AN ABSTRACT OF THE THESIS OF

BRADFORD ROBERT CRAIN (Name) for the DOCTOR OF PHILOSOPHY (Degree)

in STATISTICS (Major) presented on April 14, 1972 (Date)

Title: NONPARAMETRIC ESTIMATION OF DISTRIBUTIONS USING ORTHOGONAL EXPANSIONS

Redacted for Privacy

Abstract approved: _____
Dr. H.D. Brunk

Methods of approximation and estimation of the density and the cumulative distribution function of a distribution over a finite interval are investigated. Goodness of the methods is measured pointwise and in terms of mean integrated squared error (MISE).

If the density obeys certain regulatory conditions, i.e., continuous, positive, piecewise smooth, then the canonical exponential family of distributions serve very satisfactorily as an approximation to the true distribution. The theory has wide application to real world problems since the assumptions made are very general.

The class of matrix estimators of the density is introduced and is shown to be superior to the usual orthogonal series estimator when comparison is in terms of MISE. Large sample equivalence of the two methods is established.

Nonparametric Estimation of Distributions Using Orthogonal Expansions

by

Bradford Robert Crain

A THESIS

submitted to

Oregon State University

in partial fulfillment of
the requirements for the
degree of

Doctor of Philosophy

June 1972

APPROVED:

Redacted for Privacy

Professor of Statistics

in charge of major

Redacted for Privacy

Acting Chairman of Department of Statistics

Redacted for Privacy

Dean of Graduate School

Date thesis is presented April 11, 1972

Typed by Clover Redfern for Bradford Robert Crain

ACKNOWLEDGMENT

The author would like to express his appreciation to the many members of the Faculty of Oregon State University, with whom he has had course work or informal discussion, and to the students as well, for the plethora of valuable experience they have provided, as well as their general attitude of good will. The author has benefitted in particular from the knowledge and advice of Dr. John Lee, Dr. Lyle Calvin and Dr. Fred Ramsey.

Without the help of Dr. H.D. Brunk this thesis could not have been written. The thesis problem itself was suggested to the author by Dr. Brunk, and he has given abundantly of his time, guidance and encouragement through the period of thesis work; the author considers himself fortunate to have worked with such a person.

The author recognizes the financial assistance of the Department of Statistics and Computer Center of the university, and the support of the Public Health Service, which allowed him to complete his educational program.

Final thanks go to his wife and exuberant children.

TABLE OF CONTENTS

NONPARAMETRIC ESTIMATION OF DISTRIBUTIONS USING ORTHOGONAL EXPANSIONS

I. HISTORY

There has been a considerable amount of work done in the last ten years in the area of estimation of densities and CDF's. The impetus for this effort has been supplied to a large degree by Rosenblatt (1956), Parzen (1962) and Čencov (1962).

Rosenblatt assumes a continuous univariate density function $f(y)$ and shows that any estimate of $f(y)$ which is a symmetric function of the observations $X_1, X_2, \ldots, X_n$ is of necessity biased. He examines an estimate of $f(y)$ of the form

$$f_n(y) = \frac{F_n(y+h) - F_n(y-h)}{2h}$$

which is the difference quotient of the sample distribution function $F_n(y)$, where $h = h_n$ is a function of the sample size n and approaches zero as $n \to \infty$. The asymptotic behavior of this estimate as $n \to \infty$ is examined in terms of its mean square error, $E|f_n(y) - f(y)|^2$, which he used as a reasonable measure of the local accuracy of $f_n(y)$.

As a global measure of the goodness of f_n Rosenblatt uses the integrated mean square error

$$\int_{-\infty}^{\infty} E\left|f_n(y)-f(y)\right|^2 dy$$

Finally, Rosenblatt generalizes the estimator $f_n(y)$ to a whole class of estimates of the density function of the form

$$f_n(y) = \int_{-\infty}^{\infty} w_n(y-u)dF_n(u) = \frac{1}{n}\sum_{j=1}^{n} w_n(y-X_i)$$

where $w_n(u)$ is a nonnegative function such that

$$\int_{-\infty}^{\infty} w_n(u)du = 1$$

and such that for any $\epsilon > 0$,

$$\lim_{n\to\infty} \int_{|u|<\epsilon} w_n(u)du = 1$$

One particular form of $w_n(u)$, which is useful is

$$w_n(u) = \frac{1}{h} w(\frac{u}{h})$$

where $h = h_n \to 0$ as $n \to \infty$ and where $w(u)$ is a density. A special case is the original estimator

$$f_n(y) = \frac{F_n(y+h)-F_n(y-h)}{2h} = \int_{-\infty}^{\infty} w_n(y-u)dF_n(u)$$

where

$$w(u) = \begin{cases} \frac{1}{2}, & |u| < 1 \\ 0, & |u| \geq 1 \end{cases} .$$

This generalized estimator forms the basis for the definitive paper by Parzen and is extended to the multivariate case by Cacoullos (1964).

Parzen considers the estimator

$$f_n(x) = \int_{-\infty}^{\infty} \frac{1}{h} K(\frac{x-y}{h}) \, dF_n(y)$$

and gives conditions for f_n to be asymptotically unbiased and consistent. He discusses the asymptotic normality of the estimate $f_n(x)$ and goes on to investigate estimation of the mode of f.

In the multivariate case Cacoullos uses an estimator of the form

$$f_n(x) = \frac{1}{n} \frac{1}{h_1 \cdots h_p} \sum_{j=1}^{n} K(\frac{x_1 - X_{j1}}{h_1}, \ldots, \frac{x_p - X_{jp}}{h_p})$$

Here the density of interest is p-variate and $X_j = (X_{j1}, \ldots, X_{jp})$, $j = 1, 2, \ldots, n$ is a random sample. The $h_i = h_i(n) > 0$ satisfy $\lim_{n \to \infty} h_i(n) = 0$, $i = 1, \ldots, p$. Motivated by Parzen's work, Cacoullos gives results concerning the consistency, asymptotic unbiasedness,

and bounds for bias and mean square error of $f_n(x)$.

Most of the earlier work done in estimating densities and CDF's falls into one of two areas, the kernel method and the orthogonal series method. The work of Rosenblatt, Parzen and Cacoullos forms the nucleus of what has been done using the kernel method. Use of orthogonal functions seems to have started with an original paper by Čencov, and this approach has been taken up by Schwartz (1967), Kronmal and Tarter (1968) and Watson (1969).

Čencov assumes a random variable $\tilde{\xi}$, a measure μ in the space X of $\tilde{\xi}$-values, and a density $p(x) = \frac{dP}{d\mu}$ where P is the probability measure induced by $\tilde{\xi}$. Using a weight function $r(x)$ and by means of the inner product

$$(\phi, \psi) = \int_X \phi(x)\psi(x)r(x)\mu(dx)$$

a Hilbert space $L_2(r)$ is defined. He then considers an arbitrary n-dimensional subspace E_n with orthonormal basis $\{\phi_{kn}(x)\}_{k=1}^{n}$. The projection of p onto E_n is the best mean square approximation of p by a member of E_n, and is given by

$$\pi_n(x) = \sum_{k=1}^{n} a_{kn}\phi_{kn}(x) = \sum_{k=1}^{n} (\phi_{kn}, p)\phi_{kn}(x)$$

If $\xi^{(1)}, \ldots, \xi^{(N)}$ are independent observations of the variable $\tilde{\xi}$ then an unbiased estimate of a_{kn} is

$$a_{kn} = \frac{1}{N} \sum_{i=1}^{N} \phi_{kn}(\xi^{(i)}) r(\xi^{(i)}).$$

Čencov uses the estimator

$$p(x) = \sum_{k=1}^{n} a_{kn} \phi_{kn}(x)$$

and derives some of its properties in terms of the L_2-norm.

Schwartz considers an estimate of $f(x)$ of the form

$$f_n(x) = \sum_{j=0}^{q(n)} a_{jn} \varphi_j(x)$$

where

$$a_{jn} = \frac{1}{n} \sum_{i=1}^{n} \varphi_j(X_i),$$

φ_j is the jth Hermite function, $q(n)$ is an integer dependent on n, and $X_1, \ldots, X_n$ is a random sample from $f(x)$. Imposing certain conditions, Schwartz shows that the sequence of estimates is consistent in the sense of mean integrated squared error and also consistent in mean squared error.

Kronmal and Tarter use Fourier series methods to estimate

both the density and its CDF. In their extensive efforts they have developed not only theoretical properties of trigonometric estimators but also have used the criterion of mean integrated squared error in a Monte Carlo comparison of their series estimators with the estimators that were gotten by Watson and Leadbetter, and also with straight parametric estimation, i.e., with a parametric density whose parameters were estimated. They give a truncation rule which depends on the sample.

Watson and Leadbetter (1963) consider estimators of the form

$$f_n(x) = \frac{1}{n} \sum_{i=1}^{n} \delta_n(x-X_i)$$

which is analogous to Parzen's estimator except that they determine, for fixed n, what focusing function δ_n will give the smallest mean integrated squared error. With their method one has to invert a characteristic function and a great deal of information is needed to make the method applicable. Still, their paper is quite ingenious and of theoretical interest.

Watson (1969) assumes that the density can be written as

$$f(x) = \sum_{m=0}^{\infty} a_m \varphi_m(x)$$

where $(\varphi_m(x))$ is an orthonormal basis and considers an estimator

of the form

$$f_n(x) = \sum_{m=0}^{\infty} \lambda_m(n) a_m \varphi_m(x)$$

where a_m is the usual unbiased estimate of α_m, n is the sample size, and the weighting factors are chosen to minimize the mean integrated squared error of f_n.

To mention a few other authors among many, Pickands (1969) investigates efficient estimation of a density function, Van Ryzin (1969) addresses himself to the strong consistency of density estimates, and Bhattacharya has looked at estimation of the derivatives of a density function.

This does not exhaust the work that has been done in the area of density estimation by any means, but it does represent a survey of those results which relate most closely to what has been done here. For an excellent bibliography of density estimation see Wegman (1969).

II. INTRODUCTION

The Model

This investigation assumes that P is a probability measure over the Borel subsets of the closed interval $[-1, 1]$ and that P is absolutely continuous with respect to Legesgue measure λ. The density function $p = \frac{dP}{d\lambda}$ is assumed expressible as

$$p(x) = e^{\Sigma_{i=1}^{\infty} \tau_i \varphi_i(x) - \psi(\tau)} \tag{2.1}$$

where $\varphi_i(x)$ is the ith degree normalized Legendre polynomial over $[-1, 1]$, τ_i is a real coefficient $(i = 1, 2, \ldots)$, and $\psi(\tau)$ is defined by

$$e^{\psi(\tau)} = \int_{-1}^{1} e^{\Sigma_{i=1}^{\infty} \tau_i \varphi_i(x)} \, dx$$

It is tacitly assumed that the series $\Sigma_{i=1}^{\infty} \tau_i \varphi_i(x)$ converges uniformly on $[-1, 1]$, so $p(x)$ is continuous on $[-1, 1]$.

The normalized Legendre polynomials $\{\varphi_i\}$ satisfy

$$\int_{-1}^{1} \varphi_i(x) \varphi_j(x) dx = \delta_{ij} \qquad i, j = 1, 2, 3 \ldots$$

where δ_{ij} is the Kronecker delta, i.e.,

$$\delta_{ij} = \begin{cases} 1, & i = j \\ 0, & i \neq j \end{cases}$$

We further assume that $p(x)$ is piecewise smooth, in the sense that Churchill (1963) uses the term, so that $p(x)$ can also be expanded in an orthogonal series by

$$p(x) = \frac{1}{2} + \sum_{i=1}^{\infty} \theta_i \varphi_i(x)$$

where θ_i is the ith Fourier type coefficient given by

$$\theta_i = \int_{-1}^{1} \varphi_i(x) p(x) dx$$

$$= E[\varphi_i(X)]$$

In the spirit of approximation, define the function $p_m(x|\tau)$ (or $p_m(x)$ for short), which is a function of x and the m variables $\tau_1, \tau_2, \ldots, \tau_m$, by the formula

$$p_m(x|\tau) = e^{\Sigma_{i=1}^{m} \tau_i \varphi_i(x) - \psi_m(\tau)}$$

The normalizing function $\psi_m(\tau)$ is of course defined by

$$e^{\psi_m(\tau)} = \int_{-1}^{1} e^{\Sigma_{i=1}^{m} \tau_i \varphi_i(x)} dx.$$

Definition 2.1. Let $\theta = (\theta_1, \theta_2, \ldots, \theta_m)'$. The function $Q_\theta(\tau)$ is defined by

$$Q_\theta(\tau) = Q_\theta(\tau_1, \tau_2, \ldots, \tau_m)$$

$$= \sum_{i=1}^{m} \tau_i \theta_i - \psi_m(\tau)$$

More generally, for any vector $a \in R^m$ we define $Q_a(\tau)$ by

$$Q_a(\tau) = \sum_{i=1}^{m} \tau_i a_i - \psi_m(\tau).$$

In this chapter we will see that $\psi_m(\tau)$ is a strictly convex function of τ. We will also show that $p(x)$ may be uniformly approximated by a density of the form $p_m(x|\tau)$. Since $p_m(x|\tau)$ is a member of the canonical exponential family $\mathcal{F}_m$ generated by $\{\varphi_1, \varphi_2, \ldots, \varphi_m\}$ and the uniform distribution over $[-1, 1]$, we will see that $\mathcal{F}_m$ is a good approximating class of functions.

In Chapter III the reader will find the analytical results developed or needed in this thesis. Properties of the mappings ψ_m, Q_a, $\nabla\psi_m$ (gradient map), $\phi_m = (\nabla\psi_m)^{-1}$, etc. are established.

Chapter IV discusses the estimation of $p(x)$ and $F(y) = \int_{-1}^{y} p(x)dx$ in terms of the analytical tools developed earlier. In particular, we consider estimates of $p(x)$ of the form

$$p_m(x|\tau_n^*) = e^{\sum_{i=1}^{m} \tau_{ni}^* \varphi_i(x) - \psi_m(\tau_n^*)}, \quad -1 \leq x \leq 1$$

and estimates of $F(x)$ of the form

$$F_{mn}(x) = \int_{-1}^{x} p_m(y|\tau_n^*)dy, \quad -1 \leq x \leq 1$$

To do this we first investigate approximations of $p(x)$ and $F(x)$ of the form

$$\hat{p}(x) = p_m(x|\tau^*) = e^{\sum_{i=1}^{m} \tau_i^* \varphi_i(x) - \psi_m(\tau^*)}$$

and

$$\hat{F}(x) = \int_{-1}^{x} p_m(y|\tau^*)dy$$

In Chapter V, investigation of estimators of $p(x)$ is undertaken where the criterion of goodness of the estimate is mean integrated squared error (MISE). Several results on MISE are given and two new estimators of $p(x)$ are developed using this criterion (the estimators are chosen to yield a small MISE).

We will conclude the introductory chapter by giving some of the basic analytic results and then will pursue the analysis much further in the next chapter.

Using the Lebesgue Dominated Convergence Theorem, we have

by the definition of $\psi_m(\tau)$ that

$$\frac{\partial}{\partial \tau_j} e^{\psi_m(\tau)} = \frac{\partial \psi_m(\tau)}{\partial \tau_j} e^{\psi_m(\tau)}$$

$$= \frac{\partial}{\partial \tau_j} \int_{-1}^{1} e^{\sum_{i=1}^{m} \tau_i \varphi_i(x)} dx$$

$$= \int_{-1}^{1} \varphi_j(x) e^{\sum_{i=1}^{m} \tau_i \varphi_i(x)} dx$$

from which we see that

$$\frac{\partial \psi_m(\tau)}{\partial \tau_j} = \int_{-1}^{1} \varphi_j(x) e^{\sum_{i=1}^{m} \tau_i \varphi_i(x) - \psi_m(\tau)} dx$$

$$= \int_{-1}^{1} \varphi_j(x) p_m(x|\tau) dx$$

$$= E[\varphi_j(x) | p_m]$$

Similarly we can show that

$$\frac{\partial^2 \psi_m(\tau)}{\partial \tau_j \partial \tau_k} = \int_{-1}^{1} [\varphi_j(x) - E[\varphi_j(x) | p_m]][\varphi_k(x) - E[\varphi_k(x) | p_m]] p_m(x|\tau) dx$$

$$= \mathrm{Cov}[\varphi_j(x), \varphi_k(x) | p_m]$$

Let $H_{\psi_m}(\tau)$ be the Hessian matrix of ψ_m, i.e., the $m \times m$ matrix whose jk-th element is $\frac{\partial^2 \psi_m(\tau)}{\partial \tau_j \partial \tau_k}$. It is well known that if

$\{\varphi_1, \varphi_2, \ldots, \varphi_m\}$ are linearly independent —λ, then $H_{\psi_m}(\tau)$ is positive definite. We give a short proof:

Lemma 2.1. $H_{\psi_m}(\tau)$ is positive definite for any $\tau \in R^m$.

Proof. Suppose that $H_{\psi_m}(\tau)$ is not positive definite for some $\tau \in R^m$. Then there exists a non-zero vector $\lambda \in R^m$ such that

$$\lambda' H_{\psi_m}(\tau)\lambda = 0$$

But

$$\lambda' H_{\psi_m}(\tau)\lambda = \mathrm{Var}\Big[\sum_{i=1}^{m} \lambda_i \varphi_i(x) \,\Big|\, p_m\Big]$$

$$\therefore \quad 0 = \int_{-1}^{1} \Big(\sum_{i=1}^{m} \lambda_i \varphi_i(x) - E\Big[\sum_{i=1}^{m} \lambda_i \varphi_i(x) \,\Big|\, p_m\Big] \Big)^2 p_m(x)\,dx$$

This is impossible since the integrand is > 0 a.e. (Lebesgue measure). Hence $H_{\psi_m}(\tau)$ must be positive definite.

The density $p(x)$ has been expressed in two different ways, both of which involve an infinite number of terms. We now address ourselves to an approximation of $p(x)$ whose expansion involves only a finite number of terms:

An Approximation Lemma

(Approximation)Lemma 2.2. Let $p(x) = e^{\sum_{i=1}^{\infty} \tau_i \varphi_i(x) - \psi(\tau)}$, where the series converges uniformly on $[-1, 1]$. Then the following hold:

(i) $e^{\sum_{i=1}^{m} \tau_i \varphi_i(x)} \to e^{\sum_{i=1}^{\infty} \tau_i \varphi_i(x)}$ as $m \to \infty$, uniformly on $[-1,1]$

(ii) $e^{\psi_m(\tau)} \to e^{\psi(\tau)}$ as $m \to \infty$ i.e., $\psi_m(\tau) \to \psi(\tau)$

(iii) $e^{\sum_{i=1}^{m} \tau_i \varphi_i(x) - \psi_m(\tau)} \to e^{\sum_{i=1}^{\infty} \tau_i \varphi_i(x) - \psi(\tau)}$ as $m \to \infty$, uniformly on $[-1, 1]$

(iv) $\int_{-1}^{y} e^{\sum_{i=1}^{m} \tau_i \varphi_i(x) - \psi_m(\tau)} dx \to \int_{-1}^{y} e^{\sum_{i=1}^{\infty} \tau_i \varphi_i(x) - \psi(\tau)} dx$ as $m \to \infty$, uniformly in y

(v) $\int_{-1}^{1} \varphi_j(x) e^{\sum_{i=1}^{m} \tau_i \varphi_i(x) - \psi_m(\tau)} dx \to \int_{-1}^{1} \varphi_j(x) e^{\sum_{i=1}^{\infty} \tau_i \varphi_i(x) - \psi(\tau)} dx$ as $m \to \infty$, uniformly in j.

