A NONPARAMETRIC TWO-STAGE PROCEDURE FOR ESTIMATING A PROBABILITY DENSITY FUNCTION

Matthew P. Wand
Department of Mathematics
The University of Wollongong

MATHEMATICS HONOURS PROJECT

1985

ABSTRACT

Given a sample of continuous random variables $X_1, X_2, \ldots, X_n$ taken from a population with unknown probability density function f, the problem considered in this project is that of finding an estimate of f based only on the sample. A procedure for estimating f is proposed which constructs an estimate in two stages. The first stage involves using the order statistics to form a preliminary estimate which takes the form of a step function. This function is then "smoothed" using the normalised Hermite functions.

ACKNOWLEDGEMENTS

Special thanks must go to my supervisor, Dr Vic Drastik, for his helpful ideas and comments and for the encouragement and guidance he has given me while I have been working on this project.

I would also like to express gratitude to Jeff Dewynne, Adam Kucera and Tony van Ravenstein for their support and advice during the year. Special mention must go to Jeff for the help he has given in the production of this thesis using the TEX document processing system. Thanks must also go to Mr. Peter Castle for his help in the final printing of this thesis.

Finally, I must thank my family and friends and the staff of the Department of Mathematics for their support and understanding throughout 1985.

Matthew Wand

Matthew Wand

CHAPTER ONE

THEORETICAL BACKGROUND

1.1 Introduction

The problem of nonparametric probability density estimation arises in probability and mathematical statistics. In this chapter we shall firstly present a brief coverage of probability theory. This will allow us to define the features of mathematical statistics which are relevant to the problem.

1.2 Probability Spaces

A statistical experiment is an experiment with the following properties:

1. each of the possible outcomes of the experiment is known in advance,
2. the outcome of each performance of the experiment is not known in advance,
3. the experiment can be performed repeatedly under identical conditions.

Consider the set of all possible outcomes of the experiment. This set is called the sample space and will be denoted by the Greek letter Ω.

Let S be a family of subsets of Ω satisfying

1. $\varnothing \in S$, where $\varnothing$ is the null set.
2. If $A \in S$ then $\Omega - A \in S$.

3. If $A_1, A_2, \ldots\ldots$ is a disjoint sequence of subsets of S then $\bigcup_{n=1}^{\infty} A_n \in S$.

Then S is a σ-algebra on Ω and the sets in S are called events. The pair (Ω, S) is called the sample space of the statistical experiment. With respect to the pair (Ω, S) we define a probability measure P as a function from Ω into [0, 1] such that

1. $P(\Omega) = 1$.

2. If $A_1, A_2, \ldots\ldots$ is a disjoint sequence of events in S then

$$P\left(\bigcup_{n=1}^{\infty} A_n\right) = \sum_{n=1}^{\infty} P(A_n).$$

The triple (Ω, S, P) is called a probability space.

1.3 Random Variables and their Probability Distributions

A useful concept for measuring probabilities associated with (Ω, S, P) is that of a random variable. A random variable X is a real-valued function on Ω with the requirement that

$$X^{-1}(B) \in S \tag{1.3.1}$$

for any Borel set B in $\Re$ where $\Re$ denotes the set of real numbers.

The cumulative distribution function (c.d.f.) F of a random variable X is a real-valued function defined on $\Re$ such that

$$F(x) = P\{\omega : X(\omega) \leq x\} = P\{X \leq x\}, \qquad x \in \Re. \tag{1.3.2}$$

$F(x)$ is a nondecreasing function of x. Also

$$\lim_{x \to -\infty} F(x) = 0 \tag{1.3.3}$$

while

$$\lim_{x \to \infty} F(x) = 1. \tag{1.3.4}$$

We are now in a position to define two important types of random variables. A random variable X is said to be discrete if there is some countable set C in $\Re$ such that

$$P\{\omega : X(\omega) \in C\} = P\{X \in C\} = 1. \tag{1.3.5}$$

X is a continuous random variable if there is a nonnegative function f so that

$$F(x) = \int_{-\infty}^{x} f(t)\,dt. \tag{1.3.6}$$

We call f the probability density function (p.d.f.) of X.

Note that $F'(x) = f(x)$ when the derivative exists and

$$P\{X \in B\} = \int_B f(x)\,dx \tag{1.3.7}$$

for any set Borel set B.

Two noteworthy properties of f are

1. $f \geq 0$.
2. $\int_{-\infty}^{\infty} f(x)\,dx = 1$.

The basic difference between discrete and continuous random variables is that while a discrete random variable can assume only a countable number of points, a continuous random variable may assume a continuum of points in $\Re$.

It is the p.d.f. of a continuous random variable in which we are interested and so for the remainder of this thesis we will be dealing only with continuous random variables.

1.4 Expected Value and Variance of a Random Variable

Let X be a continuous random variable. Assuming that $xf(x)$ is absolutely integrable, the expected value of X exists and is equal to

$$\mu = E(X) = \int_{-\infty}^{\infty} xf(x)\,dx\,. \tag{1.4.1}$$

The variance of X is given by

$$\mathrm{Var}(X) = \int_{-\infty}^{\infty} (x-\mu)^2 f(x)\,dx\,, \tag{1.4.2}$$

assuming that this integral is finite.

Note that $E(X)$ and $\mathrm{Var}(X)$ can be defined analogously for discrete random variables.

1.5 Random Samples

We say that the n random variables $X_1, X_2, \ldots, X_n$ are independent if

$$P\left[\bigcap_{i=1}^{n} \{X_i \in B_i\}\right] = \prod_{i=1}^{n} P\{X_i \in B_i\} \tag{1.5.1}$$

for any Borel sets $B_1, B_2, \ldots, B_n$.

If $X_1, X_2, \ldots, X_n$ are independent continuous random variables such that each has the same p.d.f. f then $X_1, X_2, \ldots, X_n$ is called a random sample of size n.

Suppose we relabel the sample as $X_{(1)}, X_{(2)}, \ldots, X_{(n)}$ where

$$X_{(1)} \leq X_{(2)} \leq \cdots \leq X_{(n)}\,. \tag{1.5.2}$$

We call the $X_{(i)}$'s the order statistics of the sample.

CHAPTER TWO

NONPARAMETRIC PROBABILITY DENSITY FUNCTION ESTIMATION

2.1 Introduction

The problem of nonparametric probability density function estimation can now be defined. Let $X_1, X_2, \ldots, X_n$ be a random sample of continuous random variables from a population having an unknown p.d.f. f. An estimate of f, usually denoted by $\hat{f}$, is to be found using only the n sample points. Since no assumptions are made about the distributional form of f, any procedure which attempts to solve this problem is labelled as nonparametric.

The most widely known estimate of a population density is that provided by the classical histogram. Section 2.2 takes a historical viewpoint by outlining this method.

Modern nonparametric probability density function estimation was pioneered by Rosenblatt (1956) with his paper entitled: "Remarks on Some Nonparametric Estimates of a Density Function". Since then the literature on the subject has seen three basic types of density estimates emerge. These are

1. kernel estimates,
2. orthogonal series estimates, and
3. maximum likelihood estimates.

In sections 2.3, 2.4 and 2.5 we shall discuss each of these estimates in turn.

2.2 Histogram Estimates

The classical histogram was first used in the 17th century by John Graunt, a London haberdasher, who used it to represent birth and death data.

The histogram estimate is obtained by partitioning the interval between the lowest sample point and the highest sample point into p subintervals $I_1, I_2, \ldots, I_p$. Consider some interval I_j. Let ℓ_j denote its length and n_j denote the number of sample points in I_j. The value of the p.d.f. estimate over this interval is given by

$$\hat{f}(x) = \frac{n_j}{n\ell_j}. \tag{2.2.1}$$

This is a simple way of estimating f although it is somewhat inefficient since much of the information in the sample is ignored. Another problem with the histogram estimate is that of choosing the partition intervals, a choice which is often crucial.

2.3 Kernel Estimates

The basic form of a kernel estimate is given by

$$\hat{f}(x) = \frac{1}{nh}\sum_{i=1}^{n} K\left(\frac{x - X_i}{h}\right), \tag{2.3.1}$$

where K is some measurable function such that

$$\int_{-\infty}^{\infty} K(t)\,dt = 1, \tag{2.3.2}$$

and is called the kernel. The parameter h is often referred to as the window width.

The kernel can be viewed as a means of smoothing the probability "spikes" of the empirical density function $\hat{f}_n$ defined to be

$$\hat{f}_n(x) = \frac{1}{n}\sum_{i=1}^{n} \delta(x - X_i), \tag{2.3.3}$$

where δ is the Dirac delta function.

Parzen (1962) introduced the kernel estimate when he generalised Rosenblatt's "naive estimate". The naive estimate was treated as a kernel estimate with kernel

$$K(y) = \begin{cases} \frac{1}{2}, & |y| \leq 1 \\ 0, & |y| > 1 . \end{cases} \tag{2.3.4}$$

The literature has since seen various kernels receiving consideration. Two notable examples are the Gaussian kernel (e.g. Specht (1971)),

$$K_1(y) = \frac{1}{\sqrt{2\pi}} e^{(-1/2)y^2}, \tag{2.3.5}$$

and the Fourier integral kernel (Davis (1975)),

$$K_2(y) = \frac{\sin y}{\pi y} . \tag{2.3.6}$$

Simulation studies (e.g. Wegman (1972)) have shown that the choice of the window width is critical whichever kernel is used whereas the choice of the actual kernel is relatively unimportant. It is for this reason that many of the later papers on kernel estimation have concentrated on estimating the optimal window width.

2.4 Orthogonal Series Estimates

For an orthogonal series estimate it is assumed that with respect to some weight function w, defined over the interval (a, b),

$$\int_a^b [f(x)]^2 w(x)\, dx < \infty , \tag{2.4.1}$$

so that f can be expanded as

$$f(x) = \sum_{j \in I} a_j \psi_j(x) , \tag{2.4.2}$$

where I is some index set and $\{\psi_j : j \in I\}$ is a set of orthogonal functions with respect to w.

