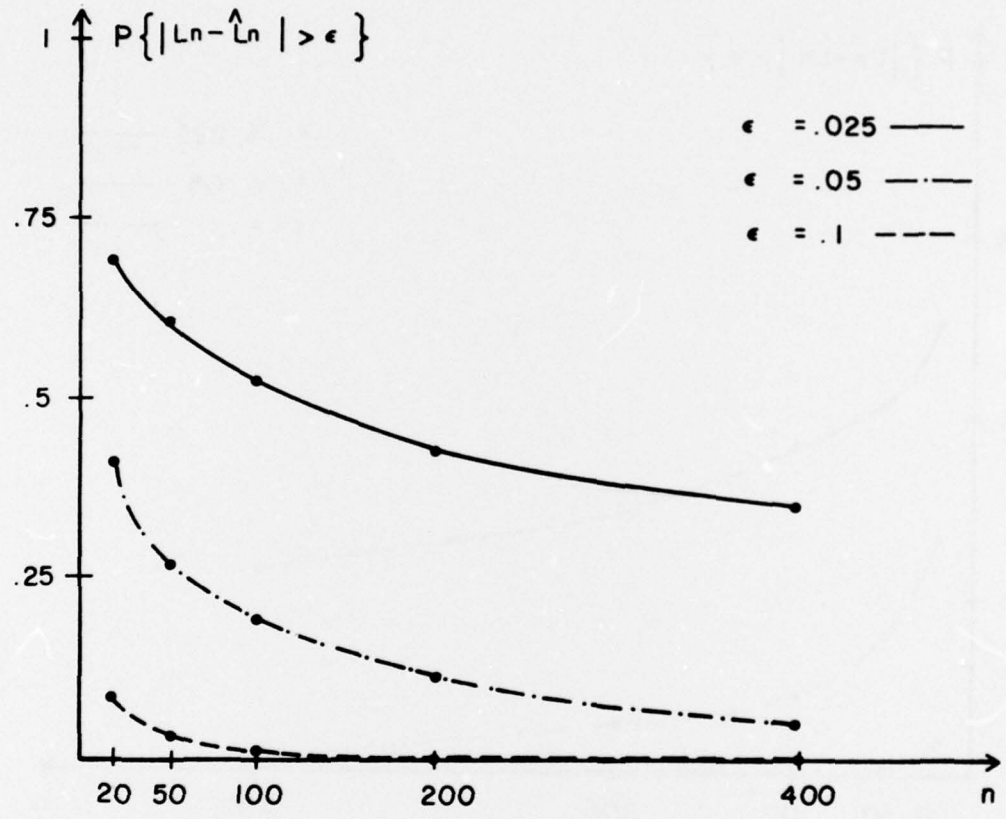


Figure 26. Performance of the Holdout Estimate for Example 3, k=3

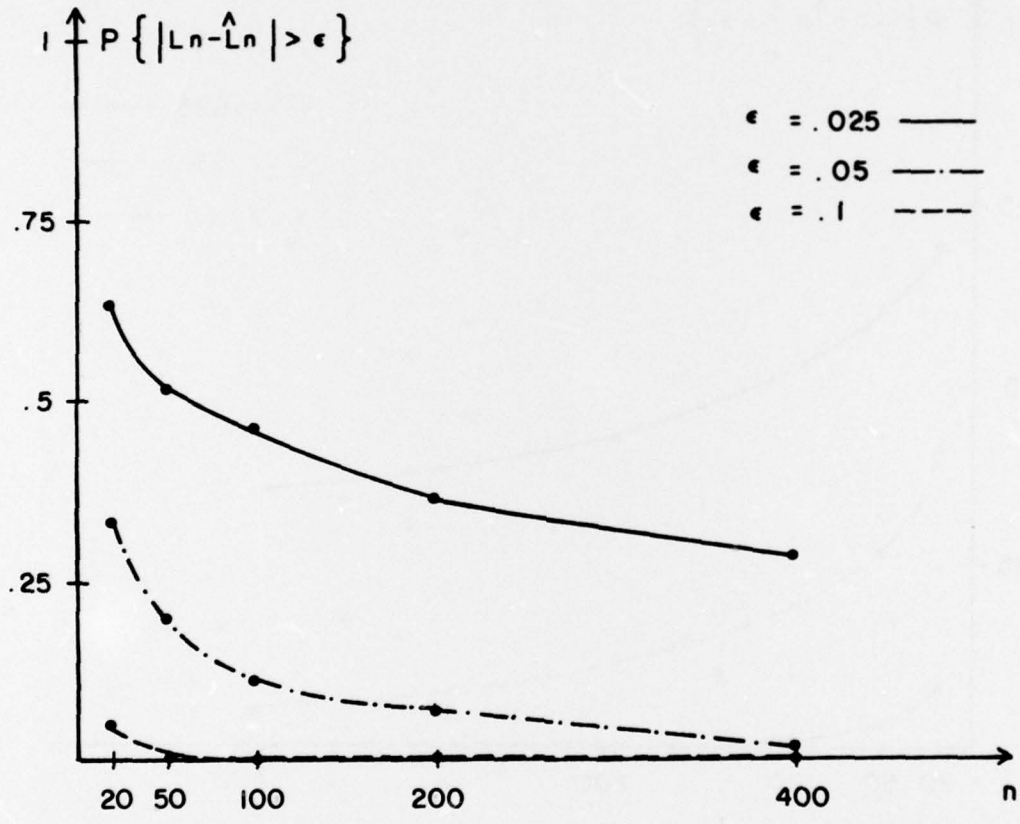


Figure 27. Performance of the Holdout Estimate for Example 3, k=5

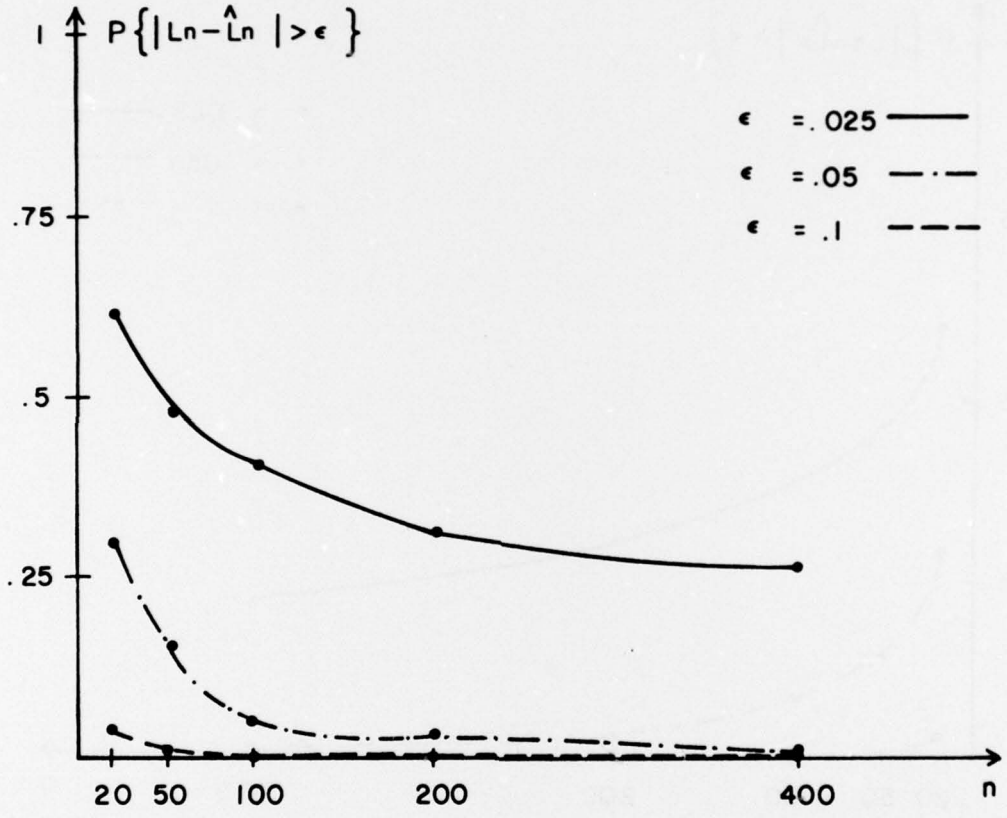


Figure 28. Performance of the Holdout Estimate for Example 3, k=7

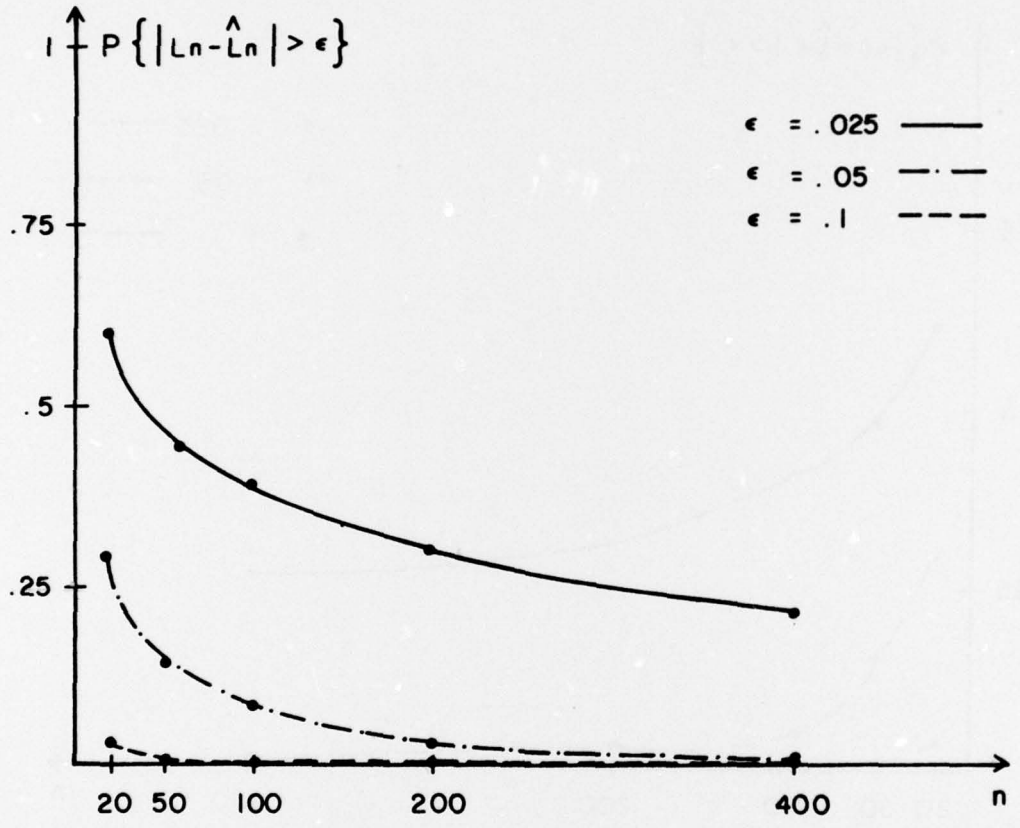


Figure 29. Performance of the Holdout Estimate for Example 3, k=9

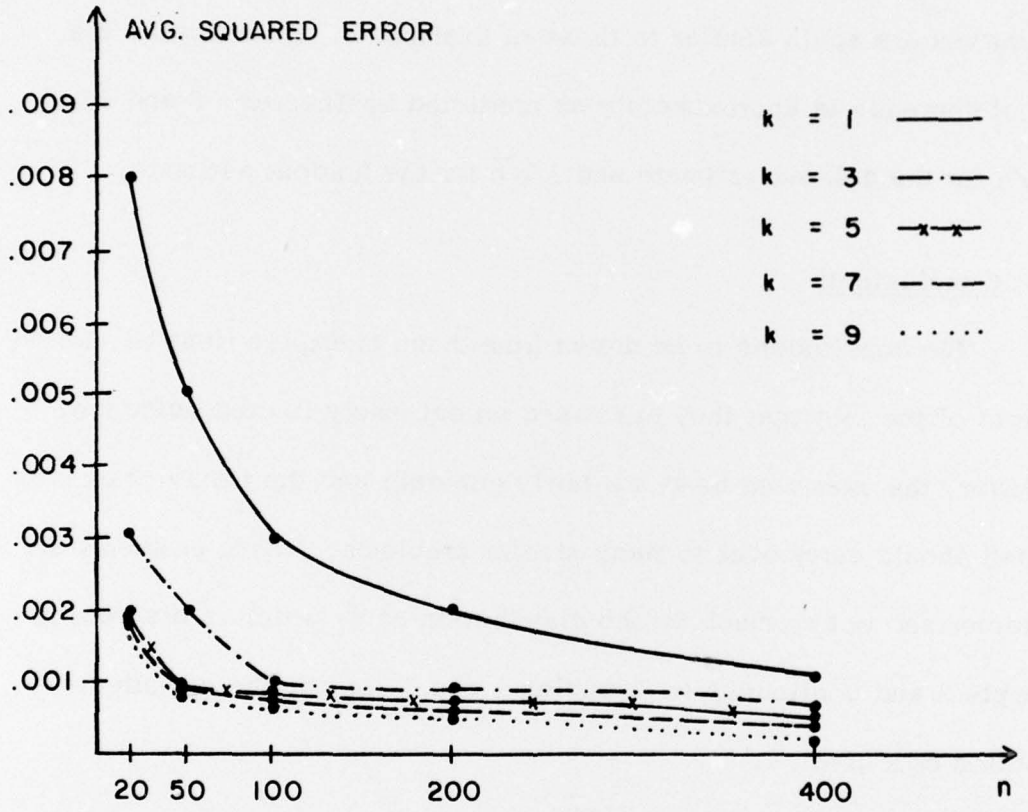


Figure 30. Average Squared Error of the Holdout Estimate for Example 3

then the performance of the estimates would be expected to approach that of Example 2 as σ becomes small.