Proof. By uniform convergence the function

$$\phi(x) = \sum_{i=1}^{\infty} \tau_i \varphi_i(x)$$

is continuous, hence uniformly continuous on $[-1, 1]$. Clearly

$\phi([-1, 1])$ is compact. Let $\rho > 0$ and define

$$L_1 = \left\{ \inf_{x \in [-1, 1]} \phi(x) \right\} - \rho$$

$$L_2 = \left\{ \sup_{x \in [-1, 1]} \phi(x) \right\} + \rho$$

By uniform convergence there exists an integer $M(\rho)$ such that $m \geq M$ implies

$$\sum_{i=1}^{m} \tau_i \varphi_i(x) \in [L_1, L_2] \quad , \quad x \in [-1, 1].$$

The function $g(y) = e^y$ is uniformly continuous on $[L_1, L_2]$. Let $\epsilon > 0$. Then there exists a $\delta > 0$ such that $x, y \in [L_1, L_2]$, $|x-y| < \delta$ implies $|g(x)-g(y)| < \epsilon$. There exists $M_1 > M$ such that for any $m \geq M_1$ and for any $x \in [-1, 1]$,

$$\left| \sum_{i=1}^{\infty} \tau_i \varphi_i(x) - \sum_{i=1}^{m} \tau_i \varphi_i(x) \right| < \delta$$

which implies

$$\left| g\left(\sum_{i=1}^{\infty} \tau_i \varphi_i(x) \right) - g\left(\sum_{i=1}^{m} \tau_i \varphi_i(x) \right) \right| < \epsilon$$

This provies (i). (ii) follows from (i) since

$$\left| \int_{-1}^{1} e^{\sum_{i=1}^{m} \tau_i \varphi_i(x)} dx - \int_{-1}^{1} e^{\sum_{i=1}^{\infty} \tau_i \varphi_i(x)} dx \right|$$

$$\leq \int_{-1}^{1} \left| e^{\sum_{i=1}^{m} \tau_i \varphi_i(x)} - e^{\sum_{i=1}^{\infty} \tau_i \varphi_i(x)} \right| dx$$

$$< 2\epsilon, \quad m > M_1.$$

To get (iii) note that for ϵ_1 such that

$$0 < \epsilon_1 < \int_{-1}^{1} e^{\sum_{i=1}^{\infty} \tau_i \varphi_i(x)} dx = e^{\psi(\tau)}$$

there exists $M_1(\epsilon_1)$ such that $m \geq M_1$ implies $\left| e^{\psi_m(\tau)} - e^{\psi(\tau)} \right| < \epsilon_1$. Then

$$-\epsilon_1 + e^{\psi(\tau)} < e^{\psi_m(\tau)} < e^{\psi(\tau)} + \epsilon_1$$

and

$$\frac{1}{e^{\psi(\tau)} + \epsilon_1} < \frac{1}{e^{\psi_m(\tau)}} < \frac{1}{e^{\psi(\tau)} - \epsilon_1}$$

Let $\epsilon_2 > 0$. There exists $M_2(\epsilon_2) > M_1$ such that $m \geq M_2$ implies

$$\left| e^{\sum_{i=1}^{m} \tau_i \varphi_i(x)} - e^{\sum_{i=1}^{\infty} \tau_i \varphi_i(x)} \right| < \epsilon_2$$

for any $x \in [-1, 1]$, or what is the same thing,

$$e^{\sum_{i=1}^{\infty} \tau_i \varphi_i(x)} - \epsilon_2 < e^{\sum_{i=1}^{m} \tau_i \varphi_i(x)} < e^{\sum_{i=1}^{\infty} \tau_i \varphi_i(x)} + \epsilon_2, \; x \in [-1,1].$$

Then multiplying through, we have

$$\frac{e^{\sum_{i=1}^{\infty}\tau_i\varphi_i(x)-\epsilon_2}}{e^{\psi(\tau)+\epsilon_1}} \leq \frac{e^{\sum_{i=1}^{m}\tau_i\varphi_i(x)}}{e^{\psi_m(\tau)}} \leq \frac{e^{\sum_{i=1}^{\infty}\tau_i\varphi_i(x)+\epsilon_2}}{e^{\psi(\tau)-\epsilon_1}} \qquad m \geq M_2$$

Define ϵ_3 and ϵ_4 by

$$\epsilon_1 = \epsilon_3 e^{\psi(\tau)}$$

$$\epsilon_2 = \epsilon_4 e^{\psi(\tau)}$$

Then we have

$$\frac{(1+\epsilon_3)e^{\sum_{i=1}^{\infty}\tau_i\varphi_i(x)} - \epsilon_3 e^{\sum_{i=1}^{\infty}\tau_i\varphi_i(x)}}{(1+\epsilon_3)e^{\psi(\tau)}} - \frac{\epsilon_4 e^{\psi(\tau)}}{(1+\epsilon_3)e^{\psi(\tau)}}$$

$$\leq \frac{e^{\sum_{i=1}^{m}\tau_i\varphi_i(x)}}{e^{\psi_m(\tau)}}$$

$$\leq \frac{(1-\epsilon_3)e^{\sum_{i=1}^{\infty}\tau_i\varphi_i(x)} + \epsilon_3 e^{\sum_{i=1}^{\infty}\tau_i\varphi_i(x)}}{(1-\epsilon_3)e^{\psi(\tau)}} + \frac{\epsilon_4 e^{\psi(\tau)}}{(1-\epsilon_3)e^{\psi(\tau)}}$$

Recalling the exponential form of $p(x)$, we have

$$p(x) - \frac{\epsilon_3}{1+\epsilon_3}p(x) - \frac{\epsilon_4}{1+\epsilon_3} \leq p_m(x) \leq p(x) + \frac{\epsilon_3}{1-\epsilon_3}p(x) + \frac{\epsilon_4}{1-\epsilon_3}$$

Now let $\rho^* = \max_{-1 \leq x \leq 1} p(x)$. Then we have

$$|p_m(x) - p(x)| \leq \max\left\{\left(\frac{\epsilon_3\rho^*}{1-\epsilon_3} + \frac{\epsilon_4}{1-\epsilon_3}\right), \left(\frac{\epsilon_3\rho^*}{1+\epsilon_3} + \frac{\epsilon_4}{1+\epsilon_3}\right)\right\}$$

$$|p_m(x) - p(x)| \leq \frac{\epsilon_3\rho^*}{1-\epsilon_3} + \frac{\epsilon_4}{1-\epsilon_3}, \quad m \geq M_2, \quad -1 \leq x \leq 1$$

Since this bound can be made arbitrarily small, (iii) is proven.

Part (iv) follows directly from (iii). To get (v) note that

$$\left|\int_{-1}^{1} \varphi_j(x)p_m(x)dx - \int_{-1}^{1} \varphi_j(x)p(x)dx\right| \leq \int_{-1}^{1} |\varphi_j(x)|\,|p_m(x) - p(x)|\,dx$$

$$\leq \left\{\int_{-1}^{1} \varphi_j^2(x)dx\right\}^{1/2} \left\{\int_{-1}^{1} (p_m(x)-p(x))^2 dx\right\}^{1/2}$$

$$\leq 1\cdot\sqrt{2\epsilon^2} = \epsilon\sqrt{2}$$

for $m > M(\epsilon)$. This bound is good for all indices j.

Remark. There are several reasons why the statistician might look to the class of estimators of $p(x)$ which have the form $p_m(x)$ for some m and some values $\tau_1, \tau_2, \ldots, \tau_m$. The density $p(x)$ has (possibly) a countable infinity of parameters and that is a lot of parameters to estimate. If an estimate of the form $p_m(x)$ would suffice then so much the better. Also, $p(x)$ is by assumption continuous and strictly positive on its support. The class of

p_m-type functions also has these characteristics. It has the property that for m not too large, the p_m are very "smooth," that is, if one is interested in "smooth" estimates then this class offers that prerogative. Finally, there is a world of difference in working with functions like $p_m(x)$ than working with functions that have the structure that $p(x)$ has. This last remark refers to the mathematics involved.

A final remark is made in the way of an apology. The symbols $\tau_1, \tau_2, \ldots$ and $\theta_1, \theta_2, \ldots$ are sometimes used as parameters and sometimes as real variables. It is hoped that the intention will usually be clear by context. Also, vectors will not be especially notated. Consequently an expression like τ may mean $(\tau_1, \tau_2, \ldots, \tau_m) \in R^m$ or $(\tau_1, \tau_2, \ldots, \tau_n, \ldots) \in R^\infty$.

III. AN OPTIMIZATION PROBLEM AND PROPERTIES OF THE MAPPING $\psi_m(\tau)$

In this chapter we deal mainly with various properties of the function $\psi_m(\tau)$. Most of the results given are already known; most of the proofs contained herein are thought by the author to be original. Although most of the results can be found as statements or theorems in the references, proofs are considerably harder to come by. The proofs we give are, in most cases, simple and direct, and are essentially self-contained. More importantly, however, the proofs we give here allow us to make some extensions which seem to not be in the literature. Therefore we call the reader's attention to the following: the author considers most proofs to be his own; Lemmas 3.9, 3.11, 3.13, 3.14 and Theorems 3.1, 3.2 are thought to be new; all other results can at least be found as statements in Rockafellar (1970), Zangwill (1969), or Barndorff-Nielsen (1970).

The Mapping ψ_m

Lemma 3.1. The mapping $\psi_m: R^m \to R^1$ is strictly convex, that is, for any two distinct vectors τ_1 and τ_2 in R^m and $\lambda \in (0, 1)$ we have

$$\psi_m(\lambda\tau_1+(1-\lambda)\tau_2) < \lambda\psi_m(\tau_1) + (1-\lambda)\psi_m(\tau_2)$$

Proof. The Hessian matrix for $\psi_m(\tau)$, $H_{\psi_m}(\tau)$, is positive definite for any $\tau \in R^m$. This is sufficient for strict convexity (Zangwill, 1969).

Lemma 3.2. For any vector $a \in R^m$, the function $Q_a(\tau) = \sum_{i=1}^{m} \tau_i a_i - \psi_m(\tau)$ is strictly concave, i.e., for any two distinct vectors τ_1 and τ_2 in R^m, we have

$$Q_a(\lambda\tau_1+(1-\lambda)\tau_2) > \lambda Q_a(\tau_1) + (1-\lambda)Q_a(\tau_2)$$

whenever $\lambda \in (0, 1)$.

Proof. $H_{Q_a}(\tau) = H_{-\psi_m}(\tau) = -H_{\psi_m}(\tau)$ which is negative definite, therefore Q_a is strictly concave.

The symbol ∇ will stand for gradient. The next lemma is also found in Zangwill (1969):

Lemma 3.3. Let h be a differentiable concave function on R^m. Then $\nabla h(x^*) = 0$ if and only if x^* maximizes h over R^m.

The function $\sum_{i=1}^{m} \tau_i a_i - \psi_m(\tau)$ is a differentiable concave function. Then Lemma 3.3 implies that this function has a maximum at τ^* if and only if $a = \nabla\psi_m(\tau^*)$. This means that the vector a must be in the range of $\nabla\psi_m$ for the function to have a maximum. This we state as a lemma:

Lemma 3.4. The function $Q_a(\tau) = \sum_{i=1}^{m} \tau_i a_i - \psi_m(\tau)$ has a maximum point if and only if $a \in \mathcal{R}_{\nabla\psi_m}$, where $\mathcal{R}_{\nabla\psi_m}$ denotes $\{\nabla\psi_m(\tau): \tau \in R^m\}$, the range of $\nabla\psi_m$.

The next lemma asserts the uniqueness of the point of maximality of a strictly concave function:

Lemma 3.5. Suppose h is strictly concave on R^m and x_1^* and x_2^* both maximize h over R^m. Then $x_1^* = x_2^*$.

Proof. Suppose $x_1^* \neq x_2^*$. We have $h(x_1^*) = h(x_2^*)$. For $0 < \lambda < 1$ we have

$$h(\lambda x_1^* + (1-\lambda)x_2^*) > \lambda h(x_1^*) + (1-\lambda)h(x_2^*) = h(x_1^*).$$

This contradicts the maximality of x_1^*, therefore $x_1^* = x_2^*$.

The function $Q_a(\tau)$ is seen to have a point of maximality if and only if $a \in \mathcal{R}_{\nabla\psi_m}$ and this maximum point is unique.

Suppose now that a random sample $y_1, y_2, \ldots, y_n$ is obtained from the distribution whose density is $p(x)$, where $p(x)$ is given by (2.1). The likelihood function is

$$L(y_1, y_2, \ldots, y_n) = p(y_1)p(y_2)\cdots p(y_n)$$

$$= \prod_{j=1}^{n} p(y_j)$$

$$= \prod_{j=1}^{n} e^{\sum_{i=1}^{\infty} \tau_i \varphi_i(y_j) - \psi(\tau)}$$

$$= e^{\sum_{i=1}^{\infty} \tau_i \sum_{j=1}^{n} \varphi_i(y_j) - n\psi(\tau)}$$

$$= \left\{ e^{\sum_{i=1}^{\infty} \tau_i \bar{\varphi}_i - \psi(\tau)} \right\}^n$$

where

$$\bar{\varphi}_i = \frac{1}{n} \sum_{j=1}^{n} \varphi_i(y_j) \qquad i = 1, 2\ldots \tag{2.2}$$

Define the restricted-likelihood function L_m by

$$L_m(y_1, y_2, \ldots, y_n) = \prod_{j=1}^{n} p_m(y_j)$$

$$= \left\{ e^{\sum_{i=1}^{m} \tau_i \bar{\varphi}_i - \psi_m(\tau)} \right\}^n$$

This would have been the likelihood function had the sample come from p_m instead of p. If we maximize this expression with respect to the τ_i's we would be doing some kind of <u>restricted maximum likelihood estimation</u>. Clearly $\tau_n^* \in R^m$ maximizes L_m if

and only if τ_n^* maximizes $Q_{\bar{\varphi}}(\tau) = \sum_{i=1}^{m} \tau_i \bar{\varphi}_i - \psi_m(\tau)$. As the sample size $n \to \infty$, $\bar{\varphi}_i \to \theta_i$ almost surely and then

$$Q_{\bar{\varphi}}(\tau) \to Q_\theta(\tau) = \sum_{i=1}^{m} \tau_i \theta_i - \psi_m(\tau)$$

almost surely as $n \to \infty$. In this chapter it will be asserted that there exists a vector $\tau^* \in R^m$ which maximizes $Q_\theta(\tau)$ and usually there is a vector $\tau_n^* \in R^m$ which maximized $Q_{\bar{\varphi}}(\tau)$. Of course these vectors, when they exist, will be unique. In order to establish their existence, some additional lemmas will be helpful.

Lemma 3.6. For any continuous concave function h on R^m and real number a the set $\{\tau \in R^m : h(\tau) \geq a\}$ is a closed convex set.

Proof. See Rockafellar (1970).

The next lemma concerns an equivalent condition for a closed convex set in R^m to be a bounded set. Clearly a set $C \subseteq R^m$ is closed, convex, bounded if and only if the set $C - x_0 = \{c - x_0 : c \in C\}$ is closed, convex, bounded. The point x_0 is arbitrary so it could be chosen in C. Hence without loss of generality assume that the set contains the origin. In other words, let $x_0 \in C$. Then we will

prove that

(i) C is convex, closed and bounded if and only if

(ii) $C - x_0$ is convex, closed and bounded if and only if

(iii) $C - x_0$ contains no rays if and only if

(iv) C contains no rays

A ray will be a point set of the form $\{x+\rho c: \rho \geq 0\}$ where $x, c \in R^m$ and $\|c\| = 1$. That (i) and (ii) are equivalent, and that (iii) and (v) are equivalent, is immediate. The lemma asserts that (ii) and (iii) are equivalent; one may find an equivalent statement in Rockafellar (1970).

Lemma 3.7. Let C be a closed convex set in R^m which contains the origin. Then C is bounded if and only if C contains no rays.

Proof. Suppose C contains the ray $\{x+\rho c: \rho \geq 0\}$. By the triangle inequality, $\|\rho c\| - \|x\| \leq \|\rho c + x\|$ which implies $\rho - \|x\| \leq \|\rho c + x\|$, so $\|\rho c + x\| \to \infty$ as $\rho \to \infty$, hence C is unbounded.

Next suppose C is unbounded. It will follow that C contains a ray. Since C is unbounded, there exists $\{x_n\}_{n=1}^{\infty} \subseteq C$ such that $\|x_n\| \to \infty$ as $n \to \infty$. By convexity of C we can assume that $\|x_n\| = n$. We will construct a ray emanating from x_1 that lies

within C.

Consider the sequence $\{x_n - x_1\}_{n=2}^{\infty}$.

$$x_n - x_1 = \|x_n - x_1\| \left[\frac{(x_n - x_1)_1}{\|x_n - x_1\|}, \ldots, \frac{(x_n - x_1)_m}{\|x_n - x_1\|} \right]$$

Since

$$-1 \leq \frac{(x_n - x_1)_i}{\|x_n - x_1\|} \leq 1 \quad i = 1, 2, \ldots, m \quad n = 2, 3, \ldots$$

the sequence $\left\{ \frac{(x_n - x_1)_1}{\|x_n - x_1\|} \right\}_{n=2}^{\infty}$ contains a convergent subsequence $\left\{ \frac{(x_{n_k} - x_1)_1}{\|x_{k_n} - x_1\|} \right\}_{k=1}^{\infty}$ with the property that

$$\lim_{k \to \infty} \frac{(x_{n_k} - x_1)}{\|x_{n_k} - x_1\|} = \limsup_n \frac{(x_n - x_1)}{\|x_n - x_1\|} = x_1^* .$$

Again, the sequence $\left\{ \frac{(x_{n_k} - x_1)_2}{\|x_{n_k} - x_1\|} \right\}_{k=1}^{\infty}$ contains a subsequence $\left\{ \frac{(x_{n_{k_j}} - x_1)_2}{\|x_{n_{k_j}} - x_1\|} \right\}_{j=1}^{\infty}$ with the property that

$$\lim_{j \to \infty} \frac{(x_{n_{k_j}} - x_1)_2}{\|x_{n_{k_j}} - x_1\|} = \limsup_k \frac{(x_{n_k} - x_1)_2}{\|x_{n_k} - x_1\|} = x_2^* .$$

Repeating the argument m times, there exists a subsequence

$\{x_{a_i}-x_1\}_{i=1}^{\infty}$ with the property that

$$\frac{(x_{a_i}-(x_1)_1}{\|x_{a_i}-x_1\|} \to x_1^* \quad \text{as} \quad i \to \infty$$

$$\vdots$$

$$\frac{(x_{a_i}-x_1)_m}{\|x_{a_i}-x_1\|} \to x_m^* \quad \text{as} \quad i \to \infty$$

Since $\|x_n\| - \|x_1\| \leq \|x_n-x_1\|$ implies $n-1 \leq \|x_n-x_1\|$ and $i \leq a_i$, we have $i-1 \leq a_i-1 = \|x_{a_i}\| - \|x_1\| \leq \|x_{a_i}-x_1\|$ so that $\|x_{a_i}-x_1\| \to \infty$ as $i \to \infty$.

Now for all i,

$$1 = \sum_{j=1}^{m} \frac{(x_{a_i}-x_1)_j^2}{\|x_{a_i}-x_1\|^2} = \sum_{j=1}^{m} \left\{ \frac{(x_{a_i}-x_1)}{\|x_{a_i}-x_1\|} \right\}^2$$

so then

$$1 = \lim_{i \to \infty} \sum_{j=1}^{m} \left\{ \frac{(x_{a_i}-x_1)}{\|x_{a_i}-x_1\|} \right\}^2$$

$$1 = \sum_{j=1}^{m} (x_j^*)^2$$

The claim is that C contains the ray $\{x_1+\rho(x_1^*, x_2^*, \ldots, x_m^*): \rho \geq 0\}$. Certainly $x_1 \in C$. Recall that C is closed. For any $\rho_0 > 0$, it will be shown that

$x_0 = x_1 + \rho_0(x_1^*, \ldots, x_m^*)$ is a limit of points in C. There exists an integer I_0 such that $i \geq I_0$ implies $\|x_{a_i} - x_1\| \geq \rho_0$. Convexity of C will now be used to generate a sequence $\{z_n\}_{I_0}^{\infty}$ of points in C such that $z_n \to x_0$ as $n \to \infty$. For $i \geq I_0$, the point

$$z_i = x_1 + \rho_0\left[\frac{(x_{a_i} - x_1)_1}{\|x_{a_i} - x_1\|}, \ldots, \frac{(x_{a_i} - x_1)_m}{\|x_{a_i} - x_1\|}\right]$$

is in C. Noticing that

$$\|x_0 - z_i\| = \|\rho_0(x_1^*, \ldots, x_m^*) - \rho_0\left[\frac{(x_{a_i} - x_1)_1}{\|x_{a_i} - x_1\|}, \ldots, \frac{(x_{a_i} - x_1)_m}{\|x_{a_i} - x_1\|}\right]\|,$$

it is clear that $z_i \to x_0$ as $i \to \infty$, hence $x_0 \in C$ and so C contains the ray $\{x_1 + \rho(x_1^*, \ldots, x_m^*) : \rho \geq 0\}$. This concludes the lemma.