The orthogonal series estimate is of the form

$$\hat{f}(x) = \sum_{j \in J} \hat{a}_j \psi_j(x), \tag{2.4.3}$$

where J is a finite subset of I and $\hat{a}_j$ is some estimate of a_j. The usual estimate of the coefficients is

$$\hat{a}_j = \frac{1}{n} \sum_{i=1}^{n} \psi_j(X_i) w(X_i), \tag{2.4.4}$$

which is unbiased.

Kronmal and Tarter (1968) have considered the use of Fourier series and Schwartz (1967) has investigated the use of normalised Hermite functions.

Associated with an orthogonal series estimate is a "stopping rule" for determining the number of terms in the expansion. The stopping rule is an important component of an orthogonal series estimate since the number of terms usually affects the performance of the estimate to a high degree.

2.5 Maximum Likelihood Estimates

Maximum likelihood estimation in statistical estimation theory is a widely used technique which involves choosing an estimator that maximises the likelihood L, where

$$L = \prod_{i=1}^{n} f(X_i). \tag{2.5.1}$$

In the case of nonparametric probability density function estimation we define the likelihood that a function ν is the true p.d.f. corresponding to the sample $X_1, X_2, \ldots, X_n$ to be

$$L(\nu) = \prod_{i=1}^{n} \nu(X_i). \tag{2.5.2}$$

If $L(v)$ is maximised over $v \in L^1(R)$ then the estimate will be

$$\hat{f}(x) = \frac{1}{n} \sum_{i=1}^{n} \delta(x - X_i) \tag{2.5.3}$$

which is the empirical density function mentioned in Section 2.3. This solution, however, is unacceptable due to its over-roughness.

To overcome this problem Good and Gaskins (1971) proposed the inclusion of a penalty function $\Phi(v)$ which would penalise roughness in v. This involves replacing the likelihood by

$$\hat{L}(v) = \prod_{i=1}^{n} v(X_i) e^{-\Phi(v)}. \tag{2.5.4}$$

The estimate of Good and Gaskins is the one which minimises $\hat{L}(v)$ and is called the maximum penalised-likelihood estimate.

This idea was extended by de Montricher, Tapia and Thompson (1975). They considered the solution of

$$\begin{aligned} &\text{Maximise} && \hat{L}(v) \\ &\text{subject to } v \in H, && \int v(t)\,dt = 1 \text{ and } v \geq 0, \end{aligned}$$

where H is a particular function space.

The practical implementation of this idea was investigated by Scott, Tapia and Thompson (1980). They proposed a numerical solution to the above constrained optimisation problem which involves meshing techniques. The resultant estimate is called the discrete maximum penalised-likelihood estimate.

CHAPTER THREE

A TWO-STAGE ESTIMATE OF THE PROBABILITY DENSITY FUNCTION

3.1 Introduction

In this chapter we shall propose a method for finding an estimate of an unknown p.d.f. f which uses only the sample $X_1, X_2, \ldots, X_n$ taken from the density f.

The method is one which constructs the estimate in two stages. The first stage involves obtaining an "empirical" estimate of f based on the order statistics of the sample. We will call this estimate the preliminary estimate. This stage is described in Section 3.2.

In the second stage, discussed in Section 3.3, the normalised Hermite functions are used as a means of "smoothing" the preliminary estimate.

3.2 Construction of the Preliminary Estimate

Let $X_{(1)}, X_{(2)}, \ldots, X_{(n)}$ be the order statistics of the sample. We define the inter-order statistic difference Δ_i as

$$\Delta_i = X_{(i+1)} - X_{(i)}, \quad i = 1, 2, \ldots, n-1. \tag{3.2.1}$$

We now examine a fundamental characteristic of a random sample having p.d.f. f. Consider an interval in $\Re$ over which f assumes relatively large function values. Since the integral of f over this interval is also large there is a greater probability of sample points occurring in the interval. We would therefore expect there to be a clustering of points in this region. This would mean that the inter-order statistic differences are small when f is large. On the other hand, on an interval

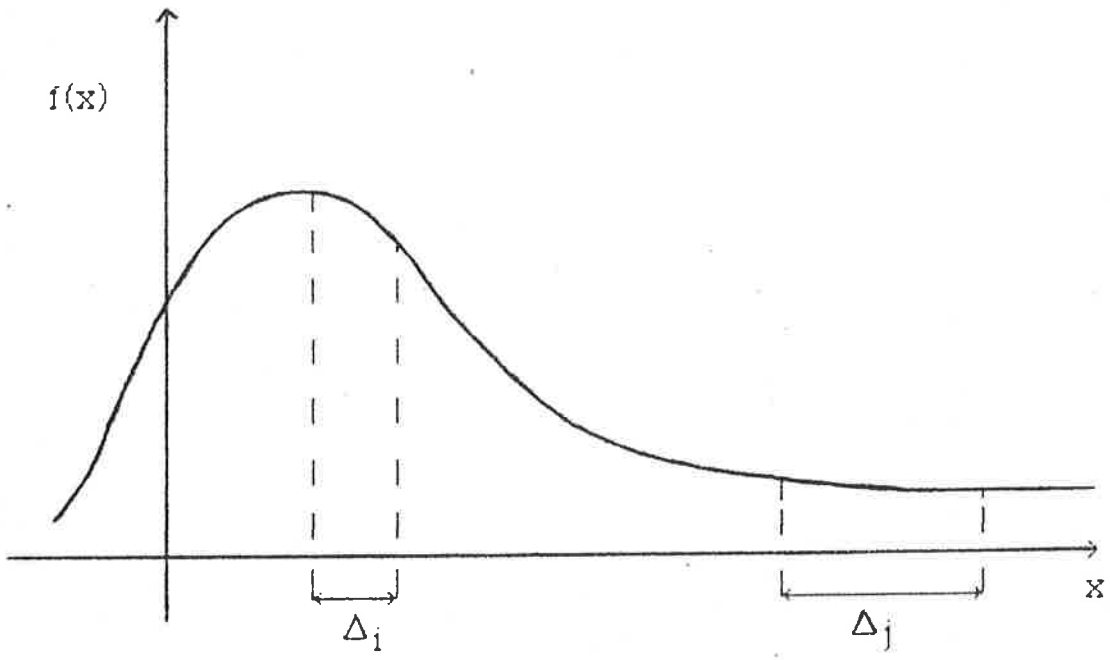


Figure 3.2.1

where f is small the sample points should be sparsely dispersed and subsequently there will be large inter-order statistic differences (see Figure 3.2.1).

It is therefore apparent that there is an inverse relationship between f and the Δ_j values. We may write this as

$$\frac{1}{\overline{\Delta}} \approx f\,, \tag{3.2.2}$$

where $\overline{\Delta}$ denotes some average of the Δ_j values. An attempt will now be made to form a preliminary estimate of f based on this idea.

A simple approach is one for which $\overline{\Delta} = \Delta_i$ so that the preliminary estimate $\hat{f}_p$ is

$$\hat{f}_p(x) = \frac{1}{(n-1)\Delta_i}\,, \qquad X_{(i)} \leq x < X_{(i+1)}\,, \quad i = 1, 2, \ldots, n-1\,, \tag{3.2.3}$$

with the $n-1$ inserted to ensure that $\hat{f}_p$ integrates to unity.

The problem with this estimate is that whenever two adjacent sample points occur very close to each other the value of $\hat{f}_p(x)$ will be far too high. Consequently $\hat{f}_p$, in this case, is usually very rough and so a smoother version must be sought.

The problem with the first proposal stems from the fact that the value of $\hat{f}_p(x)$ over $(X_{(i)}, X_{(i+1)})$ involves only Δ_i and none of the neighbouring inter-order statistic differences. A smoother estimate should be achieved with an estimate of the form

$$\hat{f}_p(x) = \frac{1}{\sum_{j \in K} \alpha_{ij} \Delta_j}\,, \qquad X_{(i)} \leq x < X_{(i+1)} \tag{3.2.4}$$

where K is some set of integers neighbouring i.

An obvious problem with this modification is that of choosing the weighting coefficients denoted here by α_{ij}. The heaviest weighting should of course be on Δ_i and the weighting should monotonically decrease with movement away from the ith position. We now attempt to find suitable weightings by considering an estimate of the quantity $E\left(\frac{1}{f(X_{(i)})}\right)$.

If we let F denote the cumulative distribution function corresponding to f and B be the beta function then the following result may be derived,

$$E\left(\frac{1}{f(X_{(i)})}\right) = \frac{1}{B(i, n+1-i)} \int_{-\infty}^{\infty} F^{i-1}(t)[1-F(t)]^{n-i}\, dt\,. \tag{3.2.5}$$

An appropriate estimate of F is that proposed by Read (1972),

$$\hat{C}(x) = \frac{i}{n+1} + \frac{x - X_{(i)}}{(n+1)\Delta_i}\,, \qquad , X_{(i)} \leq x < X_{(i+1)} \tag{3.2.6}$$

which has the property

$$\hat{C}(X_{(i)}) = \frac{i}{n+1}\,. \tag{3.2.7}$$

Let $X_{(0)}$ and $X_{(n+1)}$ be estimators of the lower and upper truncation points of the density. We can then estimate $E\left(\frac{1}{f(X_{(i)})}\right)$ as

$$\hat{E}\left(\frac{1}{f(X_{(i)})}\right) = (n+1)\sum_{j=0}^{n} d_{ij}\Delta_j\,, \quad i = 1, 2, \ldots, n\,, \tag{3.2.8}$$

where

$$d_{ij} = \frac{1}{B(i, n+1-i)} \int_{X_{(j)}}^{X_{(j+1)}} \hat{C}^{i-1}(x)[1-\hat{C}(x)]^{n-i}\, dx\,, \quad i = 1, \ldots, n\,,\ j = 0, \ldots, n\,. \tag{3.2.9}$$

The integral in this expression can be evaluated using the binomial theorem and because of (3.2.7), the d_{ij} values depend only upon n and not on the sample values.