Figs. 24 and 30 show the average squared error of the deleted and holdout estimates respectively, for the same values of n and ϵ. The curves are again similar to those of Example 2. Once again, the rate of decrease is approximately as predicted by Theorems 5 and 6, at $1/n$ for the deleted estimate and $1/\sqrt{n}$ for the holdout estimate.

IV.5 Conclusions

The conclusions to be drawn from these examples must be viewed in light of the fact that they are based on extremely limited evidence. However, the examples used are fairly general, and the behavior exhibited should carry over to many similar problems. These problems are characterized not so much by the distribution of θ, which is discrete in Example 2 and continuous in Examples 1 and 3, but by the smooth distribution of x given θ.

The most obvious suggestion made by these examples is that fairly large values or k can be used without risking a reduction in performance either of the rule or of the deleted or holdout estimates of the risk. In fact, for the estimation problem in Example 3, larger values of k not only improved L_n, but they also improved the deleted and holdout estimates. For Example 3, both L_n and the estimate of L_n were still improving with $k = 9$ in the case where $n = 20$.

We also note that in every case, the average value of L_n was quite near its limiting value long before the deleted estimate was achieving acceptable performance levels. Although this does not indicate that L_n itself is converging that rapidly, it does seem to offer some evidence that L_n may be converging somewhat more rapidly than the error estimates. It may be possible to take advantage of this to improve the performance of the holdout estimate by holding out a larger portion of the data. In the examples above, $\ell = \sqrt{n}$ observations were used to compute the holdout estimate. In cases where L_n converges quickly, $L_{n-\ell}$ will be close to L_n for larger values of ℓ. Since the holdout estimate is essentially an unbiased estimate of $L_{n-\ell}$, and the variance of the estimate declines with ℓ, performance could be improved by increasing ℓ.