An Optimization Problem

Now that some basic tools have been presented, the function $Q_\theta(\tau) = \sum_{i=1}^{m} \tau_i \theta_i - \psi_m(\tau)$ can be looked at in more detail. Recall that θ_i is the i-th Fourier coefficient of p(x), i.e.,

$$\theta_i = \int_{-1}^{1} \varphi_i(x) p(x) dx$$

$$= E[\varphi_i(x) | p]$$

Lemma 3.2 asserted that $Q_a(\tau) = \sum_{i=1}^{m} \tau_i a_i - \psi_m(\tau)$ is a strictly concave function of τ, so $Q_\theta(\tau)$ is strictly concave. By Lemma 3.4, $Q_\theta(\tau)$ will have a point of maximality (unique) if and only if $\theta = (\theta_1, \ldots, \theta_m)' \in \mathcal{R}_{\nabla\psi_m}$, the range of $\nabla\psi_m$. But we can also appeal to the continuity of ψ_m to get some information.

By its definition, $\psi_m(0) = \log 2 > 0$. Let n be a positive integer and define the set S_n by

$$S_n = \{\tau \in R^m : Q_\theta(\tau) \geq -n\}$$

S_n is not empty since it contains the point $\tau = 0$. Also we have that

$$\sup_{\tau \in R^m} Q_\theta(\tau) = \sup_{\tau \in S_n} Q_\theta(\tau).$$

Since $Q_\theta(\tau)$ is a continuous function of τ, Lemma 3.6 implies that S_n is a closed, convex subset or R^m. If S_n were also bounded, then $Q_\theta(\tau)$ would actually attain its supremum on S_n, because $Q_\theta(\tau)$ would then be continuous on a compact set. The next lemma sets out to show that S_n is actually bounded. (This is also discussed in Barndorff-Nielsen (1970).)

Lemma 3.8. For arbitrary positive integer n the set $S_n = \{\tau \in R^m : Q_\theta(\tau) \geq -n\}$ is bounded.

Proof. We make use of Lemma 3.7 and show that S_n contains no rays. To do this consider an arbitrary ray $x + \rho c$, $\rho \geq 0$, and the directional derivative $D_{Q_\theta}(x+\rho c, c)$ of $Q_\theta(\tau)$ at the point $\tau = x + \rho c$ and in the direction c. If the directional derivative becomes negative as $\rho \to \infty$, use of Taylor's theorem will imply that $Q_\theta(x+\rho c) \to -\infty$ as $\rho \to \infty$, and this would of course mean that the ray $x + \rho c$, $\rho \geq 0$, cannot be contained in S_n. (We point out that because $Q_\theta(\tau)$ is strictly concave, the function $f(\rho) = Q_\theta(x+\rho c)$ is also strictly concave and therefore $f'(\rho) = D_{Q_\theta}(x+\rho c, c)$ is a strictly monotone decreasing function of ρ.)

Now by the definition of the directional derivative we have

$$D_{Q_\theta}(x+\rho c, c) = \frac{d}{d\rho} Q_\theta(x+\rho c)$$

$$= \frac{d}{d\rho} \{ \sum_{i=1}^{m} (x_i+\rho c_i)\theta_i - \psi_m(x+\rho c) \}$$

$$= \sum_{i=1}^{m} c_i\theta_i - \sum_{i=1}^{m} c_i \frac{\partial \psi_m}{\partial \tau_i} (x+\rho c)$$

Let $\theta_i^m(x+\rho c) = \frac{\partial \psi_m}{\partial \tau_i}(x+\rho c)$ (we shall use the two interchangeably). Now the m-degree polynomial $\sum_{i=1}^{m} c_i \varphi_i(x)$ takes on its maximum value on the interval $[-1, 1]$ no more than m times. Recall that $p(x)$ is positive and continuous. Then because

$$\sum_{i=1}^{m} c_i\theta_i = \sum_{i=1}^{m} c_i \int_{-1}^{1} \varphi_i(x)p(x)dx$$

$$= \int_{-1}^{1} \Big(\sum_{i=1}^{m} c_i\varphi_i(x) \Big) p(x)dx$$

We see that

$$\sum_{i=1}^{m} c_i\theta_i < \max_{-1 \leq x \leq 1} \sum_{i=1}^{m} c_i\varphi_i(x)$$

for all vectors c in R^m with $\|c\| = 1$. The claim is that

$$\lim_{\rho \to \infty} D_{Q_\theta}(x+\rho c, c) = \sum_{i=1}^{m} c_i\theta_i - \max_{-1 \leq x \leq 1} \sum_{i=1}^{m} c_i\varphi_i(x) < 0$$

The proof of this claim will actually be contained in the next lemma, so for now it will be assumed that the claim is true.

Let ρ_0 be large enough so that $D_{Q_\theta}(x+\rho_0 c, c) < 0$. For $\rho > \rho_0$ and some $\lambda = \lambda(\rho)$ with $0 \leq \lambda \leq 1$,

$$Q_\theta(x+\rho c) = Q_\theta(x+\rho_0 c) + D_{Q_\theta}(x+[\rho_0+\lambda(\rho-\rho_0)]c, c)(\rho-\rho_0),$$

$$< Q_\theta(x+\rho_0 c) + D_{Q_\theta}(x+\rho_0 c, c)(\rho-\rho_0)$$

and hence $Q_\theta(x+\rho c) \to -\infty$ as $\rho \to \infty$. Therefore S_n cannot contain the ray $x + \rho c$, $\rho \geq 0$ and so S_n is bounded. This concludes

the proof of Lemma 3.8.

The claim contained in Lemma 3.8 needs to be substantiated; namely that

$$\lim_{\rho \to \infty} \sum_{i=1}^{m} c_i \theta_i^m(x+\rho c) = \max_{-1 \leq x \leq 1} \sum_{i=1}^{m} c_i \varphi_i(x)$$

Algebracially we have that

$$\sum_{i=1}^{m} c_i \theta_i^m(x+\rho c) = \sum_{i=1}^{m} c_i \frac{\partial \psi_m}{\partial \tau_i}(x+\rho c)$$

$$= \sum_{i=1}^{m} c_i \int_{-1}^{1} \varphi_i(x) e^{\Sigma_{j=1}^{m}(x_j+\rho c_j)\varphi_j(x) - \psi_m(x+\rho c)} dx$$

$$= \frac{\int_{-1}^{1} \sum_{i=1}^{m} c_i \varphi_i(x) \left[e^{\Sigma_{j=1}^{m}(x_j+\rho c_j)\varphi_j(x)} \right] dx}{e^{\psi_m(x+\rho c)}}$$

$$= \frac{\int_{-1}^{1} \sum_{i=1}^{m} c_i \varphi_i(x) \left[e^{\Sigma_{j=1}^{m} x_j \varphi_j(x)} \right] \left[e^{\Sigma_{j=1}^{m} c_j \varphi_j(x)} \right]^{\rho} dx}{\int_{-1}^{1} \left[e^{\Sigma_{j=1}^{m} x_j \varphi_j(x)} \right] \left[e^{\Sigma_{j=1}^{m} c_j \varphi_j(x)} \right]^{\rho} dx}$$

The limiting behavior of this last expression as $\rho \to \infty$ will be covered by Lemma 3.9:

Lemma 3.9. Suppose $\tilde{y}$ is a bounded random variable and let $K = \text{ess sup } \tilde{y}$. If $b(\tilde{y})$ is a continuous function of $\tilde{y}$ then

$$\lim_{\rho \to \infty} \left\{ \frac{E[b(\tilde{y})e^{\tilde{y}\rho}]}{E[e^{\tilde{y}\rho}]} \right\} = b(K)$$

Proof. There exists for any $\epsilon > 0$ a $\delta > 0$ such that $|y-K| \leqq \delta$ implies $|b(y)-b(K)| < \frac{\epsilon}{2}$. For such a δ we have

$$\left| \frac{E[b(\tilde{y})e^{\tilde{y}\rho}]}{E[e^{\tilde{y}\rho}]} - b(K) \right|$$

$$= \left| \frac{E[(b(\tilde{y})-b(K))e^{\tilde{y}\rho}]}{E[e^{\tilde{y}\rho}]} \right|$$

$$\leqq \frac{\int_{|y-K| \leqq \delta} |b(y)-b(K)| e^{y\rho} dP(y)}{\int e^{y\rho} dP(y)} + \frac{\int_{|y-K| > \delta} |b(y)-b(K)| e^{y\rho} dP(y)}{\int e^{y\rho} dP(y)}$$

$$\leqq \frac{\epsilon}{2} + \frac{\int_{|y-K| > \delta} |b(y)-b(K)| e^{y\rho} dP(y)}{\int e^{y\rho} dP(y)}$$

Now

$$\frac{\int_{|y-K| > \delta} |b(y)-b(K)| e^{y\rho} dP(y)}{\int e^{y\rho} dP(y)} \leqq \frac{2M \int_{|y-K| > \delta} e^{y\rho} dP(y)}{\int e^{y\rho} dP(y)}$$

where $M = \text{ess sup } b(\tilde{y})$. To finish the proof it is sufficient to show

that

$$\lim_{\rho \to \infty} \left\{ \frac{\int e^{y\rho} dP(y)}{\int_{|y-K| > \delta} e^{y\rho} dP(y)} \right\} = \infty$$

We have

$$\frac{\int e^{y\rho} dP(y)}{\int_{|y-K| > \delta} e^{y\rho} dP(y)} = 1 + \frac{\int_{|y-K| \leq \delta} e^{y\rho} dP(y)}{\int_{|y-K| > \delta} e^{y\rho} dP(y)}$$

$$= 1 + \frac{\int_{|y-K| \leq \delta} e^{(y-K)\rho} dP(y)}{\int_{|y-K| > \delta} e^{(y-K)\rho} dP(y)}$$

For $|y-K| \leq \delta$ we have $K-y \overset{a.e.}{\leq} \delta$ or $-\delta \leq y-K$, and then

$$\int_{|y-K| \leq \delta} e^{(y-K)\rho} dP(y) \geq e^{-\delta\rho} \int_{|y-K| \leq \delta} dP(y)$$

$$\geq e^{-\delta\rho} P_r\{|y-K| \leq \delta\}$$

$$> 0$$

From this we have

$$\frac{\int_{|y-K| \leq \delta} e^{(y-K)\rho} dP(y)}{\int_{|y-K| > \delta} e^{(y-K)\rho} dP(y)} \geq \frac{P_r\{|y-K| \leq \delta\}}{\int_{|y-K| > \delta} e^{(y-K+\delta)\rho} dP(y)}$$

But on the set where $|y-K| > \delta$, we have

$$e^{(y-K+\delta)} < 1$$

almost surely, and therefore

$$\lim_{\rho \to \infty} \int_{|y-K| > \delta} e^{(y-K+\delta)\rho} dP(y) = 0$$

and this concludes the proof.

With the completion of the last three lemmas, we have shown that if θ is the vector whose elements are the first m Fourier-type coefficients of the true density $p(x)$, then the function $Q_\theta(\tau) = \sum_{i=1}^{m} \tau_i \theta_i - \psi_m(\tau)$ actually attains its supremum at some unique point. This motivates the following definition.

Definition 3.1. By the vector τ^* we mean that unique point in R^m such that $Q_\theta(\tau)$ attains its supremum at that point.

Remarks. By Lemma 3.4, $\theta \in \mathcal{R}_{\nabla\psi_m}$ and by Lemma 3.3 $\nabla Q_\theta(\tau)$ is zero at $\tau = \tau^*$. This means that

$$\nabla Q_\theta(\tau) = \nabla \{ \sum_{i=1}^{m} \tau_i \theta_i - \psi_m(\tau) \}$$

$$= \theta - \nabla \psi_m(\tau)$$

satisfies

$$0 = \theta - \nabla\psi_m(\tau^*)$$

or

$$\nabla\psi_m(\tau^*) = \theta$$

We give further results on τ^* later. For now we need some preliminary lemmas. The lemma coming next is stated in Barndorff-Nielsen.

The Range of the Gradient of ψ_m

Lemma 3.10. $\mathcal{R}_{\nabla\psi_m}$ is open, bounded and convex.

Proof. The mapping $\nabla\psi_m : E^m \to E^m$ is given the usual way by

$$\nabla\psi_m(\tau) = \left(\frac{\partial\psi_m(\tau)}{\partial\tau_1}, \ldots, \frac{\partial\psi_m(\tau)}{\partial\tau_m}\right),$$

where

$$\frac{\partial\psi_m(\tau)}{\partial\tau_j} = \int_{-1}^{1} \varphi_j(x) e^{\sum_{i=1}^{m}\tau_i\varphi_i(x) - \psi_m(\tau)}\,dx \qquad j = 1, 2, \ldots, m$$

The boundedness of $\mathcal{R}_{\nabla\psi_m}$ follows from the boundedness of the functions $\varphi_i(x)$ on $[-1, 1]$.

The property of openness is a direct consequence of the inverse function theorem (see Apostol (1960, p. 144)) and the fact that the

Jacobian $J_{\nabla\psi_m}(\tau) = \det(H_{\psi_m}(\tau)) \neq 0$ for all $\tau \in R^m$.

To prove convexity let $\theta = \nabla\psi_m(\tau)$ and $\theta' = \nabla\psi_m(\tau')$ be two points of $\mathcal{R}_{\nabla\psi_m}$. Then

$$\begin{bmatrix} \theta_1 \\ \vdots \\ \theta_m \end{bmatrix} = \begin{bmatrix} \int_{-1}^{1} \varphi_1(x) e^{\sum_{i=1}^{m} \tau_i \varphi_i(x) - \psi_m(\tau)} dx \\ \vdots \\ \int_{-1}^{1} \varphi_m(x) e^{\sum_{i=1}^{m} \tau_i \varphi_i(x) - \psi_m(\tau)} dx \end{bmatrix}$$

and

$$\begin{bmatrix} \theta'_1 \\ \vdots \\ \theta'_m \end{bmatrix} = \begin{bmatrix} \int_{-1}^{1} \varphi_i(x) e^{\sum_{i=1}^{m} \tau'_1 \varphi_i(x) - \psi_m(\tau')} dx \\ \vdots \\ \int_{-1}^{1} \varphi_m(x) e^{\sum_{i=1}^{m} \tau'_i \varphi_i(x) - \psi_m(\tau')} dx \end{bmatrix}$$

Let

$$f_1(x) = e^{\sum_{i=1}^{m} \tau_i \varphi_i(x) - \psi_m(\tau)}$$

$$f_2(x) = e^{\sum_{i=1}^{m} \tau'_i \varphi_i(x) - \psi_m(\tau')}$$

Let $0 < \lambda < 1$. We have to show that there exists

$$f_3(x) = e^{\sum_{i=1}^{m} \tau''_i \varphi_i(x) - \psi_m(\tau'')}$$

such that

$$\begin{bmatrix} \theta_1'' \\ \vdots \\ \theta_m'' \end{bmatrix} = \lambda \begin{bmatrix} \theta_1 \\ \vdots \\ \theta_m \end{bmatrix} + (1-\lambda) \begin{bmatrix} \theta_1' \\ \vdots \\ \theta_m' \end{bmatrix}$$

$$= \begin{bmatrix} \int_{-1}^{1} \varphi_1(x)[\lambda f_1(x)+(1-\lambda)f_2(x)]dx \\ \vdots \\ \int_{-1}^{1} \varphi_m(x)[\lambda f_1(x)+(1-\lambda)f_2(x)]dx \end{bmatrix}$$

$$= \begin{bmatrix} \int_{-1}^{1} \varphi_1(x)f_3(x)dx \\ \vdots \\ \int_{-1}^{1} \varphi_m(x)f_3(x)dx \end{bmatrix} \in \mathcal{R}_{\nabla\psi_m}$$

Now consider the optimization problem

$$\sup_{\tau \in E^m} \left\{ \sum_{i=1}^{m} \tau_i \theta_i'' - \psi_m(\tau) \right\}$$

This is the same problem that was considered before, except the function being maximized is $Q_{\theta''}(\tau)$ instead of $Q_{\theta}(\tau)$. The problem has an exact and unique solution if and only if $\nabla\psi_m = \theta''$ has a solution in τ. $Q_{\theta''}(\tau)$ is strictly concave by a previous lemma and so

$$S_n = \{\tau \in R^m : Q_{\theta''}(\tau) \geq -n\}$$

is a nonempty, closed, convex set. We must show that S_n is

bounded, so again we try to show that S_n contains no rays.

Let $D_Q(x+\rho c, c)$ be the directional derivative of $Q_{\theta''}(\tau)$ at the point $x+\rho c$ and in the direction c. Just as in Lemma 3.8, we get

$$D_Q(x+\rho c,c) = \sum_{i=1}^{m} c_i\theta_i'' - \frac{\int_{-1}^{1} \sum_{i=1}^{m} c_i\varphi_i(x) \left[e^{\Sigma_{i=1}^{m} x_i\varphi_i(x)}\right]\left[e^{\Sigma_{i=1}^{m} c_i\varphi_i(x)}\right]^{\rho} dx}{\int_{-1}^{1} \left[e^{\Sigma_{i=1}^{m} x_i\varphi_i(x)}\right]\left[e^{\Sigma_{i=1}^{m} c_i\varphi_i(x)}\right]^{\rho} dx}$$

Lemma 3.9 implies that

$$\lim_{\rho \to \infty} D_Q(x+\rho c, c) = \left[\sum_{i=1}^{m} c_i\theta_i'' - \max_{-1 \leq x \leq 1} \left\{\sum_{i=1}^{m} c_i\varphi_i(x)\right\}\right]$$

Also we have that

$$\begin{aligned}
\sum_{i=1}^{m} c_i\theta_i'' &= \sum_{i=1}^{m} c_i\{\lambda\theta_i+(1-\lambda)\theta_i'\} \\
&= \sum_{i=1}^{m} c_i\{\lambda \int_{-1}^{1} \varphi_i(x)f_1(x)dx+(1-\lambda)\int_{-1}^{1} \varphi_i(x)f_2(x)dx\} \\
&= \sum_{i=1}^{m} c_i\int_{-1}^{1} \varphi_i(x)[\lambda f_1(x)+(1-\lambda)f_2(x)]dx \\
&= \int_{-1}^{1} (\sum_{i=1}^{m} c_i\varphi_i(x))[\lambda f_1(x)+(1-\lambda)f_2(x)]dx \\
&< \max_{-1 \leq x \leq 1} \left\{\sum_{i=1}^{m} c_i\varphi_i(x).\right\}
\end{aligned}$$

Therefore $D_Q(x+\rho c, c)$ is negative for all sufficiently large ρ and decreasing as a function of ρ, which implies that

$$Q_{\theta''}(x+\rho c) < -1 \quad \text{for} \quad \rho \quad \text{large.}$$

Hence S_n contains no rays and then S_n is bounded. It then follows that $Q_{\theta''}(\tau)$ attains its supremum at some point $\tau'' \in R^m$. By Lemma 3.3. it must hold that $\nabla Q_{\theta''}(\tau'') = 0$. Then

$$\theta'' - \nabla \psi_m(\tau'') = 0$$

$$\theta'' = \nabla \psi_m(\tau'')$$

and so $\theta'' = \lambda\theta + (1-\lambda)\theta'$ is an element of $\mathcal{R}_{\nabla\psi_m}$, hence $\mathcal{R}_{\nabla\psi_m}$ is convex. This concludes the lemma.