An estimate for f can now be obtained as the reciprocal of $\hat{E}\left(\frac{1}{f(X_{(i)})}\right)$. Thus we get

$$\hat{f}_p(x) = \frac{1}{(n+1)\sum_{j=0}^{n} d_{ij}\Delta_j}\,, \quad X_{(i)} \leq x < X_{(i+1)}\,. \tag{3.2.10}$$

It should be observed that the value of $\hat{f}_p(x)$ over the interval $(X_{(i)}, X_{(i+1)})$ now uses all of the Δ_j values ($j = 0, 1, \ldots, n$). If we let $\mathbf{d}_i$ be the vector of d_{ij} values, $j = 0, 1, \ldots, n$, and $\mathbf{v}$ be the vector of Δ_j values then the reciprocal of our estimate over $(X_{(i)}, X_{(i+1)})$ is

$$(n+1)\mathbf{d}_i^T\mathbf{v}\,. \tag{3.2.11}$$

Let us now examine the vector $\mathbf{d}_i$. Firstly it can be shown that the entries of $\mathbf{d}_i$ are positive and add up to one. Also there is the relationship

$$d_{ij} = d_{n-i,n-j} \tag{3.2.12}$$

which means that the vector $\mathbf{d}_{n-i}$ is the same as $\mathbf{d}_i$ with its entries reversed.

From the above comments the largest entry of $\mathbf{d}_i$ should ideally be d_{ii} and the other d_{ij} values should monotonically decrease as j moves away from i. This indeed is the case and consequently this estimate is a great improvement on the one initially proposed. Unfortunately there is a problem with the $\mathbf{d}_i$ vector as n increases.

For the estimate to be consistent (that is, converge to the true p.d.f. as the sample size increases) we require an averaging over an increasing number of the Δ_j's. This follows from the consistency theorem for the histogram (see Tapia and Thompson (1978) Theorem 3, pp. 46-48). For certain values of i the $\mathbf{d}_i$ vector does not satisfy this requirement since, as n increases, the vector approaches a constant vector which has zero entries a fixed distance from the ith position.

If the entries of $\mathbf{d}_i$ are plotted against $j = 0, 1, \ldots, n$ then an intrinsic property of the plot is that the points form a skewed bell shape with peak occurring at $j = i$. A common probability density with this property is the gamma distribution and this connection motivates our final adjustment to the estimate.

We shall replace the $\mathbf{d}_i$ vectors by $\mathbf{g}_i$, $i = 1, 2, \ldots, n-1$, where

$$g_{ij} = N j^{\alpha-1} \exp(-j/\beta_i), \qquad i, j = 1, 2, \ldots, n-1, \tag{3.2.14}$$

where $N = N(i, \alpha, \beta_i)$ is chosen to ensure that the vector entries sum to one. Note that the right hand side of (3.2.14) has the same form as the gamma p.d.f. .

We want g_{ii} to be the maximum entry of $\mathbf{g}_i$. Note that

$$\ln(g_{ij}) = (\alpha - 1)\ln(j) - \frac{j}{\beta_i} + ln(N). \tag{3.2.15}$$

We shall derive the appropriate constraint by treating j as a continuous variable. Thus

$$\frac{\partial}{\partial j}\ln(g_{ij}) = \frac{\alpha - 1}{j} - \frac{1}{\beta_i}. \tag{3.2.16}$$

Setting this derivative to zero we obtain the constraint

$$\alpha - 1 = \frac{i}{\beta_i} \tag{3.2.17}$$

for g_{ii} to be the maximum.

The only thing remaining is the determination of the β_i values. To do this we shall treat the entries of $\mathbf{d}_i$ as the vector of probability masses of some discrete random variable X_i^d. We shall also define a gamma random variable X_i^g having parameters α and β_i. The value of β_i will be chosen by solving

$$\mathrm{Var}(X_i^d) = \mathrm{Var}(X_i^g). \tag{3.2.18}$$

Experimentation shows that $\mathrm{Var}(X_i^d)$ is approximately equal to $\frac{n}{4}$. Thus we get the equation

$$\frac{n}{4} = \alpha\beta_i^2. \tag{3.2.19}$$

Using the constraint (3.2.17) we get

$$\beta_i = \frac{1}{2}[-i + \sqrt{i^2 + n}], \qquad i = 1, 2, \ldots, n-1. \tag{3.2.20}$$

Finally we recall that $d_{ij} = d_{n-i,n-j}$. To ensure that the g_{ij} have this property we calculate $\mathbf{g}_i$ for values of i from 1 through to $\frac{n}{2}$ if n is even and $\frac{n-1}{2}$ if n is odd. The remaining $\mathbf{g}_i$ vectors are generated by

$$g_{n-i,n-j} = g_{ij}. \tag{3.2.21}$$

The $\mathbf{g}_i$ vectors do not have the defect of the $\mathbf{d}_i$ vectors and their use sees the estimate improving for increasing sample sizes. We shall therefore accept the $\mathbf{g}_i$ vectors as suitable smoothing coefficients. Our preliminary estimate of f takes the form

$$\hat{f}_p(x) = \frac{1}{(n-1)\mathbf{g}_i^T\mathbf{v}}, \qquad X_{(i)} \le x < X_{(i+1)}. \tag{3.2.22}$$

3.3 Fitting the Hermite Series

At this stage we have a nonparametric estimate of the probability density function. This estimate $\hat{f}_p$ takes the form of a step function with the individual steps occurring between each of the order statistics. The problem with this estimate is that, because of random error in the data, some of the step heights vary quite substantially from the corresponding function values of the true p.d.f.. These fluctuations occur over comparatively small intervals and so a "global" averaging of the step function should be beneficial.

We shall smooth the preliminary estimate with the use of the normalised Hermite functions. These will now be defined.

The jth Hermite polynomial is given by

$$H_j(x) = (-1)^j e^{x^2} \frac{d^j}{dx^j} e^{-x^2} \tag{3.3.1}$$

and the jth normalised Hermite function is

$$\phi_j(x) = \frac{1}{\sqrt{2^j j!\pi^{\frac{1}{2}}}} e^{-(1/2)x^2} H_j(x) \quad j = 0, 1, \ldots\ldots \tag{3.3.2}$$

These functions form a complete orthonormal set in $L^2(\Re)$, the set of real-valued functions which are square integrable over the real line.

Let q be a function in $L^2(\Re)$. Then q has the Hermite expansion

$$q(x) = \sum_{j=0}^{\infty} a_j \phi_j(x) \tag{3.3.3}$$

where

$$a_j = \int_{-\infty}^{\infty} q(x)\phi_j(x)\,dx\,. \tag{3.3.4}$$

It should be noted here that the Hermite series fit will make the final estimate a continuous function. This is intuitively appealing since most of the commonly used p.d.f.'s are in fact continuous functions rather than step functions.

Recall that two basic properties of the p.d.f. f are

1. $f \geq 0$.

2. $\int_{-\infty}^{\infty} f(x)\,dx = 1$.

To ensure that our final estimate has the first of these properties we shall fit a Hermite series to $\hat{f}_p^{\frac{1}{2}}$. Let s denote the function $\hat{f}_p^{\frac{1}{2}}$. We seek an approximation to s of the form

$$s_m = \sum_{j=0}^{m} a_j \phi_j \,. \tag{3.3.5}$$

Here m indicates some truncation point of the series and the choice of its value will be discussed later.

We shall now choose the a_j coefficients so that the L^2 distance between s and s_m is minimised. Also, to satisfy property 2., we require s_m^2 to integrate to one. Therefore we have to solve the constrained minimisation problem

$$\text{Minimise} \quad \int_{-\infty}^{\infty} (s(x) - s_m(x))^2\,dx$$

$$\text{subject to} \quad \int_{-\infty}^{\infty} s_m^2(x)\,dx = 1 \,.$$

To solve this problem we introduce a Lagrange multiplier λ and define an auxiliary function Ψ as

$$\Psi = \int_{-\infty}^{\infty} (s(x) - s_m(x))^2\,dx + \lambda\left[\int_{-\infty}^{\infty} s_m^2(x)\,dx - 1\right]. \tag{3.3.6}$$

Now

$$\int_{-\infty}^{\infty} s_m^2(x)\,dx = \sum_{j=0}^{m} a_j^2 \,, \tag{3.3.7}$$

using the orthormality property

$$\int_{-\infty}^{\infty} \phi_j(x)\phi_k(x)\,dx = \begin{cases} 1 \,, & j = k \\ 0 \,, & j \neq k \,. \end{cases} \tag{3.3.8}$$

We can thus rewrite Ψ as

$$\Psi = \sum_{j=0}^{m} a_j^2 - 2\sum_{j=0}^{m}\left[a_j \int_{-\infty}^{\infty} \phi_j(x)s(x)\,dx\right] + \int_{-\infty}^{\infty} s^2(x)\,dx + \lambda\left[\sum_{j=0}^{m} a_j^2 - 1\right]. \tag{3.3.9}$$

Hence

$$\frac{\partial \Psi}{\partial a_r} = 2\left[a_r - \int_{-\infty}^{\infty} \phi_r(x) s(x)\, dx\right] + 2\lambda a_r \,. \tag{3.3.10}$$

Setting this partial derivative to zero we get

$$a_r = \frac{b_r}{1+\lambda}, \qquad r = 0, 1, \ldots, m \,, \tag{3.3.11}$$

where

$$\begin{aligned} b_r &= \int_{-\infty}^{\infty} \phi_r(x) s(x)\, dx \\ &= \int_{-\infty}^{\infty} \phi_r(x) \hat{f}_p^{\frac{1}{2}}(x)\, dx \,. \end{aligned} \tag{3.3.12}$$

Next, we substitute (3.3.7) into the constraint to give

$$\sum_{j=0}^{m} a_j^2 = 1 \,. \tag{3.3.13}$$

Combining the result with (3.3.11) we obtain the solution

$$a_r = \frac{b_r}{\sqrt{\sum_{j=0}^{m} b_j^2}}, \qquad r = 0, 1, \ldots, m \,. \tag{3.3.14}$$

Our final estimate of the p.d.f. f is

$$\hat{f}(x) = \left[\sum_{j=0}^{m} a_j \phi_j(x)\right]^2 . \tag{3.3.15}$$

We are now left with the problem of choosing the value of the truncation parameter m. If a very high number of terms is included in the series then $\hat{f}$ will approximate $\hat{f}_p$ too closely and the "noise" inherent in the preliminary estimate will not be optimally smoothed. On the other hand, the number of terms in the series should be large enough to ensure that the estimate possesses the global properties of $\hat{f}_p$. The value of m should be chosen so that there is a compromise between each of these two states. We now present a procedure for choosing m which attempts to reach such a compromise.