Finally, we note that for each value of k, n, and ϵ used in all the above examples, the bounds given by Theorems 5 and 6 are equal to one, and so are extremely pessimistic, as expected. However, the rate of decrease of the average squared error is consistent with the rate of decrease of the theoretical bounds.

V. SUMMARY

An attempt has been made to identify the questions which are of genuine importance to a statistician who is interested in evaluating nonparametric estimation rules. It was pointed out that asymptotic performance is an important criterion in many cases, even though data sets are always finite in practical situations. Considering the reasons behind interest in asymptotic results, it appears that the important question in this area concerns what statements can be made about the performance of the rule as the size of a single data set is increased. The importance of estimating finite sample performance was discussed and the value of a distribution-free bound on the error of such estimates was indicated.

The asymptotic performance of various nearest neighbor rules and loss functions was analyzed, and the results concerning the convergence of the conditional risk are quite strong considering the simple nature of the rules and the minimal assumptions made concerning the problem structure. In addition, distribution-free bounds on the error of two different estimates of finite sample performance are derived for a class of estimation rules which include k-nearest neighbor rules. The distribution-free nature of these bounds indicates that they are undoubtedly not tight for many specific choices of the distribution

function. For comparison purposes, a simulation study was carried out so that the bounds for a few specific distributions could be compared with the theoretical bounds.

BIBLIOGRAPHY

1. T. M. Cover, "Estimation by the Nearest Neighbor Rule," IEEE Trans. Information Theory, Vol. IT-14, January 1968, pp. 50-55.

2. E. Fix and J. L. Hodges, Jr., "Discriminatory Analysis, Nonparametric Discrimination, Consistency Properties," USAF School of Aviation Medicine, Randolph Field, Tex., Project 21-49-004, Rept. 4, Contract AF41(128)-31, February 1951.

3. E. Fix and J. L. Hodges, Jr., "Discriminatory Analysis, Small Sample Performance," USAF School of Aviation Medicine, Randolph Field, Tex., Project 21-49-004, Rept. 11, August 1952.

4. T. J. Wagner, "Convergence of the Nearest Neighbor Rule," IEEE Trans. Information Theory, Vol. IT-17, September 1971, pp. 566-571.

5. G. T. Toussaint, "Bibliography on Estimation of Misclassification," IEEE Trans. Information Theory, Vol. IT-20, July 1974, pp. 472-479.

6. T. M. Cover and P. E. Hart, "Nearest Neighbor Pattern Classification," IEEE Trans. Information Theory, Vol. IT-13, January 1967, pp. 21-27.

7. L. Breiman, Probability. Reading, Mass.: Addison-Wesley, 1968.

8. M. Loeve, Probability Theory, 3rd ed.,Princeton, N.J.: Van Nostrand, 1963.

9. W. H. Rogers and T. J. Wagner, "A Finite Sample Distribution-Free Performance Bound for Local Discrimination Rules," to appear in Annals of Statistics.

10. W. Hoeffding, "Probability Inequalities for Sums of Bounded Random Variables," J. Amer. Statist. Ass., Vol. 58, 1963, pp. 13-30.

11. T. J. Wagner, "Deleted Estimates of the Bayes Risk," Ann. Statist., Vol. 1, 1973, pp. 359-362.

12. T. M. Cover, "Rates of Convergence of Nearest Neighbor Procedures," Proc. of 1st Annual Hawaii Conf. on System Science, 1968, pp. 413-415.

13. W. Feller, An Introduction to Probability Theory and Its Applications, 3rd ed., New York, N. Y.: John Wiley and Sons, 1968.

14. R. O. Duda and P. E. Hart, Pattern Classification and Scene Analysis, New York, N. Y.: John Wiley and Sons, 1973.

15. T. S. Ferguson, Mathematical Statistics: A Decision Theoretic Approach, New York, N. Y.: Academic Press, 1967.