Definition 3.2. Let $\mathcal{B}$ be the Borel subsets of R^1 and let $\mathcal{B}[-1, 1] = [-1, 1] \cap \mathcal{B}$. If P is a probability measure on $\mathcal{B}[-1, 1]$ then a point $x \in [-1, 1]$ is a point of support for P if for every open interval (a, b) which contains x, we have $P((a, b) \cap [-1, 1]) > 0$.

The next lemma gives an alternate characterization of the set $\mathcal{R}_{\nabla\psi_m}$ in terms of points of support. Accordingly, define the set $\mathcal{H}_m$ to be the collection of $\theta \in R^m$ such that

$$\theta_i = \int_{-1}^{1} \varphi_i(x)dP(x)$$

where P is a probability measure on $([-1, 1], \mathcal{B}[-1, 1])$ which has more than m distinct points of support in $[-1, 1]$.

Lemma 3.11. $\Theta_m = \mathcal{R}_{\nabla\psi_m}$.

Proof. Let $\theta \in \mathcal{R}_{\nabla\psi_m}$. Then $\theta_i = \int_{-1}^{1} \varphi_i(x)dP(x)$ where $dP(x) = e^{\sum_{i=1}^{m}\tau_i\varphi_i(x)-\psi_m(\tau)}dx$ for some $\tau \in R^m$. Clearly every point in $[-1, 1]$ is a point of support for P, so that $\theta \in \Theta_m$. Hence $\mathcal{R}_{\nabla\psi_m} \subseteq \Theta_m$.

On the other hand, suppose $\theta \in \Theta_m$. Consider the optimization problem

$$\sup_{\tau \in R^m} \sum_{i=1}^{m} \tau_i\theta_i - \psi_m(\tau).$$

Just as before, the function $Q_\theta(\tau) = \sum_{i=1}^{m} \tau_i\theta_i - \psi_m(\tau)$ is strictly concave and attains its supremum if and only if the set $S_n = \{\tau \in R^m : Q_\theta(\tau) \geq -n\}$ is bounded. As before, we show that S_n contains no rays. Again letting $D_Q(x+\rho c, c)$ be the directional derivative of $Q_\theta(\tau)$ along the ray $x + \rho c$, we have by Lemma 3.9 that

$$\lim_{\rho \to \infty} D_Q(x+\rho c, c) = \sum_{i=1}^{m} c_i\theta_i - \max_{-1 \leq x \leq 1} \sum_{i=1}^{m} c_i\varphi_i(x).$$

Since $\sum_{i=1}^{m} c_i\varphi_i(x)$ is a polynomial of degree $\leq m$ and

$$\sum_{i=1}^{m} c_i\theta_i = \int_{-1}^{1} \sum_{i=1}^{m} c_i\varphi_i(x)\, dP(x)$$

$$< \max_{-1 \leq x \leq 1} \sum_{i=1}^{m} c_i\varphi_i(x)$$

we know that $D_Q(x+\rho c, c)$ is negative for ρ sufficiently large. Just as in Lemma 3.8, the set S_n must be bounded, which implies that $Q_\theta(\tau)$ attains its supremum at some vector $\tau^* \in R^m$. Then by Lemma 3.3 we must have that $\theta = \nabla\psi_m(\tau^*)$. But this means that $\theta \in \mathcal{R}_{\nabla\psi_m}$. Hence $\Theta_m \subseteq \mathcal{R}_{\nabla\psi_m}$ and so $\Theta_m = \mathcal{R}_{\nabla\psi_m}$.

A Note on the Closure of the Range of the Gradient of ψ_m

Define a continuous arc in R^m by the mapping $L:[-1, 1] \to R^m$ which is given by

$$L(x) = (\varphi_1(x), \varphi_2(x), \ldots, \varphi_m(x)).$$

The claim is that for every $x_0 \in [-1, 1]$, the vector $L(x_0)$ is in $\overline{\Theta}_m = \overline{\mathcal{R}}_{\nabla\psi_m}$. (This again is contained in Barndorff-Nielsen (1970).)

Lemma 3.12. For every $x_0 \in [-1, 1]$, $L(x_0) \in \overline{\Theta}_m$.

Proof. Assume that $x_0 \in (-1, 1)$. Define the function $p_n(x)$ by

$$p_n(x) = \begin{cases} 2n, & |x-x_0| \leq \frac{1}{n} \\ 0, & |x-x_0| > \frac{1}{n} \end{cases}$$

For all sufficiently large n, p_n will be a density on $[-1, 1]$. Let $\theta_n \in R^m$ be defined by

$$\theta_n = \begin{bmatrix} \theta_{n1} \\ \vdots \\ \theta_{nm} \end{bmatrix} \qquad n = 1, 2, \ldots$$

where $\theta_{ni} = \int_{-\infty}^{\infty} \varphi_i(x) p_n(x)\, dx$. Let $\epsilon > 0$. Since $\varphi_1(x), \ldots, \varphi_m(x)$ are polynomials, hence continuous, there exists $\delta > 0$ such that $|x-x_0| < \delta$ implies $|\varphi_i(x)-\varphi_i(x_0)| < \frac{\epsilon}{m}$, $i = 1, 2, \ldots, m$.

For sufficiently large n, $\theta_n \in \Theta_m$ since $p_n(x)$ has an infinite number of points of support. Also

$$\| L(x_0)-\theta_n \| = \{ \sum_{i=1}^{m} (\varphi_i(x_0)-\theta_{ni})^2 \}^{1/2} \leq \sum_{i=1}^{m} |\varphi_i(x_0)-\theta_{ni}| .$$

For

$$n > \frac{1}{\delta} + \frac{1}{1-x_0} + \frac{1}{x_0+1}$$

we have

$$|\theta_{ni}-\varphi_i(x_0)| \leq \int_{x_0-\frac{1}{n}}^{x_0+\frac{1}{n}} |\varphi_i(x)-\varphi_i(x_0)|\, 2n\, dx$$

$$< 2n\frac{\epsilon}{m}\int_{x_0-\frac{1}{n}}^{x_0+\frac{1}{n}} dx$$

$$= \frac{\epsilon}{m}$$

Then for $n > \frac{1}{\delta} + \frac{1}{1-x_0} + \frac{1}{x_0+1}$, $\|L(x_0)-\theta_n\| < \epsilon$ and hence $\|L(x_0)-\theta_n\| \to 0$ as $n \to \infty$. Therefore $L(x_0) \in \overline{\Theta}_m = \overline{\mathcal{R}}_{\nabla\psi_m}$.

The proof when $x_0 = \pm 1$ is practically the same. Alternately, if $x_0 = \pm 1$, let $\{x_n\}$ be a sequence in $(-1, 1)$ such that $x_n \to x_0$ as $n \to \infty$. Clearly $L(x_n) \to L(x_0)$ as $n \to \infty$ so that $L(x_0) \in \overline{\overline{\Theta}}_m$. But $\overline{\overline{\Theta}}_m = \overline{\Theta}_m$, hence $L(x_0) \in \overline{\Theta}_m$.

In the Context of Sampling

We now turn back to the sampling problem. Suppose a random sample of size n is obtained and the values given by $y_1, y_2, \ldots, y_n$. These values are all distinct, with probability one. We had previously defined $\bar{\varphi}_j$, $j = 1, 2, \ldots, m$, where $\bar{\varphi} = \frac{1}{n}\sum_{i=1}^{n} \varphi_j(y_i)$. The function of interest here will be $Q_{\bar{\varphi}}(\tau) = \sum_{i=1}^{m} \tau_i \bar{\varphi}_i - \psi_m(\tau)$. The claim is that $\bar{\varphi} = (\bar{\varphi}_1, \ldots, \bar{\varphi}_m)'$ will be in the range of $\nabla\psi_m$, with probability one.

Lemma 3.13. Let $\bar{\varphi} = (\bar{\varphi}_1, \ldots, \bar{\varphi}_m)'$ and suppose the sample

size is n where $n \geq m+1$. Then with probability one, $\bar{\varphi} \in \mathcal{R}_{\nabla\psi_m}$.

Proof. Suppose the sample values are $x_1, x_2, \ldots, x_n$. With probability one these values are all different. Then the empirical distribution function has n points of support and the proof follows from Lemma 3.11.

Lemma 3.14. If $n \geq m+1$ then the function $Q_{\bar{\varphi}}(\tau) = \sum_{i=1}^{m} \tau_i \bar{\varphi}_i - \psi_m(\tau)$ attains its supremum over R^m, with probability one.

Proof. With probability one, $\bar{\varphi} \in \mathcal{R}_{\nabla\psi_m}$.

Remark. If $\bar{\varphi} \in \mathcal{R}_{\nabla\psi_m}$ then there exists $\tau_n^* \in R^m$ such that $\bar{\varphi} = \nabla\psi_m(\tau_n^*)$. The vector τ_n^* would be the MLE of $\tau = (\tau_1, \ldots, \tau_m)'$ if we were sampling from the canonical exponential density p_m. There exists a vector $\tau^* \in R^m$ such that $\theta = \nabla\psi_m(\tau^*)$. Since $\bar{\varphi} \xrightarrow{a.s.} \theta$ as $n \to \infty$, $\nabla\psi_m(\tau_n^*) \xrightarrow{a.s.} \nabla\psi_m(\tau^*)$. A natural question arises: does $\tau_n^* \xrightarrow{a.s.} \tau^*$? The answer is in the affirmative but to give a proof, some more machinery is needed. (The following can be found in Barndorff-Nielsen (1970). The proof is due to Morgan (1969).)

Lemma 3.15. The mapping $\nabla\psi_m : R^m \to \mathcal{R}_{\nabla\psi_m}$ is 1-1.

Proof. Suppose there are two different vectors τ_0 and τ_1 such that $\nabla\psi_m(\tau_0) = \nabla\psi_m(\tau_1)$. Let $c = (\tau_1 - \tau_0)/\|\tau_1 - \tau_0\|$ and define

a real valued function $g(t)$ by

$$g(t) = \psi_m(\tau_0+ct)$$

Since ψ_m is a strictly convex function it follows that g is a strictly convex function of t, hence $g'(t)$ is strictly monotone increasing. The formula for $g'(t)$ is

$$g'(t) = \sum_{i=1}^{m} c_i \frac{\partial \psi_m}{\partial \tau_i} (\tau_0+ct)$$

Then

$$g'(0) = \sum_{i=1}^{m} c_i \frac{\partial \psi_m}{\partial \tau_i} (\tau_0)$$

$$g'(\|\tau_1-\tau_0\|) = \sum_{i=1}^{m} c_i \frac{\partial \psi_m}{\partial \tau_i} (\tau_0+(\tau_1-\tau_0))$$

$$= \sum_{i=1}^{m} c_i \frac{\partial \psi_m}{\partial \tau_i} (\tau_1)$$

But $\nabla\psi_m(\tau_0) = \nabla\psi_m(\tau_1)$ means that $\frac{\partial \psi_m}{\partial \tau_i} (\tau_0) = \frac{\partial \psi_m}{\partial \tau_i} (\tau_1)$ for $i = 1, 2, \ldots, m$ and so $g'(0) = g'(\|\tau_1-\tau_0\|)$. This contradicts the strict monotonicity of $g'(t)$ and therefore $\nabla\psi_m$ is one-to-one onto its range.

The Inverse of the Gradient of ψ_m

The inverse of $\nabla\psi_m$ exists by the last lemma, and will be denoted by $\phi_m : \mathcal{R}_{\nabla\psi_m} \to R^m$. The next lemma discusses some properties of ϕ_m and can be found in Barndorff-Nielsen (1970).

Lemma 3.16. The mapping ϕ_m is continuous and has continuous first order partial derivatives throughout $\mathcal{R}_{\nabla\psi_m}$.

Proof. At any point $\tau \in R^m$ the Jacobian $J_{\nabla\psi_m}(\tau)$ of $\nabla\psi_m$ at τ is nonzero, since $J_{\nabla\psi_m}(\tau) = \det(H_{\psi_m}(\tau)) > 0$. The Inverse Function Theorem then applies to $\nabla\psi_m$ and the lemma is a direct consequence of this theorem.

Some Asymptotic Statements

With the function $\phi_m = (\nabla\psi_m)^{-1}$ established, consider the previous result that

$$\overline{\varphi} = \nabla\psi_m(\tau_n^*) \xrightarrow{a.s.} \nabla\psi_m(\tau^*) = \theta \quad \text{as} \quad n \to \infty$$

Applying the inverse function,

$$\phi_m(\nabla\psi_m(\tau_n^*)) \xrightarrow{a.s.} \phi_m(\nabla\psi_m(\tau^*)) \quad \text{as} \quad n \to \infty$$

But

$$\phi_m(\nabla\psi_m(\tau_n^*)) = \tau_n^*$$

and

$$\phi_m(\nabla\psi_m(\tau^*)) = \tau^*$$

so that we have proven the following theorem.

Theorem 3.1. $\tau_n^* \xrightarrow{a.s.} \tau^*$ as $n \to \infty$.

With the tools provided we can state another interesting result.

Theorem 3.2. If the true density $p(x)$ is actually of the canonical exponential family, that is, if $\tau_i = 0$ for $i > S$ where S is some positive integer, then $\tau^* = \tau$ whenever $m \geq S$, and $p_m(x|\tau^*)$ is exactly $p(x)$, i.e.,

$$\begin{aligned} p(x) &= e^{\sum_{i=1}^{\infty}\tau_i\varphi_i(x)-\psi(\tau)} \\ &= e^{\sum_{i=1}^{m}\tau_i^*\varphi_i(x)-\psi_m(\tau^*)} \\ &= p_m(x|\tau^*) \quad , \quad -1 \leq x \leq 1 \end{aligned}$$

Proof. By the remark following Definition 3.1, $\nabla\psi_m(\tau^*) = \theta$ for all m. When $m \geq S$ we also have that $\nabla\psi_m(\tau) = \theta$. By Lemma 3.15 and Lemma 3.16,

$$\phi_m(\nabla\psi_m(\tau^*)) = \phi_m(\nabla\psi_m(\tau))$$

or

$$\tau^* = \tau$$

This proves the theorem.

Remark. Throughout this chapter we have assumed that the system $\{\varphi_i\}$ was the set of normalized Legendre polynomials. However, most of the results remain valid for a more general system. We needed merely that the set of functions $\{\varphi_i\}$ be orthonormal, continuous, linearly independent (which is a condition implied by the orthonormality), and the property that any nondegenerate generalized polynomial $\sum_{i=1}^{m} c_i \varphi_i(x)$ attains its supremum (over the domain $[a, b]$) at a finite number of points or even a Lebesgue null set.

In particular, most of the results will still be valid for Fourier series expansions.

This concludes Chapter III. Most of the analytic tools concerning the functions ψ_m, Q_a, $\nabla\psi_m$, ϕ_m, etc. have been noted. The next chapter utilizes these results, where they are applied in the spirit of estimation.

IV. ESTIMATION IN EXPONENTIAL FORMS

In this chapter attention will be focused towards estimation of densities and cumulative distribution functions. Use will be made of some of the results given in the previous chapter, and some new properties will be developed as well. We assume all conditions on $p(x)$ that were mentioned in the introduction.

Recall that by assumption the density $p(x)$ has the following two representations

$$p(x) = e^{\sum_{i=1}^{\infty} \tau_i \varphi_i(x) - \psi(\tau)}, \qquad -1 \leq x \leq 1$$

$$= \sum_{i=0}^{\infty} \theta_i \varphi_i(x), \qquad -1 \leq x \leq 1$$

where θ_i is the ith Fourier coefficient of p and τ_i is the ith Fourier coefficient of $\log p$. In the usual manner the cumulative distribution function of p will be denoted by $F(y)$, where

$$F(y) = \begin{cases} 0, & y \leq -1 \\ \int_{-1}^{y} e^{\sum_{i=1}^{\infty} \tau_i \varphi_i(x) - \psi(\tau)} \, dx, & -1 < y < 1 \\ 1, & y \geq 1 \end{cases}$$

The CDF F_m

It has been previously established that there exists a unique vector $\tau^* \in R^m$ with the property that

$$\nabla \psi_m(\tau^*) = \theta$$

where $\theta = (\theta_1, \theta_2, \ldots, \theta_m)'$ is the vector of Fourier coefficients of $p(x)$. We define another cumulative distribution function $F_m(y)$ by

$$F_m(y) = \begin{cases} 0, & y \leq -1 \\ \int_{-1}^{y} e^{\sum_{i=1}^{m} \tau_i^* \varphi_i(x) - \psi_m(\tau^*)} dx, & -1 < y < 1 \\ 1, & y \geq 1 \end{cases}$$

and the corresponding density by

$$\frac{dF_m(y)}{dy} = p_m(y|\tau^*)$$

$$= e^{\sum_{i=1}^{m} \tau_i^* \varphi_i(y) - \psi_m(\tau^*)}$$

Since $p_m(y|\tau^*)$ is a smooth, even analytic, function of y on $[-1, 1]$, it too will be expressible as an orthogonal series which will be written as

$$p_m(y|\tau^*) = \sum_{i=0}^{\infty} \theta_i^m(\tau^*)\varphi_i(x)$$

where of course

$$\theta_j^m(\tau^*) = \int_{-1}^{1} \varphi_j(x) e^{\Sigma_{i=1}^m \tau_i^* \varphi_i(x) - \psi_m(\tau^*)} dx \quad j = 0, 1, 2, \ldots$$

The first theorem of this chapter relates the Fourier coefficients of $p(x)$ and $p_m(x|\tau^*)$.

Theorem 4.1. We have $\theta_i = \theta_i^m(\tau^*)$ for $i = 0, 1, 2, \ldots, m$.

Proof. The relation $\theta_0 = \theta_0^m(\tau^*)$ results from the following.

$$\begin{aligned} 1 &= \int_{-1}^{1} p(x)dx \\ &= \int_{-1}^{1} \sum_{i=0}^{\infty} \theta_i \varphi_i(x)\, dx \\ &= \sum_{i=1}^{\infty} \theta_i \int_{-1}^{1} \varphi_i(x)dx \\ &= \theta_0 \int_{-1}^{1} \varphi_0(x)dx \end{aligned}$$

by the orthogonality property of the Legendre polynomials. Similarly,

$$1 = \int_{-1}^{1} p_m(x|\tau^*)dx$$

$$= \theta_0^m(\tau^*) \int_{-1}^{1} \varphi_0(x)dx$$

This immediately implies that $\theta_0 = \theta_0^m(\tau^*)$. Now the vector τ^* satisfied the relation $\nabla\psi_m(\tau^*) = \theta$. But the vector $\nabla\psi_m(\tau^*)$ is the same as the vector $(\theta_1^m(\tau^*), \theta_2^m(\tau^*), \ldots, \theta_m^m(\tau^*))'$, that is, $\nabla\psi_m(\tau^*)$ is the expectation of the vector $(\varphi_1(x), \varphi_2(x), \ldots, \varphi_m(x))'$ where expectation is taken with respect to dF_m. Hence $\theta_i = \theta_i^m(\tau^*)$ for $i = 1, 2, \ldots, m$. This proves the theorem.