The L^2 distance between $\hat{f}$ and $\hat{f}_p$ can be approximated by the quantity

$$D_m = \frac{1}{n} \sum_{i=1}^{n} \left[\hat{f}(X_i) - \hat{f}_p(X_i)\right]^2 . \tag{3.3.16}$$

We define

$$r_m = \frac{D_m}{D_{m-1}}, \quad m = 1, 2, \ldots\ldots \tag{3.3.17}$$

so that if r_m is very close to unity then $\hat{f}$ should be a close approximation to $\hat{f}_p$. Because r_m is based on an approximation to the L^2 error we shall deal with an averaging of five adjacent r_m values and define

$$\overline{r_m} = \frac{1}{5}\sum_{j=-2}^{2} r_{m+j}, \quad m = 3, 4, \ldots\ldots \tag{3.3.18}$$

The procedure to be used for choosing the number of terms can now be given. Commence with m equal to 8 and increase its value until $\overline{r_m}$ exceeds 0.95 . When this criterion is met the procedure is stopped and the final value of m is the one used in (3.3.15).

3.4 Shifting and Scaling of the Data

Because the performance of the estimate given in this chapter may be affected by location and scale changes of the true p.d.f. it is worthwhile standardising the data before applying the density estimation procedure. We shall use the sample median, $\hat{\mu}$, as the location shift and for the scaling we use

$$\hat{\sigma} = \frac{1}{2}(X_{(n+1-\nu)} - X_{(\nu)}), \tag{3.4.1}$$

where $\nu = [\frac{n}{5}]$ and $[\cdot]$ is the truncation function. We standardise the data via the transformation

$$X_i \rightarrow \frac{X_i - \hat{\mu}}{\hat{\sigma}}, \tag{3.4.2}$$

and use the standardised sample to find the estimate. The function obtained as the estimate is transformed back to yield the final estimate.

CHAPTER 4

MONTE CARLO RESULTS AND COMPARISONS

4.1 Introduction

Comparisons among a selection of previously proposed nonparametric density estimates and the estimate proposed in Chapter 3 are now made via a Monte Carlo simulation study.

To compare a set of density estimates a measure of goodness of a particular estimate is needed. A widely used criterion for the goodness of an estimate $\hat{f}$ of the true p.d.f. f is based on the mean integrated square error (M.I.S.E.),

$$E\int_{-\infty}^{\infty}[f(x)-\hat{f}(x)]^2\,dx\,. \tag{4.1.1}$$

An approximation to the M.I.S.E. will be used to compare the estimates considered in this chapter.

4.2 A Brief Description of the Estimates to be Compared

Virtually all of the density estimates discussed in the literature depend on a single parameter which governs the smoothness, as well as the goodness, of the estimate. A defect of many of these estimates is that the choice of the smoothing parameter is not specified. The estimate described in Chapter 3 does not have this defect and so it seems reasonable that it should only be compared with other completely data-based estimates. The estimates to be used for comparison are a histogram estimate, the Gaussian kernel estimate of Scott, Tapia and Thompson (1977) and the Hermite series estimate of Schwartz (1967).

The histogram estimate to be used is one based on ten equal-lengthed intervals formed by subdividing the interval $(X_{(1)}, X_{(n)})$.

The kernel estimate of Scott, Tapia and Thompson is

$$\hat{f}(x) = \frac{1}{nh} \sum_{i=1}^{n} \phi\left(\frac{x - X_i}{h}\right), \tag{4.2.1}$$

where

$$\phi(z) = \frac{1}{\sqrt{2\pi}} e^{(-1/2)z^2}. \tag{4.2.2}$$

The window width h is chosen as the limit of the sequence (h_i) where

$$h_0 = X_{(n)} - X_{(1)}, \tag{4.2.3}$$

$$h_{i+1} = \pi^{\frac{1}{10}} \beta(h_i) n^{-\frac{1}{5}} \tag{4.2.4}$$

and

$$\beta(h)^{-5} = \frac{3}{8\sqrt{\pi} n^2 h^9} \sum_{j=1}^{n} \sum_{k=1}^{n} [h^4 - (X_j - X_k)^2 h^2 + \frac{1}{12}(X_j - X_k)^4] e^{-(X_j - X_k)^2/4h^2}. \tag{4.2.5}$$

The iteration is terminated when the difference between two adjacent h_i values is less than .05 . We shall call the estimate produced by this algorithm the kernel estimate.

The orthogonal series estimate of Schwartz is given by

$$\hat{f}(x) = \sum_{i=1}^{q(n)} \hat{a}_i \phi_i(x) \tag{4.2.6}$$

where

$$\hat{a}_i = \frac{1}{n} \sum_{k=1}^{n} \phi_i(X_k). \tag{4.2.7}$$

The smoothing parameter in this case is $q(n)$. Schwartz does not completely specify the choice of $q(n)$ although he suggests that it takes the form

$$q(n) = \alpha n^r. \tag{4.2.8}$$

In the Monte Carlo study we shall use the formula

$$q(n) = 1.8n^{\frac{1}{2}}. \tag{4.2.9}$$

The choices for α and r were made after experimentally seeing what "worked well" for the symmetric triangular density. We shall call this estimate the Hermetian estimate.

4.3 Description of the Monte Carlo Study

The densities used in the Monte Carlo study are as follows

1. Exponential $$f(x) = \begin{cases} e^{-x}, & x \geq 0 \\ 0, & \text{otherwise} \end{cases}$$

2. Logistic $$f(x) = \frac{e^x}{(e^x + 1)^2}$$

3. Cauchy $$f(x) = \frac{1}{\pi(1 + x^2)}$$

4. Laplace $$f(x) = \frac{1}{2} e^{-|x|}$$

5. Skew Triangular $$f(x) = \begin{cases} \frac{2}{3} + \frac{2}{3}x, & -1 \leq x < 0 \\ \frac{2}{3} - \frac{1}{3}x, & 0 \leq x < 2 \\ 0, & \text{otherwise.} \end{cases}$$

The sample values for each density were obtained by generating uniform pseudorandom numbers and then mapping to the appropriate density via the inverse probability integral transformation. The function used for this transformation is F^{-1}, the inverse of the corresponding c.d.f. . The sample sizes used were 25, 50, 100 and 200. For each these sample sizes and for each density 25 different samples were obtained and "fed" to each of the five density estimation procedures.

Because the integrated squared error,

$$\int_{-\infty}^{\infty} [f(x) - \hat{f}(x)]^2 \, dx \,, \tag{4.3.1}$$

is usually difficult to calculate,

$$\frac{1}{n} \sum_{i=1}^{n} [f(x_i) - \hat{f}(x_i)]^2 \,, \tag{4.3.2}$$

was used as an approximation. The estimates of the M.I.S.E. which we use is the median of the 25 approximate integrated squared error values. We shall call this quantity the median squared error of the estimate.

To compare each of the types of estimates for any fixed density and sample size a useful statistic is the relative efficiency. The estimate with the lowest median squared error is assigned a relative efficiency of 100% and the relative efficiency of each of the remaining estimates is the ratio of the lowest median square error of the particular estimate. This ratio will always be expressed as a percentage.

4.4 Analysis of Results

The complete set of median squared error values obtained from the Monte Carlo study is given in Appendix A. The relative efficiencies are tabulated in Tables 4.4.1, 4.4.2, 4.4.3 and 4.4.4 and will be used to make comparisons among each of the four types of density estimate.

For the exponential density we see that the histogram gives by far the best results. This is due to the exponential p.d.f. having a relatively large jump discontinuity at $x = 0$. The other three estimation procedures return a continuous function and there is usually a significant error about this value of x because we have a continuous curve trying to fit a discontinuity. The histogram, on the other hand, does quite well in estimating the true density near its mode and returns a perfect estimate for negative values of x.

The relative efficiencies for the logistic density see the histogram estimate falling behind. The two-stage estimate and the kernel estimate both give good estimates although the kernel estimate seems to be at its best when the sample size is large. The logistic p.d.f. is very similar in shape to the standard normal p.d.f. given in 4.2.2. and each of these estimates involve this function in some form. As can be seen in Table 4.4.3, the Hermetian estimate performs terribly for

Table 4.4.1: Relative Efficiencies of the Two-Stage Estimate.

Density	Sample size 25	50	100	200
Exponential	34	25	20	18
Logistic	100	100	100	95
Cauchy	100	100	100	100
Laplace	100	100	100	100
Triangular	68	92	100	100

Table 4.4.2: Relative Efficiencies of the Kernel Estimate.

Density	Sample size 25	50	100	200
Exponential	29	21	21	24
Logistic	70	45	56	100
Cauchy	39	61	45	90
Laplace	49	40	25	26
Triangular	42	93	77	91

Table 4.4.3: Relative Efficiencies of the Hermetian Estimate.

Density	Sample size 25	50	100	200
Exponential	51	38	30	26
Logistic	9	7	5	5
Cauchy	27	39	25	34
Laplace	48	43	24	22
Triangular	100	100	97	49

Table 4.4.4: Relative Efficiencies of the Histogram Estimate.

Density	Sample size 25	50	100	200
Exponential	100	100	100	100
Logistic	36	40	51	55
Cauchy	57	32	6	6
Laplace	51	69	39	26
Triangular	14	38	50	64

the logistic density. This is due to the high sensitivity of the smoothing parameter of the estimate. Many more than the optimal number of terms have been included causing the estimate to behave very badly (see Figure B.3, Appendix B) .

In the case of the Cauchy density, the two-stage estimate seems to do the best with the kernel estimate coming next but performing poorly for small sample sizes. It is the histogram estimate which behaves very badly for this density. This can be attributed to the fact that the extreme order statistics of the Cauchy density have infinite expected value and so typically we have $X_{(1)}$ realising a very large negative value and $X_{(n)}$ a very large positive value. Because these realisations are usually quite different in absolute value a highly distorted estimate is produced. This problem only worsens if the sample size is increased.

The two-stage estimate is shown to be by far the best estimate for the Laplace density. The problem with the kernel estimate seems to be that the algorithm for choosing a close to optimal window width fails for the Laplace. Consequently the resultant curve is usually too flat and does not properly fit the cusp of the Laplace p.d.f. (see Figure B.6, Appendix B) .

Finally we consider the results for the skew triangular density. The two-stage estimate is seen to be slightly ahead of the kernel estimate for this density. The results for the Hermetian estimate are also quite good and this is probably due to the number of terms in the Hermite expansion being closer to the optimal. This happens because, as mentioned in Section 4.2 , the value of $q(n)$ was chosen as one which is close to the optimal value for the symmetric triangular density and this is closely related to the skew triangular density. As in the case of the other continuous densities, the histogram estimate does not do well against the kernel and two-stage estimates because of its inefficient use of the data.

We shall now consider the performance of each of the estimates as the sample size, n , increases. One facet of a probability density function estimate which receives much attention in the literature is the estimate's rate of convergence. We say that an estimate has a rate of

Table 4.4.5: Median Squared Error Values for the Logistic Density.

Type of Estimate	Sample size 25	50	100	200
Two-stage	0.00297	0.00123	0.00076	0.00057
Kernel	0.00424	0.00273	0.00135	0.00054
Hermetian	0.03183	0.01675	0.01512	0.01132
Histogram	0.00829	0.00305	0.00149	0.00098

convergence of order $c(n)$ if

$$\text{MISE}(n) = O(c(n)) \text{ as } n \to \infty, \tag{4.4.1}$$

where MISE(n) is the M.I.S.E. when the sample size is equal to n. Usually $c(n)$ is of the form

$$c(n) = n^{\alpha}. \tag{4.4.2}$$

A rough idea of the true rate of convergence of an estimate can be given by considering the median squared error values for varying values of n. In Table 4.4.5 we give these values for the logistic density.

The Hermetian estimate for the logistic density is so much worse than the other estimates that we shall not concern ourselves with its rate of convergence. The best rates of convergence seem to be exhibited by the kernel and histogram estimates as there is considerable improvement in the estimate as n increases, especially for n increasing from 100 to 200. However, for low sample sizes ($n \leq 100$) the two-stage estimate is about twice as good as the kernel and histogram estimates but does not seem to improve very much when the sample size is increased to 200 and in fact is "overtaken" by the kernel estimate.

4.5 Final Comments

The two-stage estimate must be accepted as a good probability density function estimate in light of the fact that it "beats" its only serious competitor, the kernel estimate of Scott, Tapia and Thompson.

One problem with the two-stage estimate is its failure to improve very much for larger sample sizes from its good small sample performance. It is hoped that a more careful (and more time consuming) choice of some of the parameters involved in the preliminary estimate would increase the rate of convergence of the estimate.

Another problem with the two-stage estimate is its poor performance for densities with a discontinuous p.d.f.. The performance for such a density could be improved if more terms are included in the Hermite series fit. Too few terms are fitted for a discontinuous p.d.f. with the stopping rule given in Chapter 3. If a more versatile stopping rule could be formulated then the estimates for p.d.f.'s such as the exponential and uniform would probably improve. Unfortunately, due to time restrictions the stopping rule had to be developed as a somewhat simplistic one.

APPENDIX A: TABLES OF MONTE CARLO RESULTS

Table A.1: Median Squared Error Values of the Two-stage Estimate.

	Sample size			
Density	25	50	100	200
Exponential	0.09593	0.07097	0.06672	0.06112
Logistic	0.00297	0.00123	0.00076	0.00057
Cauchy	0.00389	0.00368	0.00163	0.00173
Laplace	0.00916	0.00444	0.00164	0.00115
Triangular	0.01737	0.01563	0.00629	0.00407

Table A.2: Median Squared Error Values of the Kernel Estimate.

	Sample size			
Density	25	50	100	200
Exponential	0.11212	0.08584	0.06354	0.04437
Logistic	0.00424	0.00273	0.00135	0.00054
Cauchy	0.00991	0.00603	0.00364	0.00192
Laplace	0.01879	0.01103	0.00646	0.00435
Triangular	0.02827	0.01552	0.00818	0.00445

Table A.3: Median Squared Error Values of the Hermetian Estimate.

	Sample size			
Density	25	50	100	200
Exponential	0.06449	0.04701	0.04493	0.04213
Logistic	0.03183	0.01675	0.01512	0.01132
Cauchy	0.01434	0.00932	0.00653	0.00512
Laplace	0.01923	0.01028	0.00684	0.00525
Triangular	0.01178	0.01441	0.00647	0.00832

Table A.4: Median Squared Error Values of the Histogram Estimate.

	Sample size			
Density	25	50	100	200
Exponential	0.03306	0.01805	0.01341	0.01085
Logistic	0.00829	0.00305	0.00149	0.00098
Cauchy	0.00681	0.01140	0.02654	0.03040
Laplace	0.01809	0.00640	0.00425	0.00434
Triangular	0.08684	0.03802	0.01269	0.00634

APPENDIX B: GRAPHICAL RESULTS

The following pages contain graphs which, for selected samples, show the true p.d.f. (dotted line) and a particular estimate of the sample (solid line). The samples used are of 100 observations from both the logistic and Laplace densities. For each graph the sample was specially selected so that the squared error was close to the median squared error of the estimate for the particular sample size and density. Thus, each graph is supposed to depict the "average case" behaviour of the estimate.