As a consequence of the preceding theorem it becomes evident that the expansion of $p_m(x|\tau^*)$ can be modified to read

$$p_m(x|\tau^*) = \sum_{i=0}^{m} \theta_i\varphi_i(x) + \sum_{i=m+1}^{\infty} \theta_i^m(\tau^*)\varphi_i(x) .$$

Also from the theorem we may write

$$\int_{-1}^{1} \varphi_i(x)dF_m(x) = \int_{-1}^{1} \varphi_i(x)dF(x) \qquad i = 0, 1, 2, \ldots, m.$$

Suppose we denote the vector space of all polynomials on R^1 which have real coefficients and degree less than or equal to m by the symbol $\mathcal{P}_m$. The normalized Legendre polynomials

$\{\varphi_0(x), \varphi_1(x), \ldots, \varphi_m(x)\}$ form an orthonormal basis for $\mathcal{P}_m$. Thus for an arbitrary element of $\mathcal{P}_m$, say $\sum_{i=0}^{m} a_i\varphi_i(x)$, the following holds:

$$\int_{-1}^{1} \sum_{i=0}^{m} a_i\varphi_i(x) \; dF_m(x) = \sum_{i=0}^{m} a_i \int_{-1}^{1} \varphi_i(x)dF_m(x)$$

$$= \sum_{i=0}^{m} a_i \int_{-1}^{1} \varphi_i(x)dF(x)$$

$$= \int_{-1}^{1} \sum_{i=0}^{m} a_i\varphi_i(x) \; dF(x)$$

In particular the monomials $1, x, x^2, \ldots, x^m$ are all members of $\mathcal{P}_m$ and so the following theorem has been established:

Theorem 4.2. The following moment relationship holds between F_m and F, namely that

$$\int_{-1}^{1} x^k dF_m(x) = \int_{-1}^{1} x^k dF_m(x) \qquad k = 0, 1, 2, \ldots, m.$$

As we wish to apply a result found in Wilks (1962), the corresponding notation will be adopted here. Define

$$\mu'_k = \int_{-1}^{1} x^k dF(x)$$

$$\mu'_{k,m} = \int_{-1}^{1} x^k dF_m(x)$$

For any integer (positive or zero) value of k, $\mu'_k = \mu'_{k,m}$ as soon as m is at least as large as k. Hence

$$\lim_{m \to \infty} \mu'_{k,m} = \mu'_k \qquad k = 0, 1, 2, \ldots$$

We state without proof a result given in Wilks: "If $\tilde{X}$ is a bounded random variable, then its c.d.f. $F(x)$ is uniquely determined by its moments μ'_k, $k = 0, 1, 2, \ldots$." Since our distribution is on the interval $[-1, 1]$, $F(x)$ is uniquely determined by its central moments. This means that $F(x)$ is the only c.d.f. which has the moments μ'_k, $k = 0, 1, 2, \ldots$.

The idea now is to consider a sequence $\tilde{X}_1, \tilde{X}_2, \tilde{X}_3, \ldots$ of random variables such that the distribution of $\tilde{X}_m$ is F_m on $[-1, 1]$. Again we state without proof a result given in Wilks:

> Let $(\tilde{X}_1, \tilde{X}_2, \tilde{X}_3, \ldots)$ be a sequence of random variables. Let the rth moment of $\tilde{X}_n$ be $\mu'_{r,n}$ and finite for all n and r. Let $\lim_{n \to \infty} \mu'_{r,n} = \mu'_r$, where μ'_r is finite for all r. Then if $(\tilde{X}_1, \tilde{X}_2, \tilde{X}_3, \ldots)$ converges in distribution to $F(x)$, $\mu'_0, \mu'_1, \mu'_2, \ldots$ is the moment-sequence of $F(x)$. Conversely, if this moment-sequence uniquely determines a c.d.f. $F(x)$, it is the limiting c.d.f. of $(\tilde{X}_1, \tilde{X}_2, \tilde{X}_3, \ldots)$.

Applying the converse portion of this result we get the following:

Theorem 4.3. Let $\tilde{X}_m$ be distributed as F_m and $\tilde{X}$ be distributed as F. Then $\tilde{X}_m \xrightarrow{d} \tilde{X}$ and in fact $F_m(y) \to F(y)$ uniformly as $m \to \infty$.

Proof. The only thing to prove is the uniform convergence. We have that $F_m(y) \to F(y)$ as $m \to \infty$ at all continuity points of F. Since F is continuous, $F_m(y) \to F(y)$ as $m \to \infty$ for all $y \in R^1$.

Also it is clear that $F_m(y) = F(y)$ for $|y| \geq 1$. Let $\epsilon > 0$. Choose n such that $\frac{1}{n+1} < \frac{\epsilon}{5}$. There exists a partition $y_0 = -1 < y_1 < y_2 < \ldots < y_n < y_{n+1} = 1$ such that $F(y_k) = \frac{k}{n+1}$, $k = 0, 1, 2, \ldots, n+1$.

Now suppose $y \in [-1, 1]$. Then for some k,

$$y_k \leq y \leq y_{k+1}.$$

We have then

$$\begin{aligned}
|F_m(y)-F(y)| &\leq |F_m(y)-F_m(y_k)| + |F_m(y_k)-F(y_k)| + |F(y_k)-F(y)| \\
&\leq |F_m(y_{k+1})-F_m(y_k)| + |F_m(y_k)-F(y_k)| + |F(y_k)-F(y_{k+1})| \\
&\leq |F_m(y_{k+1})-F(y_{k+1})| + |F(y_{k+1})-F(y_k)| \\
&\quad + |F(y_k)-F_m(y_k)| + |F_m(y_k)-F(y_k)| + \frac{1}{n+1} \\
&\leq |F_m(y_{k+1})-F(y_{k+1})| + 2|F_m(y_k)-F(y_k)| + \frac{2}{n+1}
\end{aligned}$$

Choose M such that $m > M$ implies $|F_m(y_k)-F(y_k)| < \frac{\epsilon}{5}$ for $k = 0, 1, 2, \ldots, n+1$. Then for $m > M$,

$$|F_m(y)-F(y)| \leq \frac{\epsilon}{5} + \frac{2\epsilon}{5} + \frac{2}{n+1} < \epsilon \quad , \quad -1 \leq y \leq 1$$

and so $F_m(y) \to F(y)$ uniformly as $m \to \infty$.

Consistency of F_{mn}

It will be recalled that there exists, with probability one, a unique vector τ_n^* such that $\nabla\psi_m(\tau_n^*) = \bar{\varphi}$ where $\bar{\varphi} = (\bar{\varphi}_1, \bar{\varphi}_2, \ldots, \bar{\varphi}_m)'$ and $n \geq m+1$ is the sample size. It was shown in Chapter III that $\tau_n^* \xrightarrow{a.s.} \tau^*$. Let $\tau_n^* = (\tau_{n1}^*, \tau_{n2}^*, \ldots, \tau_{nm}^*)'$. Then we have the following:

Corollary 4.3. If the c.d.f. F_{mn} is given by

$$F_{mn}(y) = \begin{cases} 0, & y \leq -1 \\ \int_{-1}^{1} e^{\sum_{i=1}^{m} \tau_{ni}^* \varphi_i(x) - \psi_m(\tau_n^*)} dx & -1 < y < 1 \\ 1, & y \geq 1 \end{cases}$$

then

$$\lim_{m\to\infty} \lim_{n\to\infty} F_{mn}(y) = F(y) \quad a.s.$$

Proof. Since $\tau_n^* \xrightarrow{a.s.} \tau^*$, $F_{mn}(y) \xrightarrow{a.s.} F_m(y)$ as $n \to \infty$.

To verify this we appeal to the Lebesgue Dominated Convergence Theorem. Define a function $g: R^{m+1} \to R^1$ by

$$g(x_1, x_2, \ldots, x_m, x_{m+1}) = \exp\{\sum_{i=1}^{m} x_i \varphi_i(x_{m+1}) - \psi_m(x_1, \ldots, x_m)\}$$

The function g is continuous in all $m+1$ variables, and so g will attain its supremum on any compact subset of R^{m+1}. The sequence of vectors $\{\tau_n^*\}$ is convergent to τ^*, so there is some m-dimensional closed cube C_m with center at τ^* and such that $\tau_n^* \in C_m$ for every n. Let $S = C_m \times [-1, 1]$. Then S is a compact subset of R^{m+1}. Let $g_{max} = \sup_{x \in S} g(x)$. Then we have that

$$g(\tau_{n1}^*, \ldots, \tau_{nm}^*, x) \leq g_{max} \quad -1 \leq x \leq 1$$

and

$$\lim_{n \to \infty} g(\tau_{n1}^*, \ldots, \tau_{nm}^*, x) = g(\tau_1^*, \ldots, \tau_m^*, x)$$

By the Lebesgue Dominated Convergence Theorem,

$$\lim_{n \to \infty} \int_{-1}^{y} g(\tau_{n1}^*, \ldots, \tau_{nm}^*, x)dx = \int_{-1}^{y} \lim_{n \to \infty} g(\tau_{n1}^*, \ldots, \tau_{nm}^*, x)dx$$

$$= \int_{-1}^{y} g(\tau_1^*, \ldots, \tau_m^*, x)dx$$

or

$$\lim_{n\to\infty} \int_{-1}^{y} e^{\sum_{i=1}^{m} \tau_{ni}^{*} \varphi_i(x) - \psi_m(\tau_n^{*})} dx = \int_{-1}^{y} e^{\sum_{i=1}^{m} \tau_i^{*} \varphi_i(x) - \psi_m(\tau^{*})} dx$$

$$\lim_{n\to\infty} F_{mn}(y) \overset{a.s.}{=} F_m(y)$$

This proves the corollary.

Orthogonal Expansion of F

At this point we note that $F(x)$ itself will have an orthogonal expansion given by

$$F(x) = \sum_{i=0}^{\infty} \beta_i \varphi_i(x)$$

and where

$$\beta_i = \int_{-1}^{1} \varphi_i(x) F(x) dx \qquad i = 0, 1, 2, \ldots$$

Define the functions $\Phi_i(x)$ by

$$\Phi_i(x) = \int_{-1}^{x} \varphi_i(z) dz \qquad i = 0, 1, 2, \ldots$$

There exists constants a_i and b_i such that

$$\varphi_i(x) = a_i \varphi'_{i+1}(x) + b_i x \varphi'_i(x)$$

This relation is contained in Jackson (1941) and the a_i and b_i

are given by

$$a_i = \frac{1}{i+1}\sqrt{\frac{2i+1}{2i+3}}$$

$$b_i = -\frac{1}{i+1}$$

for $i = 0, 1, 2, \ldots$. The prime symbol indicates differentiation.

The next lemma indicates a relation between the θ_i's and β_i's. The symbols θ_i and $\varphi_i(x)$ are taken to be zero whenever $i < 0$.

Lemma 4.1. For $i = 1, 2, 3, \ldots$ we have

$$\beta_i = \frac{a_i}{b_i+1}[\varphi_{i+1}(1)-\theta_{i+1}] + \frac{b_i}{b_i+1}[\varphi_i(1)-c_i\theta_{i-1}-d_i\theta_{i+1}]$$

where the values of c_i and d_i are given by

$$c_i = \frac{i}{\sqrt{2i-1}\sqrt{2i+1}}$$

$$d_i = \frac{i+1}{\sqrt{2i+1}\sqrt{2i+3}}$$

(notice that $c_{i+1} = d_i$). For the case $i = 0$ we have

$$\beta_0 = \frac{1}{\sqrt{2}} - \frac{1}{\sqrt{3}}\theta_1$$

Proof. First we solve for β_0. By its definition,

$$\beta_0 = \int_{-1}^{1} \varphi_0(x) F(x) dx$$

But $\varphi_0(x) = \frac{1}{\sqrt{2}}$ so that

$$\sqrt{2}\,\beta_0 = \int_{-1}^{1} F(x)dx$$

$$= [x\,F(x)\Big|_{-1}^{1}] - \int_{-1}^{1} x dF(x)$$

$$= 1 - \int_{-1}^{1} x p(x) dx$$

Recalling that $p(x) = \frac{1}{2} + \sum_{i=1}^{\infty} \theta_i \varphi_i(x)$ we have by orthogonality that

$$\sqrt{2}\,\beta_0 = 1 - \int_{-1}^{1} \theta_1 x \varphi_1(x) dx$$

$$= 1 - \sqrt{\frac{2}{3}}\,\theta_1 \int_{-1}^{1} (\sqrt{\frac{3}{2}}\,x)(\sqrt{\frac{3}{2}}\,x) dx$$

(recall $\varphi_1(x) = \sqrt{\frac{3}{2}}\,x$)

$$= 1 - \sqrt{\frac{2}{3}}\,\theta_1$$

and so

$$\beta_0 = \frac{1}{\sqrt{2}} - \frac{1}{\sqrt{3}}\,\theta_1$$

To prove the remainder of the lemma, we have

$$\Phi_i(x) = \int_{-1}^{x} \varphi_i(z)dz$$

$$= \int_{-1}^{x} [a_i \varphi'_{i+1}(z) + b_i z \varphi'_i(z)]dz$$

$$= a_i[\varphi_{i+1}(x) - \varphi_{i+1}(-1)] + b_i \int_{-1}^{x} z\varphi'_i(z)dz$$

$$= a_i[\varphi_{i+1}(x) - \varphi_{i+1}(-1)] + b_i[z\varphi_i(z)\Big|_{-1}^{x}] - b_i \int_{-1}^{x} \varphi_i(z)dz$$

$$= a_i[\varphi_{i+1}(x) - \varphi_{i+1}(-1)] + b_i[x\varphi_i(x) + \varphi_i(-1)] - b_i \Phi_i(x)$$

and consequently

$$\Phi_i(x) = \frac{a_i}{b_i+1}[\varphi_{i+1}(x) - \varphi_{i+1}(-1)] + \frac{b_i}{b_i+1}[x\varphi_i(x) + \varphi_i(-1)]$$

To compute β_i we have

$$\beta_i = \int_{-1}^{1} \varphi_i(x)F(x)dx$$

$$= \int_{-1}^{1} F(x)d\Phi_i(x)$$

$$= [F(x)\Phi_i(x)\Big|_{-1}^{1}] - \int_{-1}^{1} \Phi_i(x)dF(x)$$

$$= \Phi_i(1) - \int_{-1}^{1} \Phi_i(x)dF(x)$$

Now

$$\Phi_i(1) = \frac{a_i}{b_i+1}[\varphi_{i+1}(1)-\varphi_{i+1}(-1)] + \frac{b_i}{b_i+1}[\varphi_i(1)+\varphi_i(-1)]$$

In addition, we get directly that

$$\int_{-1}^{1} \Phi_i(x)dF(x) = \frac{a_i}{b_i+1}[\theta_{i+1}-\varphi_{i+1}(-1)] + \frac{b_i}{b_i+1}[\int_{-1}^{1} x\varphi_i(x)dF(x) + \varphi_i(-1)]$$

By a relation given in Apostol (1960),

$$x\varphi_i(x) = c_i\varphi_{i-1}(x) + d_i\varphi_{i+1}(x)$$

we have

$$\int_{-1}^{1} \Phi_i(x)dF(x) = \frac{a_i}{b_i+1}[\theta_{i+1}-\varphi_{i+1}(-1)]$$

$$+ \frac{b_i}{b_i+1}[\int_{-1}^{1}(c_i\varphi_{i-1}(x)+d_i\varphi_{i+1}(x))dF(x)+\varphi_i(-1)]$$

$$= \frac{a_i}{b_i+1}[\theta_{i+1}-\varphi_{i+1}(-1)] + \frac{b_i}{b_i+1}[c_i\theta_{i-1}+d_i\theta_{i+1}+\varphi_i(-1)]$$

Putting everything together, the expression for β_i becomes

$$\beta_i = \frac{a_i}{b_i+1}[\varphi_{i+1}(1)-\varphi_{i+1}(-1)] + \frac{b_i}{b_i+1}[\varphi_i(1)+\varphi_i(-1)]$$

$$+ \frac{a_i}{b_i+1}[\varphi_{i+1}(-1)-\theta_{i+1}] + \frac{b_i}{b_i+1}[-c_i\theta_{i-1}-d_i\theta_{i+1}-\varphi_i(-1)]$$

$$= \frac{a_i}{b_i+1}[\varphi_{i+1}(1)-\theta_{i+1}] + \frac{b_i}{b_i+1}[\varphi_i(1)-c_i\theta_{i-1}-d_i\theta_{i+1}]$$

and this proves the lemma.

Remark. It might be commented that with the relationship given in the previous lemma, it is easy to given an unbiased estimate for β_i. Since $E\bar{\varphi}_i = \theta_i$ it follows that

$$\hat{\beta}_i = \frac{a_i}{b_i+1}[\varphi_{i+1}(1)-\bar{\varphi}_{i+1}] + \frac{b_i}{b_i+1}[\varphi_i(1)-c_i\bar{\varphi}_{i-1}-d_i\bar{\varphi}_{i+1}]$$

has the desired property. In fact this estimate is actually the type of estimator used by Kronmal and Tarter (1968). They use

$$\hat{\hat{\beta}}_i = \int_{-1}^{1} \varphi_i(x)G_n(x)dx$$

where $G_n(x)$ is the empirical cumulative distribution function. Writing

$$\hat{\hat{\beta}}_i = \int_{-1}^{1} \varphi_i(x)G_n(x)dx = \int_{-1}^{1} G_n(x)d\Phi_i(x)$$

$$= [G_n(x)\Phi_i(x)\Big|_{-1}^{1}] - \int_{-1}^{1} \Phi_i(x)dG_n(x)$$

$$= \Phi_i(1) - \int_{-1}^{1} \Phi_i(x)dG_n(x)$$

$$= \Phi_i(1) - \int_{-1}^{1} [\frac{a_i}{b_i+1}(\varphi_{i+1}(x)-\varphi_{i+1}(-1)) + \frac{b_i}{b_i+1}(c_i\varphi_{i-1}(x)+d_i\varphi_{i+1}(x)+\varphi_i(-1)]dG_n(x) =$$

$$= \frac{a_i}{b_i+1}[\varphi_{i+1}(1)-\varphi_{i+1}(-1)] + \frac{b_i}{b_i+1}[\varphi_i(1)+\varphi_i(-1)]$$

$$- \frac{a_i}{b_i+1}[\bar{\varphi}_{i+1}-\varphi_{i+1}(-1)] - \frac{b_i}{b_i+1}[c_i\bar{\varphi}_{i-1}+d_i\bar{\varphi}_{i+1}+\varphi_i(-1)]$$

$$= \frac{a_i}{b_i+1}[\varphi_{i+1}(1)-\bar{\varphi}_{i+1}] + \frac{b_i}{b_i+1}[\varphi_i(1)-c_i\bar{\varphi}_{i-1}-d_i\bar{\varphi}_{i+1}]$$

$$= \hat{\beta}_i,$$

the equivalence is established. Of course they use cosine series and sine series instead of Legendre polynomials, but the estimator $\hat{\beta}_i = \hat{\hat{\beta}}_i$ is definitely of Kronmal-Tarter type.

Getting back to the c.d.f.

$$F_m(y) = \int_{-1}^{1} e^{\Sigma_{i=1}^{m}\tau_i^*\varphi_i(x)-\psi_m(\tau^*)}\,dx$$

We notice that F_m will also have an orthogonal series expansion on $[-1,1]$, say

$$F_m(y) = \sum_{i=0}^{\infty}\beta_i^m\varphi_i(y)$$

The next theorem relates F_m and F in terms of their Fourier coefficients.

Theorem 4.4. The c.d.f.'s F_m and F have the same first m Fourier coefficients, that is, we have $\beta_i^m = \beta_i$ for

$i = 0, 1, 2, \ldots, m-1$.

Proof. The function $\Phi_i(x)$ is a polynomial of degree $i+1$.

We have already shown that

$$\int_{-1}^{1}\left[\sum_{i=0}^{p} a_i\varphi_i(x)\right]dF_m(x) = \int_{-1}^{1}\left[\sum_{i=0}^{p} a_i\varphi_i(x)\right]dF(x)$$

for all such polynomials with $p \leq m$. Therefore it must hold that

$$\int_{-1}^{1}\Phi_i(x)dF_m(x) = \int_{-1}^{1}\Phi_i(x)dF(x) \qquad i = 0, 1, 2, \ldots, m-1$$

But

$$\begin{aligned}
\beta_i^m &= \int_{-1}^{1} F_m(x)\varphi_i(x)dx \\
&= \int_{-1}^{1} F_m(x)d\Phi_i(x) \\
&= [F_m(x)\Phi_i(x)\Big|_{-1}^{1}] - \int_{-1}^{1}\Phi_i(x)dF_m(x) \\
&= [F(x)\Phi_i(x)\Big|_{-1}^{1}] - \int_{-1}^{1}\Phi_i(x)dF(x) \\
&= \int_{-1}^{1} F(x)\varphi_i(x)dx \\
&= \beta_i
\end{aligned}$$

and this proves the theorem.

The next theorem provides additional information about the manner in which the function F_m approximates F and also relates p_m and p in a similar way.

Theorem 4.5. The function $F_m - F$ has at least $m-1$ roots in the interval $(-1, 1)$ and the function $p_m - p$ has at least m roots in the interval $(-1, 1)$.