Figure B.1: Two-stage Estimate - Logistic

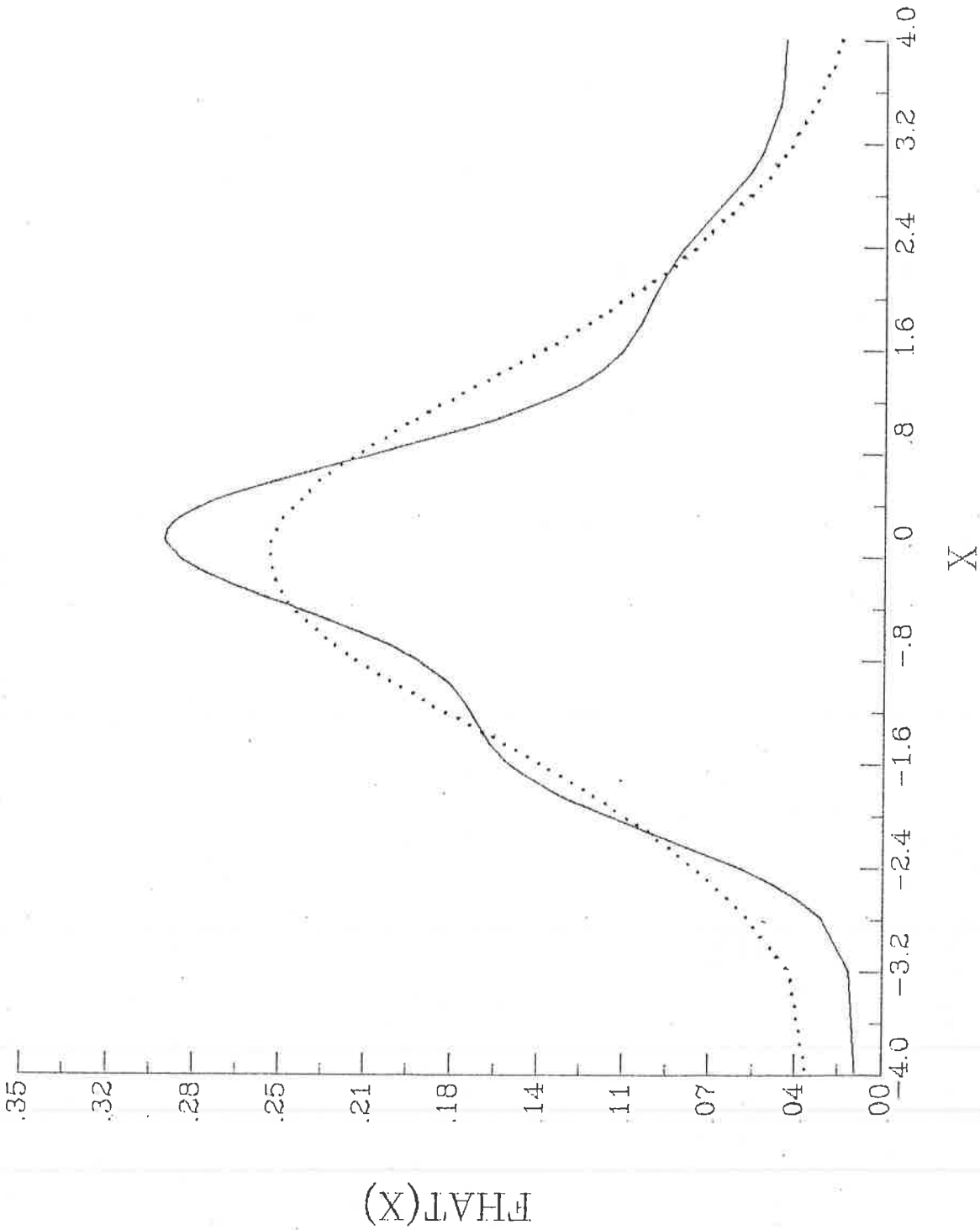


Figure B.2: Kernel Estimate - Logistic

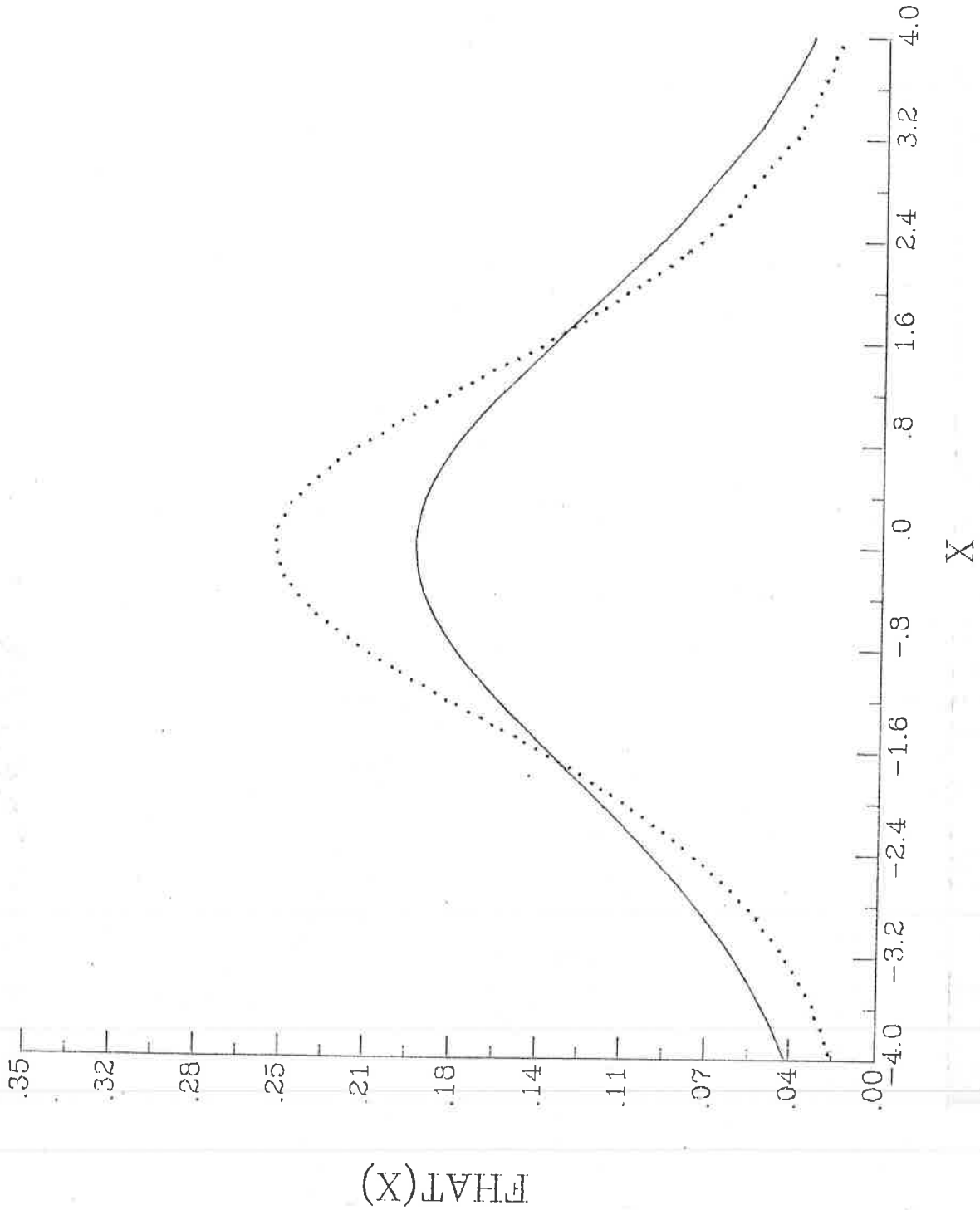


Figure B.3: Hermetian Estimate - Logistic

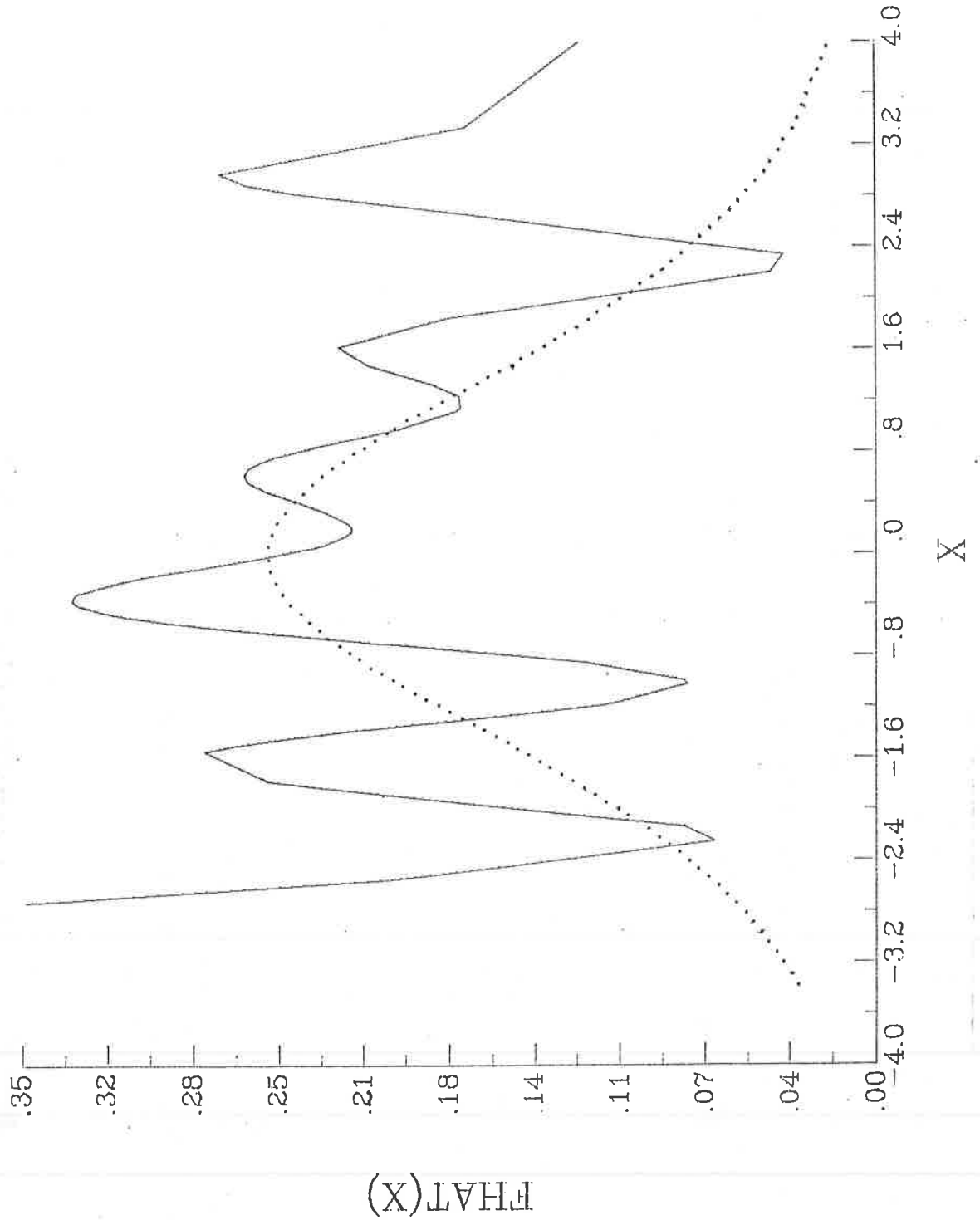


Figure B.4: Histogram Estimate - Logistic

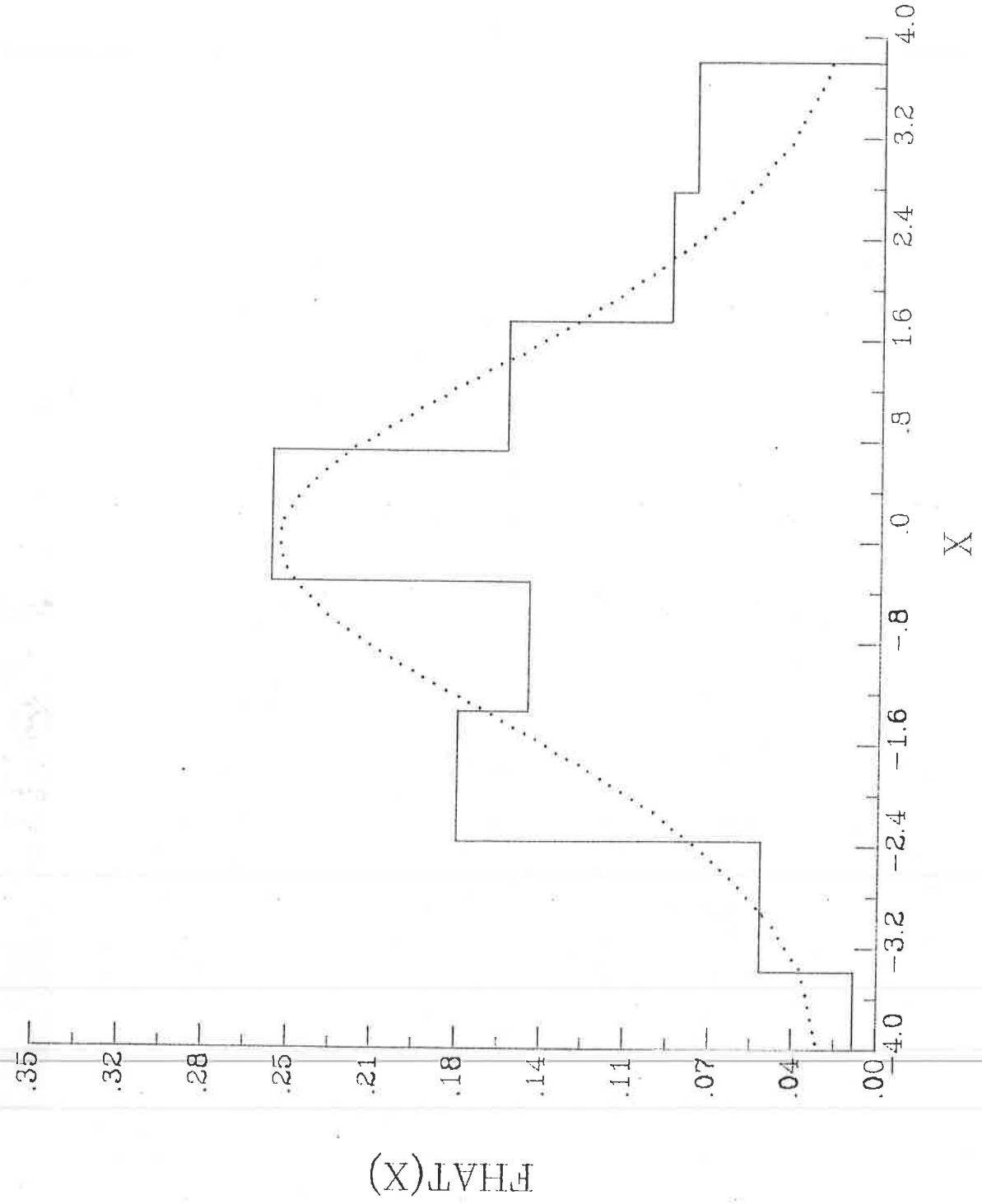


Figure B.5: Two-stage Estimate - Laplace

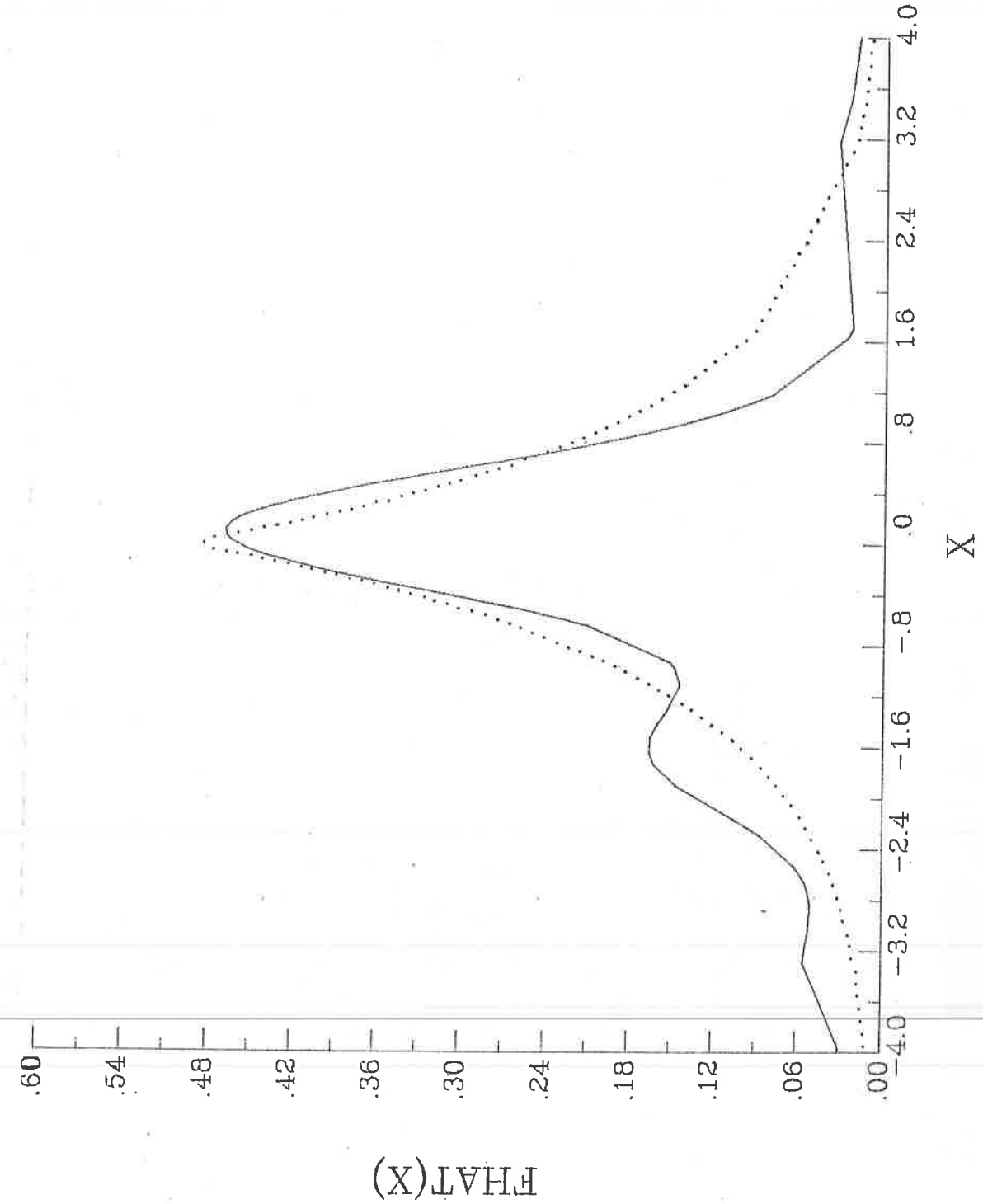


Figure B.6: Kernel Estimate - Laplace

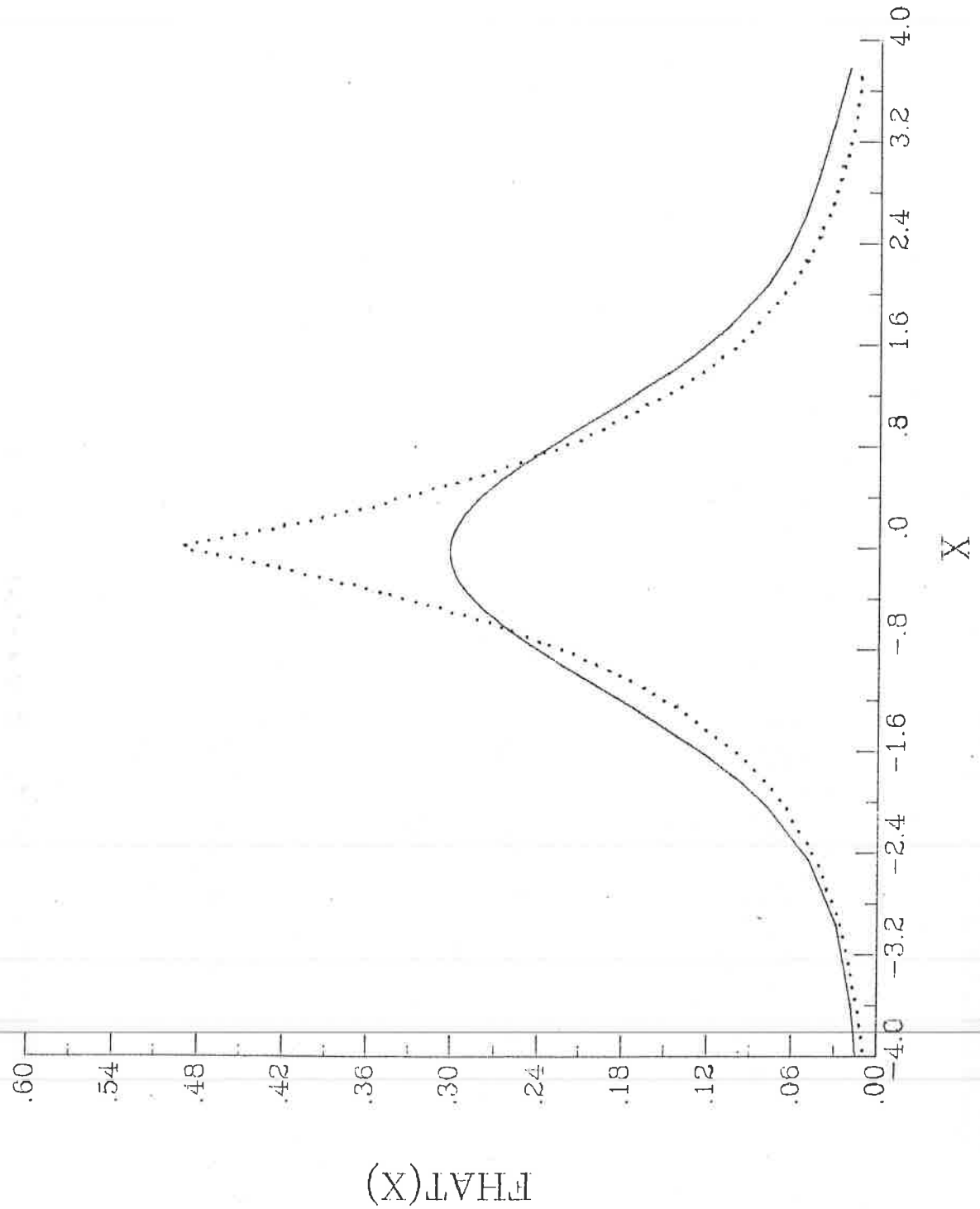


Figure B.7: Hermetian Estimate - Laplace

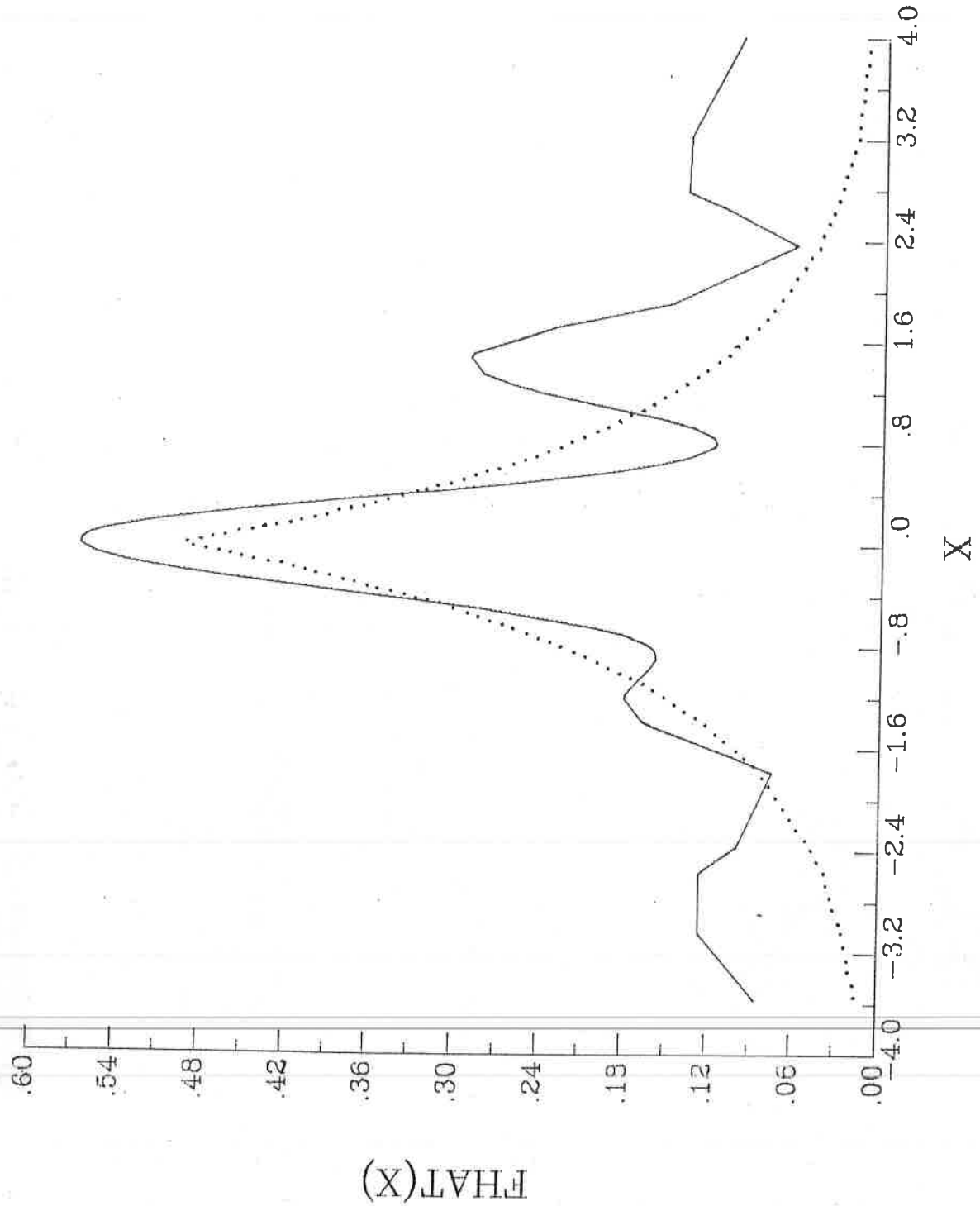
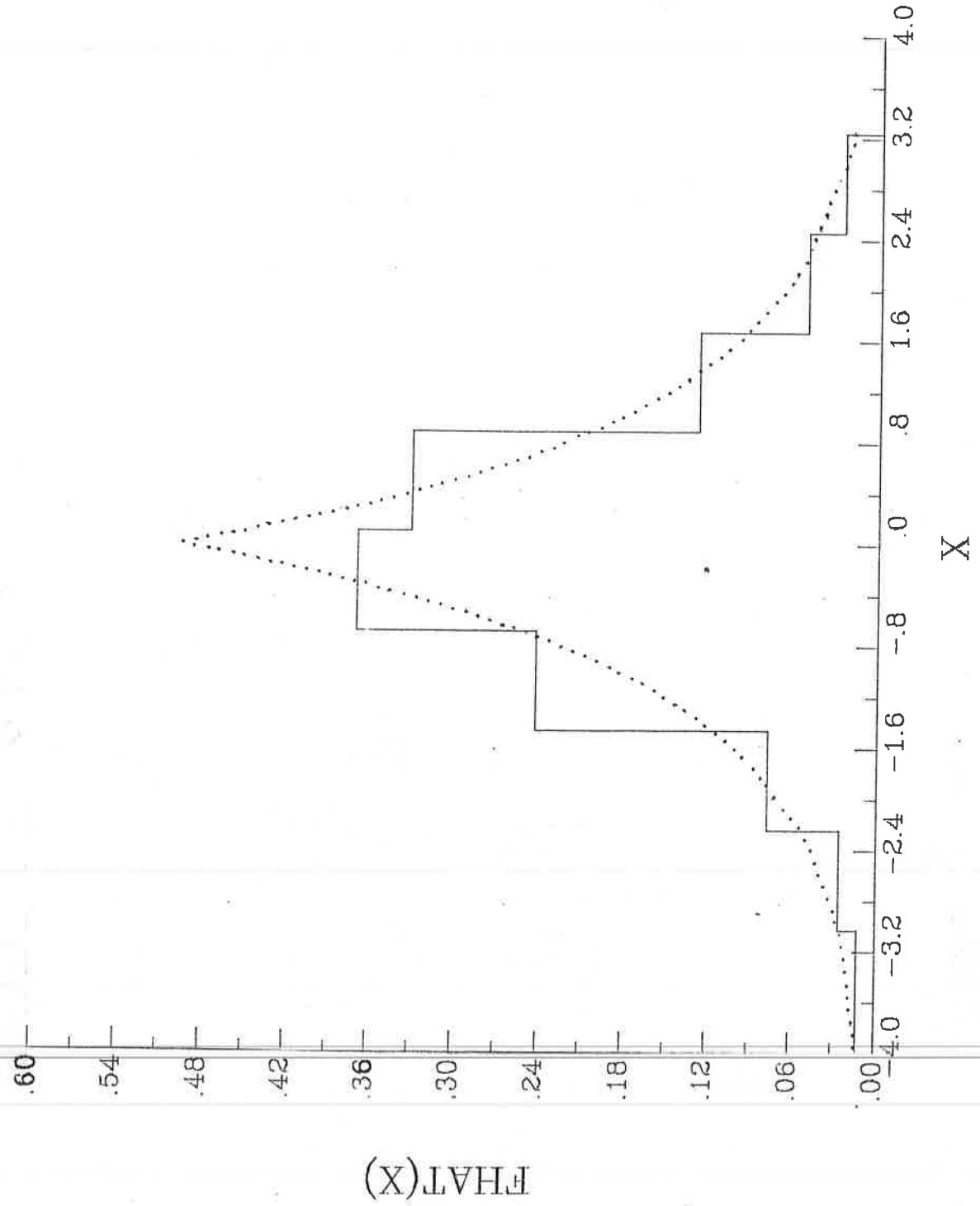


Figure B.8: Histogram Estimate - Laplace



References

Bean, S.J. and Tsokos, C.P. (1980). Developments in Nonparametric Density Estimation. *International Statistical Review*, **48**, 267-287.

Davis, K.B. (1977). Mean Integrated Square Error Properties of Density Estimates. *Annals of Statistics*, **5**, 530-535.