Proof. The function $F_m - F$ is orthogonal to the vector space $\mathcal{P}_{m-1}$ of polynomials of degree less than or equal to $m-1$. This is so because

$$\int_{-1}^{1} \varphi_i(x)[F_m(x)-F(x)]dx = 0 \quad , \quad i = 0, 1, 2, \ldots, m-1$$

and $\{\varphi_0, \varphi_1, \ldots, \varphi_{m-1}\}$ is an orthonormal basis for $\mathcal{P}_{m-1}$. The fact that $F_m - F \perp \mathcal{P}_{m-1}$ is sufficient condition for $F_m - F$ to have at least $m-1$ roots in $(-1, 1)$. Since $F_m(-1) - F(-1) = 0 = F_m(1) - F(1)$ it follows that $F_m - F$ has at least $m+1$ roots on $[-1, 1]$. Denote $m+1$ of the roots of $F_m - F$ by $y_1 < y_2 < \ldots < y_m < y_{m+1}$ where $-1 \leq y_1$ and $y_{m+1} \leq 1$. For $k = 1, 2, \ldots, m$ we have that $(F_m - F)(y_k) = (F_m - F)(y_{k+1})$. By Rolle's Theorem there exists a point x_k such that $y_k < x_k < y_{k+1}$ and $\frac{d}{dx}(F_m - F)(x) = 0$ at $x = x_k$.

This implies that $p_m(x_k|\tau^*) - p(x_k) = 0$ or $p_m(x_k|\tau^*) = p(x_k)$ for $k = 1, 2, \ldots, m$ and $p_m - p$ thus has at least m roots in the interval $(-1, 1)$. This concludes the theorem.

Estimation of the Density p

We now turn our attention to the possibility that p_m might in some sense be a reasonable approximation to the density p. We initiate the proceedings with a lemma.

Lemma 4.2. Let the random variable $\tilde{X}$ be distributed as F on the interval $[-1, 1]$ where

$$F(y) = \int_{-1}^{y} e^{\sum_{i=1}^{\infty} \tau_i \varphi_i(x) - \psi(\tau)} dx \quad , \quad -1 \leq y \leq 1$$

$$= \int_{-1}^{y} p(x)dx$$

and where $p(x)$ satisfies all the assumptions given previously in the introduction. Let the vector $\tau \in R^m$ consist of the first m τ_i's, that is, $\tau = (\tau_1, \tau_2, \ldots, \tau_m)'$. Thus τ is thought of as a vector of fixed parameters. Let $\tau^* \in R^m$ be defined as usual and define θ_j and θ_j' by

$$\theta_j = \int_{-1}^{1} \varphi_j(x) e^{\sum_{i=1}^{\infty} \tau_i \varphi_i(x) - \psi(\tau)} dx \qquad j = 1, 2, 3, \ldots$$

$$\theta_j' = \int_{-1}^{1} \varphi_j(x) e^{\Sigma_{i=1}^{m} \tau_i \varphi_i(x) - \psi_m(\tau)} dx \qquad j = 1, 2, 3, \ldots$$

The vectors θ and $\theta' \in R^m$ are given by $\theta = (\theta_1, \theta_2, \ldots, \theta_m)'$ and $\theta' = (\theta_1', \theta_2', \ldots, \theta_m')'$. Let $\|\cdot\|$ be the usual Euclidean norm on R^m, that is, for $x = (x_1, \ldots, x_m)' \in R^m$ we mean

$$\|x\| = \{\sum_{i=1}^{m} x_i^2\}^{1/2}.$$

Then there exists vectors $\gamma_1, \gamma_2, \ldots, \gamma_m \in R^m$ such that $\|\tau - \tau^*\|^2 \leq \|\theta - \theta'\|^2 \sum_{j=1}^{m} (1/\lambda_j^*)^2$, where λ_j^* is the smallest eigenvalue of $H_{\psi_m}(\gamma_j)$.

Proof. As usual $\nabla\psi_m$ is the gradient of ψ_m and ϕ_m is its continuously differentiable inverse. The function ϕ_m is vector valued so denote it by $\phi_m = (\phi_m^1, \phi_m^2, \ldots, \phi_m^m)'$. Then

$$\begin{aligned}
\tau - \tau^* &= \phi_m(\nabla\psi_m(\tau)) - \phi_m(\nabla\psi_m(\tau^*)) \\
&= \phi_m(\theta') - \phi_m(\theta) \\
&= \begin{bmatrix} \phi_m^1(\theta') - \phi_m^1(\theta) \\ \vdots \\ \phi_m^m(\theta') - \phi_m^m(\theta) \end{bmatrix}
\end{aligned}$$

Now θ' and θ are both in $\mathcal{R}_{\nabla\psi_m}$, the range of $\nabla\psi_m$,

and $\mathcal{R}_{\nabla\psi_m}$ is an open convex set by Lemma 3.10. Then the line segment between θ and θ' is contained in $\mathcal{R}_{\nabla\psi_m}$. By a well known result in advanced calculus, there exist points $p_1, p_2, \ldots, p_m$ all on the segment between θ and θ', such that

$$\phi_m^j(\theta') - \phi_m^j(\theta) = \sum_{i=1}^{m} \frac{\partial \phi_m^j}{\partial \theta_i}(p_j)\,(\theta_1' - \theta_i) \qquad j = 1, 2, \ldots, m$$

Then

$$\tau - \tau^* = \begin{bmatrix} \sum_{i=1}^{m} \frac{\partial \phi_m^1}{\partial \theta_i}(p_i)\,(\theta_i' - \theta_i) \\ \vdots \\ \sum_{i=1}^{m} \frac{\partial \phi_m^m}{\partial \theta_i}(p_m)\,(\theta_i' - \theta_i) \end{bmatrix}$$

Now define the matrix $J_{\phi_m}(z)$ by

$$J_{\phi_m}(z) = \begin{bmatrix} \frac{\partial \phi_m^1}{\partial \theta_i}(z) & \cdots & \frac{\partial \phi_m^1}{\partial \theta_m}(z) \\ \vdots & & \\ \frac{\partial \phi_m^m}{\partial \theta_1}(z) & \cdots & \frac{\partial \phi_m^m}{\partial \theta_m}(z) \end{bmatrix}, \quad z \in \mathcal{R}_{\nabla\psi_m}$$

Since ϕ_m is the inverse of $\nabla\psi_m$, the relation

$$\phi_m(\nabla\psi_m(\tau)) = \tau$$

holds for all $\tau \in R^m$. By the chain rule for differentiation,

$$J_{\phi_m}(\nabla\psi_m(\tau))J_{\nabla\psi_m}(\tau) = J_I(\tau)$$

where $I: R^m \to R^m$ is the identity function. Let I_m represent the m by m identity matrix. Then $J_I(\tau) = I_m$ and we have

$$J_{\phi_m}(\nabla\psi_m(\tau))J_{\nabla\psi_m}(\tau) = I_m$$

This equation immediately implies that

$$J_{\phi_m}(\nabla\psi_m(\tau)) = [J_{\nabla\psi_m}(\tau)]^{-1}$$

$$= H^{-1}_{\psi_m}(\tau)$$

Now each p_j is an element of $\mathcal{R}_{\nabla\psi_m}$ so that there exist vectors $\gamma_j \in R^m$ such that

$$p_j = \nabla\psi_m(\gamma_j) \qquad j = 1, 2, \ldots, m$$

In particular we can write

$$J_{\phi_m}(\nabla\psi_m(\gamma_j)) = J_{\phi_m}(p_j)$$

$$= H^{-1}_{\psi_m}(\gamma_j)$$

Let $e_j = (\delta_{1j}, \delta_{2j}, \ldots, \delta_{mj})'$, $j = 1, 2, \ldots, m$ where

$$\delta_{ij} = \begin{cases} 1, & i = j \\ 0, & i \neq j \end{cases}$$

Then the vector $\tau - \tau^*$ can be expressed as

$$\tau - \tau^* = \begin{bmatrix} e_1' J_{\phi_m}(p_1)(\theta' - \theta) \\ \vdots \\ e_m' J_{\phi_m}(p_m)(\theta' - \theta) \end{bmatrix}$$

$$= \begin{bmatrix} e_1' H^{-1}_{\psi_m}(\gamma_1)(\theta' - \theta) \\ \\ e_m' H^{-1}_{\psi_m}(\gamma_m)(\theta' - \theta) \end{bmatrix}$$

which then implies that

$$\|\tau - \tau^*\|^2 = \sum_{j=1}^{m} \{e_j' H^{-1}_{\psi_m}(\gamma_j)(\theta' - \theta)\}^2$$

$$\leq \sum_{j=1}^{m} \{e_j' H^{-1}_{\psi_m}(\gamma_j) e_j\}\{(\theta' - \theta)' H^{-1}_{\psi_m}(\gamma_j)(\theta' - \theta)\}$$

Now

$$e_j' H^{-1}_{\psi_m}(\gamma_j) e_j \leq e_j' e_j \{\text{largest eigenvalue of } H^{-1}_{\psi_m}(\gamma_j)\}$$

$$(\theta' - \theta)' H^{-1}_{\psi_m}(\gamma_j)(\theta' - \theta) \leq (\theta' - \theta)'(\theta - \theta)\{\text{largest eigenvalue of } H^{-1}_{\psi_m}(\gamma_j)\}$$

and also

$$\{\text{largest eigenvalue of } H_{\psi_m}^{-1}(\gamma_j)\} = \{\text{smallest eigenvalue of } H_{\psi_m}(\gamma_j)\}^{-1}$$

Let λ_j^* be the smallest eigenvalue of $H_{\psi_m}(\gamma_j)$, $j = 1, 2, \ldots, m$. Then since $e_j'e_j = 1$ we have

$$\|\tau - \tau^*\|^2 \leq (\theta' - \theta)'(\theta' - \theta) \sum_{j=1}^{m} (1/\lambda_j^*)^2$$

or

$$\|\tau - \tau^*\|^2 \leq \|\theta' - \theta\|^2 \sum_{j=1}^{m} (1/\lambda_j^*)^2$$

This concludes the lemma.

Lemma 4.3. Suppose the hypotheses of Lemma 4.2 hold and assume further that $\|\tau - \tau^*\| = \theta(m^{-3/2})$. Then $p_m(x|\tau^*) \to p(x)$ uniformly as $m \to \infty$.

Proof. Referring to Lemma 2.2, it would be sufficient to show that

$$\sum_{i=1}^{m} \tau_i^* \varphi_i(x) - \psi_m(\tau^*) \to \sum_{i=1}^{\infty} \tau_i \varphi_i(x) - \psi(\tau)$$

uniformly as $m \to \infty$ or even to show that

$$\sum_{i=1}^{m} \tau_i^* \varphi_i(x) \to \sum_{i=1}^{\infty} \tau_i \varphi_i(x)$$

uniformly as $m \to \infty$.

By hypothesis the series $\sum_{i=1}^{\infty} \tau_i \varphi_i(x)$ converges uniformly on $[-1, 1]$. Then for any $\epsilon > 0$ there exists an integer M such that $m > M$ implies

$$\left| \sum_{i=m+1}^{\infty} \tau_i \varphi_i(x) \right| < \epsilon/2 \quad , \quad -1 \leq x \leq 1$$

Also,

$$\left| \sum_{i=1}^{m} \tau_i \varphi_i(x) - \sum_{i=1}^{m} \tau_i^* \varphi_i(x) \right| \leq \sum_{i=1}^{m} |\tau_i - \tau_i^*| \, |\varphi_i(x)|$$

$$\leq \|\tau - \tau^*\| \sum_{i=1}^{m} |\varphi_i(x)|$$

$$\leq \|\tau - \tau^*\| \; m \sqrt{\frac{2m+1}{2}}$$

since $|\varphi_i(x)| \leq \sqrt{\frac{2i+1}{2}}$. If $\|\tau - \tau^*\| = \mathcal{O}(m^{-3/2})$ then there exists an integer $N > M$ such that if $m > N$ then

$$\|\tau - \tau^*\| \; m \sqrt{\frac{2m+1}{2}} < \frac{\epsilon}{2}$$

Hence if $m > N$ then

$$\left| \sum_{i=1}^{m} \tau_i^* \varphi_i(x) - \sum_{i=1}^{\infty} \tau_i \varphi_i(x) \right| < \epsilon \quad , \quad -1 \leq x \leq 1$$

and so

$$\sum_{i=1}^{m} \tau_i^* \varphi_i(x) \to \sum_{i=1}^{\infty} \tau_i \varphi_i(x) \qquad \text{uniformly as} \quad m \to \infty$$

This concludes the lemma.

Remark. By Theorem 3.2, if $\tau_i = 0$ for $i > S$ where S is some positive integer then the hypotheses of Lemma 4.3 hold.

The next result is stated as a theorem.

Theorem 4.6. Given the same hypotheses as in Lemma 4.2, make the additional assumption that

$$\|\theta' - \theta\| \left\{ \sum_{j=1}^{m} (1/\lambda_j^*)^2 \right\}^{1/2} = \mathcal{O}(m^{-3/2})$$

Then

$$e^{\Sigma_{i=1}^{m} \tau_i^* \varphi_i(x) - \psi_m(\tau^*)} \to e^{\Sigma_{i=1}^{\infty} \tau_i \varphi_i(x) - \psi(\tau)}$$

uniformly in x as $m \to \infty$, that is $p_m(x|\tau^*) \to p(x)$ uniformly.

Proof. Lemma 4.2 and Lemma 4.3.

Corollary 4.6. If $\|\theta'-\theta\|\left\{\sum_{j=1}^{m}(1/\lambda_j^*)^2\right\}^{1/2} = \theta(m^{-3/2})$ then

$$\lim_{m\to\infty}\lim_{n\to\infty} e^{\sum_{i=1}^{m}\tau_{ni}^*\varphi_i(x)-\psi_m(\tau_n^*)} \underset{=}{a.s.} e^{\sum_{i=1}^{\infty}\tau_i\varphi_i(x)-\psi(\tau)}$$

uniformly in x.

Proof. Since $\tau_n^* \xrightarrow{a.s.} \tau^*$ the result follows.

This concludes the chapter on what could be loosely though of as estimation of exponential type. We have assumed that the true underlying density and cumulative distribution function can be represented explicitly in the form

$$p(x) = e^{\sum_{i=1}^{\infty}\tau_i\varphi_i(x)-\psi(\tau)}, \quad -1 \underset{=}{\leq} x \underset{=}{\leq} 1$$

$$F(x) = \int_{-1}^{x} p(y)dy, \quad -1 \underset{=}{\leq} x \underset{=}{\leq} 1$$

We have investigated properties of estimators of $p(x)$ and $F(x)$. The estimators are chosen from the canonical exponential family of distributions generated by $\{\varphi_1, \varphi_2, \ldots, \varphi_m\}$ and the uniform distribution over $[-1, 1]$.

The next chapter deals with estimation of densities using mean integrated squared error as the measure of goodness of the estimate.

V. ESTIMATION USING THE CRITERION OF MEAN INTEGRATED SQUARED ERROR

The topic of this chapter has been covered previously in the literature, notably by Watson and Leadbetter (1963), Schwartz (1967), Kronmal and Tarter (1968) and Watson (1968). It is intended that what follows here will be a modest extension of what has already been published.

The general definition of mean integrated squared error is presented now. It is assumed that a random sample has been obtained from an underlying distribution which has density f and an estimator $\hat{f}$ has been constructed based on the values of the sample. The functions f and $\hat{f}$ are supposed to be squared-integrable with respect to a non-negative weight function w, that is, it is assumed that

$$\int f^2(x)w(x)dx < \infty$$

and

$$\int \hat{f}^2(x)w(x)dx < \infty.$$

Definition of MISE

The mean integrated squared error (abbreviated MISE) is defined to be

$$\text{MISE} = E \int [f(x)-f(x)]^2 w(x)dx$$

The principal case that we will consider is given by the additional assumptions that a complete orthonormal basis with respect to $w(x)dx$ exists and that the density f can be expanded in terms of that basis. In other words if the basis is $\{\phi_i(x)\}_{i=0}^{\infty}$ then f can be written as

$$f(x) = \sum_{i=0}^{\infty} \theta_i \phi_i(x)$$

where

$$\theta_i = \int f(x)\phi_i(x)w(x)dx \qquad i = 0, 1, 2, \ldots$$

To be consistent with the previous chapters we use the Legendre polynomials $\{\varphi_i\}_{i=1}^{\infty}$, which are a complete orthonormal basis on $[-1, 1]$ and $w(x)$ will be identically one on $[-1, 1]$, zero elsewhere.

Suppose the sample is denoted $\{X_1, X_2, \ldots, X_n\}$ and as before we define

$$\bar{\varphi}_i = \frac{1}{n} \sum_{j=1}^{n} \varphi_i(X_j) \qquad i = 0, 1, 2, \ldots$$

We initially consider estimators $\hat{f}_m$ which have the form

$$\hat{f}_m = \sum_{i=0}^{m} \bar{\varphi}_i \varphi_i(x) \qquad (\bar{\varphi}_0 = \varphi_0(x) = \sqrt{\frac{1}{2}})$$

This type of estimator is discussed in Watson, Schwartz, Kronmal and Tarter, among others. The MISE for $\hat{f}_m$ is given by

$$\text{MISE} = E\int_{-1}^{1}[f(x)-\hat{f}_m(x)]^2dx$$

$$= E\int_{-1}^{1}\left\{\sum_{i=0}^{m}(\theta_i-\bar{\varphi}_i)\varphi_i(x) + \sum_{i=m+1}^{\infty}\theta_i\varphi_i(x)\right\}^2 dx$$

$$= E\left\{\sum_{i=0}^{m}(\theta_i-\bar{\varphi}_i)^2 + \sum_{i=m+1}^{\infty}\theta_i^2\right\}$$

$$= E\left\{\sum_{i=1}^{m}(\theta_i-\bar{\varphi}_i)^2 + \sum_{i=m+1}^{\infty}\theta_i^2\right\}$$

since $\varphi_0(x)$ is a constant function. The mean $\bar{\varphi}_i$ is an unbiased estimator of θ_i so

$$\text{MISE} = \sum_{i=1}^{m}\text{Var}(\bar{\varphi}_i) + \sum_{i=m+1}^{\infty}\theta_i^2$$

$$= \frac{1}{n}\sum_{i=1}^{m}\text{Var}(\varphi_i(\tilde{x})) + \sum_{i=m+1}^{\infty}\theta_i^2$$

We denote var $(\varphi_i(\tilde{x}))$ by θ_{ii} and write

$$\text{MISE} = \frac{1}{n}\sum_{i=1}^{m}\theta_{ii} + \sum_{i=m+1}^{\infty}\theta_i^2 \qquad (5.1)$$

Some General Results

The first result we give has been indicated by Kronmal and Tarter to be empirically true for several sampling distributions; we show it to be true under fairly general conditions.

Theorem 5.1. Suppose f is continuous and strictly positive on $[-1, 1]$. For fixed sample size n, the MISE of f_m is a monotone increasing function of m for m sufficiently large. Furthermore, MISE $\to \infty$ as $m \to \infty$ and n is held fixed.