de Montricher, G.F., Tapia, R.A. and Thompson, J.R. (1975). Nonparametric Maximum Likelihood Estimation of Probability Densities by Penalty Function Methods. *Annals of Statistics*, **3**, 1329-1348.

Good, I.J. and Gaskins, R.A. (1971). Nonparametric Roughness Penalties for Probability Densities. *Biometrika*, **58**, 255-277.

Kronmal, R. and Tarter, M. (1968). The Estimation of Probability Densities and Cumulatives by Fourier Series Methods. *Journal of the American Statistical Association*, **63**, 925-952.

Parzen, E. (1962). On the Estimation of a Probability Density Function and the Mode. *Annals of Mathematical Statistics*, **40**, 1065-1076.

Read, P.R. (1972). The Asymptotic Inadmissibility of the Sample Distribution Function. *Annals of Mathematical Statistics*, **43**, 89-95.

Rohatgi, V.K. (1976). *An Introduction to Probability Theory and Mathematical Statistics*. Wiley.

Rosenblatt, M. (1956). Remarks on Some Nonparametric Estimates of a Density Function. *Annals of Mathematical Statistics*, **27**, 832-837.

Schwartz, S.C. (1967). Estimation of a Probability Density by an Orthogonal Series. *Annals of Mathematical Statistics*, **38**, 1261-1265.

Scott, D.W. and Factor, L.E. (1981). Monte Carlo Study of Three Data-Based Nonparametric Probability Density Estimators. *Journal of the American Statistical Association*, **76**, 9-15.

Scott, D.W., Tapia, R.A. and Thompson, J.R. (1977). Kernel Density Estimation Revisited. *Journal of Nonlinear Analysis, Theory, Methods and Applications*, **1**, 339-372.

Scott, D.W., Tapia, R.A. and Thompson, J.R. (1980). Nonparametric Probability Density Estimation by Discrete Maximum Penalised-Likelihood Criteria. *Annals of Statistics*, **8**, 820-832.

Specht, D.F. (1971). Series Estimation of a Probability Density Function. *Technometrics*, **13**, 409-424.

Tapia, R.A. and Thompson, J.R. (1978). *Nonparametric Probability Density Estimation*. John Hopkins University Press, Baltimore, Maryland.

Wegman, E.J. (1972). Nonparametric Density Estimation: II. A Comparison of Density Estimation Methods. *Journal of Statistical Computation and Simulation*, **1**, 225-245.