Proof. Let $\text{Min}(f) = \min_{-1 \leq x \leq 1} f(x) > 0$. By Bessel's inequality,

$$\sum_{i=1}^{\infty} \theta_i^2 \leq \int_{-1}^{1} f^2(x)dx < \infty$$

which implies that $\theta_i^2 \to 0$ as $i \to \infty$. Suppose we define the function $F(m, n)$ by

$$F(m, n) = \frac{1}{n} \sum_{i=1}^{m} \theta_{ii} + \sum_{i=m+1}^{\infty} \theta_i^2 \qquad \begin{matrix} m = 1, 2, \ldots \\ n = 1, 2, \ldots \end{matrix}$$

$$= \frac{1}{n} \sum_{i=1}^{m} \{E[\varphi_i(\tilde{x})]^2 - \theta_i^2\} + \sum_{i=m+1}^{\infty} \theta_i^2$$

Now

$$F(m+1,n)-F(m,n) = \frac{1}{n}\sum_{i=1}^{m+1}\{E[\varphi_i(x)]^2-\theta_i^2\} + \sum_{i=m+2}^{\infty}\theta_i^2$$

$$- \frac{1}{n}\sum_{i=1}^{m}\{E[\varphi_i(x)]^2-\theta_i^2\} + \sum_{i=m+1}^{\infty}\theta_i^2$$

$$= \frac{1}{n}\{E[\varphi_{m+1}(x)]^2-\theta_{m+1}^2\} - \theta_{m+1}^2$$

We get an easy lower bound for $E[\varphi_{m+1}(x)]^2$ by

$$E[\varphi_{m+1}(x)]^2 = \int_{-1}^{1}\varphi_{m+1}^2(x)f(x)dx$$

$$\geq \text{Min}(f)\int_{-1}^{1}\varphi_{m+1}^2(x)dx$$

$$= \text{Min}(f) > 0$$

Then

$$F(m+1,n) - F(m,n) \geq \frac{1}{n}\text{Min}(f) - \frac{n+1}{n}\theta_{m+1}^2$$

Since $\theta_{m+1}^2 \to 0$ as $m \to \infty$ and $\frac{1}{n}\text{Min}(f) > 0$ it is clear that $F(m+1,n) > F(m,n)$ for all sufficiently large m. For n fixed it also becomes clear that there must exist at least one value of m for which MISE is a minimum.

Now there exists M such that $m \geq M$ implies

$$F(m+1,n) - F(m,n) \geq \frac{1}{2n}\text{Min}(f)$$

so that for $m \geq M+1$ we have

$$F(m,n) = F(1,n) + \sum_{k=1}^{M-1} [F(k+1,n)-F(k,n)] + \sum_{k=M}^{m-1} [F(k+1,n)-F(k,n)]$$

$$\geq F(1,n) + \sum_{k=1}^{M-1} [F(k+1,n)-F(k,n)] + (m-M)\frac{\text{Min}(f)}{2n}$$

and then it must hold that $F(m,n) \to \infty$ as $m \to \infty$. This proves the theorem.

Theorem 5.2. For m fixed the MISE of $\hat{f}_m$ is a strictly monotone decreasing function of n and in fact

$$\text{MISE} = F(m,n) \to \sum_{i=m+1}^{\infty} \theta_i^2 \quad \text{as} \quad n \to \infty.$$

Proof. Obvious from

$$F(m,n) = \frac{1}{n}\sum_{i=1}^{m} \theta_{ii} + \sum_{i=m+1}^{\infty} \theta_i^2$$

From Theorem 5.1 it is clear that for fixed sample size n there will be at least one optimal value of m as far as MISE is concerned. For sample size n, let M_n^* be an optimal value of m.

Theorem 5.3. For the estimator $\hat{f}_m$ and for any determination of the sequence $M_1^*, M_2^*, \ldots$ we have

$$M_1^* \leq M_2^* \leq \cdots \leq M_n^* \leq \cdots$$

Proof. Looking at second order differences, we have

$$F(m+1,n+1) - F(m,n+1) - F(m+1,n) + F(m,n)$$

$$= \left\{ \frac{1}{n+1} \sum_{i=1}^{m+1} \theta_{ii} + \sum_{i=m+2}^{\infty} \theta_i^2 \right\} - \left\{ \frac{1}{n+1} \sum_{i=1}^{m} \theta_{ii} + \sum_{i=m+1}^{\infty} \theta_i^2 \right\}$$

$$- \left\{ \frac{1}{n} \sum_{i=1}^{m+1} \theta_{ii} + \sum_{i=m+2}^{\infty} \theta_i^2 \right\} + \left\{ \frac{1}{n} \sum_{i=1}^{m} \theta_{ii} + \sum_{i=m+1}^{\infty} \theta_i^2 \right\}$$

$$= \left(\frac{1}{n+1} - \frac{1}{n} \right) \theta_{m+1,m+1}$$

$$< 0$$

This inequality yields the following:

$$0 < F(m,n) - F(m,n+1) < F(m+1,n) - F(m+1,n+1)$$

and

$$F(m,n) - F(m+1,n) < F(m,n+1) - F(m+1,n+1)$$

Since these inequalities hold for all m and n, we can write iteratively

$$F(m,n) - F(m+1,n) < F(m,n+1) - F(m+1,n+1)$$

$$F(m+1,n) - F(m+2,n) < F(m+1,n+1) - F(m+2,n+1)$$

$$\vdots$$

$$F(m+r-1,n) - F(m+r,n) < F(m+r-1,n+1) - F(m+r,n+1)$$

where r is a positive integer. Summing corresponding sides, we obtain

$$F(m,n) - F(m+r,n) < F(m,n+1) - F(m+r,n+1) \qquad (5.2)$$

Now for sample size n the optimal value of m, M_n^*, may or may not be unique. By Theorem 5.1 there can be only a finite number of candidates for M_n^*; call them $M_1^*(n), \ldots, M_k^*(n)$ and suppose $M_1^*(n) < \ldots < M_k^*(n)$. Using the relations developed we have by (5.2) that

$$0 \leq F(m,n) - F(M_k^*(n),n) < F(m,n+1) - F(M_k^*(n),n+1)$$

$$m < M_k^*(n)$$

Thus it must be that

$$M_k^*(n) \leq M_{n+1}^*$$

This means that for any determination of M_n^* and M_{n+1}^* we have $M_n^* \leq M_{n+1}^*$. Thus

$$M_1^* \leq M_2^* \leq M_3^* \leq \ldots \leq M_n^* \leq \ldots$$

whether they are uniquely determined or not. Of course this means

that the sequence $\{M_n^*\}_{i=1}^{\infty}$ is monotone increasing. This concludes the theorem.

At this point we seek to generalize the estimator $\hat{f}_m$ so far considered. Watson suggests the estimator $\hat{g}_m$ of the density f given by

$$\hat{g}_m(x) = \sum_{i=0}^{m} \lambda_i(n)\bar{\varphi}_i\varphi_i(x)$$

where the factors λ_i depend on sample size n and are, ideally, chosen to make MISE a minimum. Watson gives these values of the λ_i by

$$\lambda_i(n) = \frac{\theta_i^2}{\theta_i^2 + \frac{\theta_{ii}}{n}}$$

and clearly $\lambda_i(n) \to 1$ as $n \to \infty$ whenever $\theta_i \neq 0$.

Matrix Estimators

Motivated by Watson's lead, we introduce a class of matrix estimators of which $\hat{f}_m$ and $\hat{g}_m$ are special cases. The general form of the matrix estimator is

$$f(x) = \frac{1}{2} + \underset{\sim}{\bar{\varphi}}'\underset{\sim}{A}\underset{\sim}{\varphi}(x)$$

where $\underset{\sim}{\bar{\varphi}} = (\bar{\varphi}_1, \bar{\varphi}_2, \ldots, \bar{\varphi}_m)'$, $\underset{\sim}{\varphi}(x) = (\varphi_1(x), \varphi_2(x), \ldots, \varphi_m(x))'$ and A is an $m \times m$ real matrix. We note that in particular,

$$\hat{f}_m(x) = \frac{1}{2} + \underset{\sim}{\bar{\varphi}}' I_m \underset{\sim}{\varphi}(x)$$

where I_m is the $m \times m$ identity matrix. Watson's estimator $\hat{g}_m$ is given by

$$\hat{g}_m(x) = \frac{1}{2} + \underset{\sim}{\bar{\varphi}}' \Lambda \underset{\sim}{\varphi}(x)$$

where Λ is diagonal and the ith diagonal element is $\lambda_i(n)$. We propose to find the matrix A which minimizes the MISE.

Suppose the variance-covariance matrix of $\underset{\sim}{\varphi}(x)$ is denoted by $\not{\Sigma}$, that is,

$$\not{\Sigma} = \begin{bmatrix} \theta_{11} & \theta_{12} & \cdots & \theta_{1m} \\ \theta_{21} & \theta_{22} & \cdots & \theta_{2m} \\ \vdots & & & \\ \theta_{m1} & \theta_{m2} & \cdots & \theta_{mm} \end{bmatrix} \tag{5.3}$$

Let $\underset{\sim}{\theta} = (\theta_1, \theta_2, \ldots, \theta_m)'$. The next theorem identifies the optimal matrix A.

Theorem 5.4. For each m the optimal matrix A exists and is given by $A = (\underset{\sim\sim}{\theta\theta}' + \frac{1}{n} \not{\Sigma})^{-1} \underset{\sim\sim}{\theta\theta}'$.

Proof. For an arbitrary matrix A we have

$$\frac{1}{2} + \underset{\sim}{\bar{\varphi}}'\mathrm{A}\underset{\sim}{\varphi}(x) = \frac{1}{2} + \sum_{i=1}^{m}\sum_{j=1}^{m}\bar{\varphi}_i\varphi_j(x)a_{ij}$$

The MISE for the matrix estimator is

$$\begin{aligned}
\mathrm{MISE} &= \mathrm{E}\int_{-1}^{1}[f(x)-(\tfrac{1}{2}+\underset{\sim}{\bar{\varphi}}'\mathrm{A}\underset{\sim}{\varphi}(x)]^2dx \\
&= \mathrm{E}\int_{-1}^{1}\left\{\left[\frac{1}{2}+\sum_{j=1}^{\infty}\theta_j\varphi_j(x)\right] - (\tfrac{1}{2}+\underset{\sim}{\bar{\varphi}}'\mathrm{A}\underset{\sim}{\varphi}(x))\right\}^2 dx \\
&= \mathrm{E}\int_{-1}^{1}\left\{\sum_{j=1}^{\infty}\theta_j\varphi_j(x) - \sum_{i=1}^{m}\sum_{j=1}^{m}\bar{\varphi}_i\varphi_j(x)a_{ij}\right\}^2 dx \\
&= \mathrm{E}\int_{-1}^{1}\left\{\sum_{j=1}^{m}\theta_j\varphi_j(x) - \sum_{j=1}^{m}\left[\sum_{i=1}^{m}\bar{\varphi}_i a_{ij}\right]\varphi_j(x) + \sum_{j=m+1}^{\infty}\theta_j\varphi_j(x)\right\}^2 dx \\
&= \mathrm{E}\sum_{j=1}^{m}\left\{\theta_j - \sum_{i=1}^{m}\bar{\varphi}_i a_{ij}\right\}^2 + \sum_{j=m+1}^{\infty}\theta_j^2
\end{aligned}$$

Let a_j = jth column of A so that $\mathrm{A} = (a_1, a_2, \ldots, a_m)$.

Then the MISE can be written as

$$\mathrm{MISE} = \mathrm{E}\sum_{j=1}^{m}[\theta_j - \underset{\sim}{\bar{\varphi}}'a_j]^2 + \sum_{j=m+1}^{\infty}\theta_j^2$$

$$= \mathrm{E}[(\theta' - \underset{\sim}{\bar{\varphi}}'\mathrm{A})(\theta - \mathrm{A}'\underset{\sim}{\bar{\varphi}})] + \sum_{j=m+1}^{\infty}\theta_j^2 =$$

$$= E[\underset{\sim}{\theta}'\underset{\sim}{\theta} - \underset{\sim}{\bar{\varphi}}'A\underset{\sim}{\theta} - \underset{\sim}{\theta}'A'\underset{\sim}{\bar{\varphi}} + \underset{\sim}{\bar{\varphi}}'AA'\underset{\sim}{\bar{\varphi}}] + \sum_{j=m+1}^{\infty} \theta_j^2$$

$$= \underset{\sim}{\theta}'\underset{\sim}{\theta} - 2\underset{\sim}{\theta}'A\underset{\sim}{\theta} + E\underset{\sim}{\bar{\varphi}}'AA'\underset{\sim}{\bar{\varphi}} + \sum_{j=m+1}^{\infty} \theta_j^2$$

Now

$$E\underset{\sim}{\bar{\varphi}}'AA'\underset{\sim}{\bar{\varphi}} = E \operatorname{tr} \underset{\sim}{\bar{\varphi}}'AA'\underset{\sim}{\bar{\varphi}}$$

$$= E \operatorname{tr} A'\underset{\sim}{\bar{\varphi}}\underset{\sim}{\bar{\varphi}}'A$$

$$= \operatorname{tr} EA'(\underset{\sim}{\bar{\varphi}} - \underset{\sim}{\theta} + \underset{\sim}{\theta})(\underset{\sim}{\bar{\varphi}} - \underset{\sim}{\theta} + \underset{\sim}{\theta})'A$$

$$= \operatorname{tr} E\{A'(\underset{\sim}{\bar{\varphi}} - \underset{\sim}{\theta})(\underset{\sim}{\bar{\varphi}} - \underset{\sim}{\theta})'A + A'\underset{\sim}{\theta}(\underset{\sim}{\bar{\varphi}} - \underset{\sim}{\theta})'A + A'(\underset{\sim}{\bar{\varphi}} - \underset{\sim}{\theta})\underset{\sim}{\theta}'A + A'\underset{\sim}{\theta}\underset{\sim}{\theta}'A\}$$

$$= \operatorname{tr} \{\frac{1}{n} A'\not{\Sigma}A + 0 + 0 + A'\underset{\sim}{\theta}\underset{\sim}{\theta}'A\}$$

$$= \frac{1}{n} \operatorname{tr} A'\not{\Sigma}A + \operatorname{tr} A'\underset{\sim}{\theta}\underset{\sim}{\theta}'A$$

$$= \frac{1}{n} \operatorname{tr} A'\not{\Sigma}A + \operatorname{tr} \underset{\sim}{\theta}'AA'\underset{\sim}{\theta}$$

$$= \frac{1}{n} \operatorname{tr}(a_1, a_2, \ldots, a_m)'\not{\Sigma}(a_1, a_2, \ldots, a_m) + \underset{\sim}{\theta}'AA'\underset{\sim}{\theta}$$

$$= \frac{1}{n} \sum_{j=1}^{m} a_j'\Sigma a_j + \underset{\sim}{\theta}'AA'\underset{\sim}{\theta}$$

Therefore the formula for MISE becomes

$$\text{MISE} = \sum_{j=1}^{\infty} \theta_j^2 - 2\underset{\sim}{\theta}'A\underset{\sim}{\theta} + \frac{1}{n}\sum_{j=1}^{m} a_j'\not{\Sigma}a_j + \underset{\sim}{\theta}'AA'\underset{\sim}{\theta}$$

$$= \sum_{j=1}^{\infty} \theta_j^2 - 2(\underset{\sim}{\theta}'a_1, \underset{\sim}{\theta}'a_2, \ldots, \underset{\sim}{\theta}'a_m)\underset{\sim}{\theta} + \frac{1}{n}\sum_{j=1}^{m} a_j'\not{\Sigma}a_j$$

$$+ (\underset{\sim}{\theta}'a_1, \underset{\sim}{\theta}'a_2, \ldots, \underset{\sim}{\theta}'a_m)(\underset{\sim}{\theta}'a_1, \underset{\sim}{\theta}'a_2, \ldots, \underset{\sim}{\theta}'a_m)'$$

$$= \sum_{j=1}^{\infty} \theta_j^2 - 2\sum_{j=1}^{m} \theta_j \underset{\sim}{\theta}'a_j + \frac{1}{n}\sum_{j=1}^{m} a_j'\not{\Sigma}a_j + \sum_{j=1}^{m} a_j'\underset{\sim\sim}{\theta\theta}'a_j$$

As a consequence of this last line we may write

$$\text{MISE} - \sum_{j=1}^{\infty} \theta_j^2 = \sum_{j=1}^{m} \{a_j'(\underset{\sim\sim}{\theta\theta}' + \frac{1}{n}\not{\Sigma})a_j - 2\theta_j \underset{\sim}{\theta}'a_j\}$$

To find the matrix $A = (a_1, a_2, \ldots, a_m)$ which minimizes the MISE it clearly is sufficient to solve for a_j, $j = 1, 2, \ldots, m$. To do so we must minimize the expression

$$a_j'(\underset{\sim\sim}{\theta\theta}' + \frac{1}{n}\not{\Sigma})a_j - 2\theta_j \underset{\sim}{\theta}'a_j$$

for each j. Suppose we define m functions $f_1, f_2, \ldots, f_m$ by

$$f_j(y) = y'(\underset{\sim\sim}{\theta\theta}' + \frac{1}{n}\not{\Sigma})y - 2\theta_j \underset{\sim}{\theta}'y$$

The matrix $(\underset{\sim\sim}{\theta\theta}' + \frac{1}{n}\not{\Sigma})$ is positive definite and $H_{f_j}(y) = (\underset{\sim\sim}{\theta\theta}' + \frac{1}{n}\not{\Sigma})$,

which implies that each f_j is strictly convex. To minimize $f_j(y)$ it is necessary and sufficient to solve the equation

$$\nabla f_j(y) = 0.$$

But $\nabla f_j(y) = 2(\underset{\sim}{\theta}\underset{\sim}{\theta}' + \frac{1}{n}\not\Sigma)y - 2\theta_j\underset{\sim}{\theta}$ and so the optimal vector y^* is

$$y^* = (\underset{\sim}{\theta}\underset{\sim}{\theta}' + \frac{1}{n}\not\Sigma)^{-1}\theta_j\underset{\sim}{\theta}$$

Consequently the optimal choice of a_j is

$$a_j = \theta_j(\underset{\sim}{\theta}\underset{\sim}{\theta}' + \frac{1}{n}\not\Sigma)^{-1}\underset{\sim}{\theta}_j\theta$$

and the A matrix which minimizes MISE becomes

$$A = (\underset{\sim}{\theta}\underset{\sim}{\theta}' + \frac{1}{n}\not\Sigma)^{-1}\underset{\sim}{\theta}(\theta_1, \theta_2, \ldots, \theta_m)$$

$$= (\underset{\sim}{\theta}\underset{\sim}{\theta}' + \frac{1}{n}\not\Sigma)^{-1}\underset{\sim}{\theta}\underset{\sim}{\theta}'$$

This proves the theorem.

Lemma 5.1. For $A = (\underset{\sim}{\theta}\underset{\sim}{\theta}' + \frac{1}{n}\not\Sigma)^{-1}\underset{\sim}{\theta}\underset{\sim}{\theta}'$ the matrix estimator $\frac{1}{2} + \underset{\sim}{\overline{\varphi}}'A\underset{\sim}{\varphi}(x)$ has

$$\text{MISE} = \sum_{i=1}^{\infty}\theta_i^2 - \underset{\sim}{\theta}'A\underset{\sim}{\theta}.$$

Proof. From the previous theorem the MISE is

$$\text{MISE} = E \sum_{j=1}^{m} \left\{ \theta_j - \sum_{i=1}^{m} (\bar{\varphi}_i a_{ij}) \right\}^2 + \sum_{j=m+1}^{\infty} \theta_j^2$$

which was shown to be equivalent to

$$\text{MISE} = \underset{\sim}{\theta}'\underset{\sim}{\theta} - 2\underset{\sim}{\theta}' A \underset{\sim}{\theta} + E\underset{\sim}{\bar{\varphi}}' AA' \underset{\sim}{\bar{\varphi}} + \sum_{j=m+1}^{\infty} \theta_j^2$$

Also from the last theorem we had

$$E\underset{\sim}{\bar{\varphi}}' AA' \underset{\sim}{\bar{\varphi}} = \frac{1}{n} \text{tr } A' \not{\Sigma} A + \text{tr } \underset{\sim}{\theta}' AA' \underset{\sim}{\theta}$$

$$= \frac{1}{n} \text{tr } A' \not{\Sigma} A + \text{tr } A' \underset{\sim}{\theta}\underset{\sim}{\theta}' A$$

$$= \text{tr } \{A' (\frac{1}{n} \not{\Sigma} + \underset{\sim}{\theta}\underset{\sim}{\theta}') A\}$$

Then the expression for MISE becomes

$$\text{MISE} = \sum_{i=1}^{\infty} \theta_i^2 - 2\underset{\sim}{\theta}' A \underset{\sim}{\theta} + \text{tr}\{A' (\underset{\sim}{\theta}\underset{\sim}{\theta}' + \frac{1}{n} \not{\Sigma}) A\}$$

$$= \sum_{i=1}^{\infty} \theta_i^2 - 2\underset{\sim}{\theta}' (\underset{\sim}{\theta}\underset{\sim}{\theta}' + \frac{1}{n} \not{\Sigma})^{-1} \underset{\sim}{\theta}\underset{\sim}{\theta}' \underset{\sim}{\theta}$$

$$+ \text{tr} \{\underset{\sim}{\theta}\underset{\sim}{\theta}' (\underset{\sim}{\theta}\underset{\sim}{\theta}' + \frac{1}{n} \not{\Sigma})^{-1} (\underset{\sim}{\theta}\underset{\sim}{\theta}' + \frac{1}{n} \not{\Sigma}) (\underset{\sim}{\theta}\underset{\sim}{\theta}' + \frac{1}{n} \not{\Sigma})^{-1} \underset{\sim}{\theta}\underset{\sim}{\theta}'\}$$

$$= \sum_{i=1}^{\infty} \theta_i^2 - 2\underset{\sim}{\theta}' (\underset{\sim}{\theta}\underset{\sim}{\theta}' + \frac{1}{n} \not{\Sigma})^{-1} \underset{\sim}{\theta}\underset{\sim}{\theta}' \underset{\sim}{\theta} + \text{tr}\{\theta\theta' (\theta\theta' + \frac{1}{n} \not{\Sigma})^{-1} \theta\theta'\} =$$

$$= \sum_{i=1}^{\infty} \theta_i^2 - 2\underset{\sim}{\theta}'(\underset{\sim}{\theta}\underset{\sim}{\theta}' + \frac{1}{n}\hat{\Sigma})^{-1}\underset{\sim}{\theta}\underset{\sim}{\theta}'\underset{\sim}{\theta} + \mathrm{tr}\{\underset{\sim}{\theta}'(\underset{\sim}{\theta}\underset{\sim}{\theta}' + \frac{1}{n}\hat{\Sigma})^{-1}\underset{\sim}{\theta}\underset{\sim}{\theta}'\underset{\sim}{\theta}\}$$

$$= \sum_{i=1}^{\infty} \theta_i^2 - 2\underset{\sim}{\theta}'\underset{\sim}{A}\underset{\sim}{\theta} + \underset{\sim}{\theta}'\underset{\sim}{A}\underset{\sim}{\theta}$$

$$= \sum_{i=1}^{\infty} \theta_i^2 - \underset{\sim}{\theta}'\underset{\sim}{A}\underset{\sim}{\theta}$$

Lemma 5.2. As $n \to \infty$, $\sum_{i=1}^{\infty} \theta_i^2 - \underset{\sim}{\theta}'\underset{\sim}{A}\underset{\sim}{\theta} \to \sum_{i=m+1}^{\infty} \theta_i^2$.

Proof. We use a formula for the inverse of $(\underset{\sim}{\theta}\underset{\sim}{\theta}' + \frac{1}{n}\hat{\Sigma})$ which is to be found in Rao (1965):

$$(\underset{\sim}{\theta}\underset{\sim}{\theta}' + \frac{1}{n}\hat{\Sigma})^{-1} = (\frac{1}{n}\hat{\Sigma})^{-1} - \frac{(\frac{1}{n}\hat{\Sigma})^{-1}\underset{\sim}{\theta}\underset{\sim}{\theta}'(\frac{1}{n}\hat{\Sigma})^{-1}}{1+\underset{\sim}{\theta}'(\frac{1}{n}\hat{\Sigma})^{-1}\underset{\sim}{\theta}}$$

Then we have

$$\underset{\sim}{\theta}'\underset{\sim}{A}\underset{\sim}{\theta} = \underset{\sim}{\theta}'(\underset{\sim}{\theta}\underset{\sim}{\theta}' + \frac{1}{n}\hat{\Sigma})^{-1}\underset{\sim}{\theta}\underset{\sim}{\theta}'\underset{\sim}{\theta}$$

$$= \underset{\sim}{\theta}'\left\{(\frac{1}{n}\hat{\Sigma})^{-1} - \frac{(\frac{1}{n}\hat{\Sigma})^{-1}\underset{\sim}{\theta}\underset{\sim}{\theta}'(\frac{1}{n}\hat{\Sigma})^{-1}}{1+\underset{\sim}{\theta}'(\frac{1}{n}\hat{\Sigma})^{-1}\underset{\sim}{\theta}}\right\}\underset{\sim}{\theta}\underset{\sim}{\theta}'\underset{\sim}{\theta}$$

$$= \left\{n\underset{\sim}{\theta}'\hat{\Sigma}^{-1}\underset{\sim}{\theta} - \frac{(n\underset{\sim}{\theta}'\hat{\Sigma}^{-1}\underset{\sim}{\theta})(n\underset{\sim}{\theta}'\hat{\Sigma}^{-1}\underset{\sim}{\theta})}{1+n\underset{\sim}{\theta}'\hat{\Sigma}^{-1}\underset{\sim}{\theta}}\right\}\underset{\sim}{\theta}'\underset{\sim}{\theta}$$

$$= \left\{\frac{n\underset{\sim}{\theta}'\Sigma^{-1}\underset{\sim}{\theta}}{1+n\underset{\sim}{\theta}'\Sigma^{-1}\underset{\sim}{\theta}}\right\}\underset{\sim}{\theta}'\underset{\sim}{\theta} \to \underset{\sim}{\theta}'\underset{\sim}{\theta} \quad \text{as} \quad n \to \infty$$

which proves the lemma.

Next, if A is the optimal choice of matrix with regard to MISE then $\frac{1}{2} + \underset{\sim}{\bar{\varphi}}'A\underset{\sim}{\varphi}(x)$ and $\frac{1}{2} + \underset{\sim}{\bar{\varphi}}'I_m\underset{\sim}{\varphi}(x)$ are shown to be asymptotically equivalent.

Theorem 5.5. For fixed m, we have

$$\mathrm{MISE}\left(\tfrac{1}{2}+\underset{\sim}{\bar{\varphi}}'I_m\underset{\sim}{\varphi}(x)\right) - \mathrm{MISE}\left(\tfrac{1}{2}+\underset{\sim}{\bar{\varphi}}'A\underset{\sim}{\varphi}(x)\right) \to 0 \quad \text{as} \quad n \to \infty,$$

Proof. Immediate from (5.1), (5.3) and Lemmas 5.1, 5.2 and

$$\sum_{i=m+1}^{\infty} \theta_i^2 \leq \mathrm{MISE}\left(\tfrac{1}{2}+\underset{\sim}{\bar{\varphi}}'A\underset{\sim}{\varphi}(x)\right) \leq \mathrm{MISE}\left(\tfrac{1}{2}+\underset{\sim}{\bar{\varphi}}'I_m\underset{\sim}{\varphi}(x)\right)$$

$$= \frac{1}{n}\operatorname{tr} \not{\Sigma} + \sum_{i=m+1}^{\infty} \theta_i^2$$

Remark. From this inequality it is clear that the reduction in MISE is bounded by $\frac{1}{n}\operatorname{tr} \not{\Sigma}$. Thus for very large n the reduction is quite imperceptible but for large m the savings may be significant if n is relatively small. Further work on this would seem to be in order.

Thus far we have considered three matrix type estimates of the density, namely $\frac{1}{2} + \underset{\sim}{\bar{\varphi}}'I_m\underset{\sim}{\varphi}(x)$, $\frac{1}{2} + \underset{\sim}{\bar{\varphi}}'\Lambda\underset{\sim}{\varphi}(x)$, and $\frac{1}{2} + \underset{\sim}{\bar{\varphi}}'A\underset{\sim}{\varphi}(x)$. We introduce still another matrix type estimate of f, which we denote by $\frac{1}{2} + \underset{\sim}{\bar{\varphi}}'B\underset{\sim}{\varphi}(x)$.

We cite a result found in Lindley (1965),

> If the random variable $\tilde{x}$ has a multivariate normal distribution with $E(\tilde{x}) = A\theta$ where A is known but $\theta = (\theta_1, \theta_2, \ldots, \theta_p)'$ is not, and known dispersion matrix C; and if the prior distribution $\tilde{\theta}$ is multivariate normal with $E(\tilde{\theta}) = \underset{\sim}{\mu}_0$ and dispersion matrix C_0; then the posterior distribution $\tilde{\theta}$ is also multivariate normal with mean $\underset{\sim}{\mu}_1 = \{C_0^{-1}+A'C^{-1}A\}^{-1}\{C_0^{-1}\underset{\sim}{\mu}_0+A'C^{-1}x\}$ and dispersion matrix $\{C_0^{-1}+A'C^{-1}A\}^{-1}$.

Applying this result, we notice that $\underset{\sim}{\bar{\varphi}}$ is approximately multivariate normal with mean = $\underset{\sim}{\theta}$ and dispersion matrix = $\frac{1}{n}\Sigma$ (the author is indebted to G. S. Watson for suggesting a Bayesian approach using the approximate normality of $\underset{\sim}{\bar{\varphi}}$). Then if we put a MVN (μ_0, Σ_0) prior on $\tilde{\theta}$, the posterior distribution of $\tilde{\theta}$ is approximately MVN(μ_1, Σ_1), where

$$\underset{\sim}{\mu}_1 = (\Sigma_0^{-1} + n\Sigma^{-1})^{-1}(\Sigma_0^{-1}\underset{\sim}{\mu}_0 + n\Sigma^{-1}\underset{\sim}{\bar{\varphi}})$$

$$\Sigma_1 = (\Sigma_0^{-1} + n\Sigma^{-1})^{-1}$$

The particular prior we wish to consider here is for $\underset{\sim}{\mu}_0 = 0$. Then look at the estimate $\underset{\sim}{\mu}_1$ of $\underset{\sim}{\theta}$ of the form $\underset{\sim}{\mu}_1 = (\Sigma_0^{-1} + n\Sigma^{-1})^{-1} n\Sigma^{-1}\underset{\sim}{\bar{\varphi}}$ and the matrix estimator (of the density) $\frac{1}{2} + \underset{\sim}{\mu}_1'\underset{\sim}{\varphi}(x) = \frac{1}{2} + \underset{\sim}{\bar{\varphi}}'B\underset{\sim}{\varphi}(x)$ where

$$B = n\Sigma^{-1}(\Sigma_0^{-1} + n\Sigma^{-1})^{-1}$$

$$= (\tfrac{1}{n}\Sigma)^{-1}(\Sigma_0^{-1} + n\Sigma^{-1})^{-1}$$

$$= [(\Sigma_0^{-1} + n\Sigma^{-1})(\tfrac{1}{n}\Sigma)]^{-1}$$

$$= (I + \tfrac{1}{n}\Sigma_0^{-1}\Sigma)^{-1}$$

Clearly $B \to I$ as $n \to \infty$, so for large samples $\frac{1}{2} + \underset{\sim}{\bar{\varphi}}' B \underset{\sim}{\varphi}(x)$ is approximately the same as $\frac{1}{2} + \underset{\sim}{\bar{\varphi}}' I_m \underset{\sim}{\varphi}(x)$; thus the Bayesian estimator has the same MISE properties as the other matrix estimators considered, when sample size is large.

As an analogue to real valued series, it is not hard to show that for n sufficiently large, one may write

$$(I + \Sigma_0^{-1}\Sigma/n)^{-1} = I - (\Sigma_0^{-1}\Sigma/n) + (\Sigma_0^{-1}\Sigma/n)^2 - (\Sigma_0^{-1}\Sigma/n)^3 + \ldots$$

$$= \sum_{k=0}^{\infty} (-1)^k (\Sigma_0^{-1}\Sigma/n)^k$$

Then additionally

$$\tfrac{1}{2} + \underset{\sim}{\bar{\varphi}}'(I + \Sigma_0^{-1}\Sigma/n)^{-1} \underset{\sim}{\varphi}(x) = \tfrac{1}{2} + \underset{\sim}{\bar{\varphi}}' B \underset{\sim}{\varphi}(x)$$

$$= \tfrac{1}{2} + \underset{\sim}{\bar{\varphi}}' \left\{ \sum_{k=0}^{\infty} (-1)^k (\Sigma_0^{-1}\Sigma/n)^k \right\} \underset{\sim}{\varphi}(x) \quad =$$

$$= \frac{1}{2} + \sum_{k=0}^{\infty} (-1)^k \underset{\sim}{\bar{\varphi}}'(\not{\Sigma}_0^{-1}\not{\Sigma}/n)^k \underset{\sim}{\varphi}(x)$$

$$= \frac{1}{2} + \underset{\sim}{\bar{\varphi}}' I_m \underset{\sim}{\varphi}(x) + \sum_{k=1}^{\infty} (-1)^k \underset{\sim}{\bar{\varphi}}'(\not{\Sigma}_0^{-1}\not{\Sigma}/n)^k \underset{\sim}{\varphi}(x)$$

$$= \frac{1}{2} + \underset{\sim}{\bar{\varphi}}' I_m \underset{\sim}{\varphi}(x) - \underset{\sim}{\bar{\varphi}}'(\not{\Sigma}_0^{-1}\not{\Sigma}/n) \sum_{k=1}^{\infty} (-1)^{k-1} (\not{\Sigma}_0^{-1}\not{\Sigma}/n)^{k-1} \underset{\sim}{\varphi}(x)$$

$$= \frac{1}{2} + \underset{\sim}{\bar{\varphi}}' I_m \underset{\sim}{\varphi}(x) - \underset{\sim}{\bar{\varphi}}'(\not{\Sigma}_0^{-1}\not{\Sigma}/n)(I+\not{\Sigma}_0^{-1}\not{\Sigma}/n)^{-1} \underset{\sim}{\varphi}(x)$$

Thus we can state the following corollary.

Corollary 5.5.1. For each (fixed) value of m,

$$\Delta = \{[\frac{1}{2} + \underset{\sim}{\bar{\varphi}}'(I+\not{\Sigma}_0^{-1}\not{\Sigma}/n)^{-1} \underset{\sim}{\varphi}(x)] - [\frac{1}{2} + \underset{\sim}{\bar{\varphi}}' I_m \underset{\sim}{\varphi}(x)]\} = \mathcal{O}(\frac{1}{n})$$

uniformly for $-1 \leq x \leq 1$.

Proof. $\Delta = \underset{\sim}{\bar{\varphi}}'(\not{\Sigma}_0^{-1}\not{\Sigma}/n)(I+\not{\Sigma}_0^{-1}\not{\Sigma}/n)^{-1} \underset{\sim}{\varphi}(x)$. The proof of the corollary follows from $\underset{\sim}{\bar{\varphi}} \xrightarrow{a.s.} \underset{\sim}{\theta}$, $(I+\not{\Sigma}_0^{-1}\not{\Sigma}/n)^{-1} \to I$, and the continuity of $\underset{\sim}{\varphi}(x)$.

Thus the Bayesian matrix estimator has appeal from a large simple point of view. One more result which may or may not have computational advantage in the following:

Corollary 5.5.2. (Bayesian estimator with remainder term):

For $N = 1, 2, 3, \ldots$ we have

$$\frac{1}{2} + \underset{\sim}{\bar{\varphi}}'(I+\Sigma_0^{-1}\Sigma/n)^{-1}\underset{\sim}{\varphi}(x) = \frac{1}{2} + \sum_{j=0}^{N}(-1)^j \underset{\sim}{\bar{\varphi}}'(\Sigma_0^{-1}\Sigma/n)^j \underset{\sim}{\varphi}(x) + R_N(x)$$

where

$$R_N(x) = (-1)^{N+1}\underset{\sim}{\bar{\varphi}}'(\Sigma_0^{-1}\Sigma/n)^{N+1}(I+\Sigma_0^{-1}\Sigma/n)^{-1}\underset{\sim}{\varphi}(x).$$

Proof. Follows from

$$\frac{1}{2} + \underset{\sim}{\bar{\varphi}}'(I+\Sigma_0^{-1}\Sigma/n)\underset{\sim}{\varphi}(x) = \frac{1}{2} + \sum_{j=0}^{N}(-1)^j \underset{\sim}{\bar{\varphi}}'(\Sigma_0^{-1}\Sigma/n)^j \underset{\sim}{\varphi}(x)$$

$$+ \sum_{j=N+1}^{\infty}(-1)^j \underset{\sim}{\bar{\varphi}}'(\Sigma_0^{-1}\Sigma/n)^j \underset{\sim}{\varphi}(x)$$

and

$$\sum_{j=N+1}^{\infty}(-1)^j \underset{\sim}{\bar{\varphi}}'(\Sigma_0^{-1}\Sigma/n)^j \underset{\sim}{\varphi}(x) = (-1)^{N+1}\underset{\sim}{\bar{\varphi}}'(\Sigma_0^{-1}\Sigma/n)^{N+1}$$

$$\times \sum_{j=N+1}^{\infty}(-1)^{j-(N+1)}(\Sigma_0^{-1}\Sigma/n)^{j-N+1}\underset{\sim}{\varphi}(x)$$

$$= (-1)^{N+1}\underset{\sim}{\bar{\varphi}}'(\Sigma_0^{-1}\Sigma/n)^{N+1}(I+\Sigma_0^{-1}\Sigma/n)^{-1}\underset{\sim}{\varphi}(x)$$

$$= R_N(x).$$

This concludes the chapter on estimation of densities using mean integrated squared error as criterion for goodness. A few pertinent comments might be in order. First and foremost, it can be noted that every result in this chapter holds true for a continuous positive density on an interval $[a,b]$ and practically any complete orthonormal basis $\{\varphi_i\}_{i=0}^{\infty}$ for $L_2[a,b]$. Secondly, as far as we know, no researcher has ever done a comparative Monte Carlo study of these matrix type estimators, chiefly, we imagine, because as far as we know, they are hitherto now unknown. Such a study would be interesting.

BIBLIOGRAPHY

Apostol, Tom M. 1960. Mathematical analysis. Reading, Mass., Addison-Wesley. 559 p.

Barndorff-Nielsen, Ole. 1970. Exponential families, exact theory. Aarhus, Denmark, Aarhus Universitet, Various Publication Series No. 19.

Bhattacharya, P.K. 1967. Estimation of a probability density function and its derivatives. Sankhya, ser. A, part 4, 29:373-382.

Buck, R. Creighton. 1965. Advanced calculus. New York, Mcgraw-Hill. 527 p.

Buehler, Robert J. 1965. The limit of the nth power of a density. The Annals of Mathematical Statistics 36:1878-1882.

Cacoullos, Theophilos. 1966. Estimation of a multivariate density. Annals of the Institute of Statistical Mathematics 18:179-190.

Čencov, N.N. 1962. Evaluation of an unknown distribution density from observations. Soviet Math. 3:1559-1562.

Cheney, E.W. 1966. Introduction to approximation theory. New York, McGraw-Hill. 259 p.

Churchill, Ruel V. 1963. Fourier series and boundary value problems. New York, McGraw-Hill. 248 p.

Dawid, A.P. 1970. On the limiting normality of posterior distributions. Proc. Camb. Phil. Soc. 67:625-633.

Elkins, T.A. 1968. Cubical and spherical estimation of a multivariate probability density. Journal of the American Statistical Association 63:1495-1513.

Jackson, Dunham. 1941. Fourier series and orthogonal polynomials. Carus Mathematical Monograph No. 6. Menasha, Wisc., Banta. 234 p.

Johnson, Richard A. 1970. Asymptotic expansions associated with posterior distributions. The Annals of Mathematical Statistics 41:851-864.

Kronmal, R. and Michael Tarter. 1968. The estimation of probability densities and cumulatives by Fourier series methods. Journal of the American Statistical Association 63:925-952.

Lindley, D. V. 1965. Introduction to probability and statistics. Part 2. Inference. London, Cambridge University Press. 292 p.

Loftsgaarden, D. O. and C. P. Quensenberry. 1965. A nonparametric estimate of a multivariate density function. The Annals of Mathematical Statistics 38:1261-1265.

Morgan, Ronnie L. 1969. A class of conjugate prior distributions, and optimal allocation. Unpublished Ph. D. Thesis at Univ. Missouri.

Neyman, Jerzy. 1937. 'Smooth' test for goodness of fit. Skandinavisk Aktuarietidskrift 20:149-199.

Parzen, Emanuel. 1962. On estimation of a probability density function and mode. The Annals of Mathematical Statistics 33:1065-1076.

Pickards, James III. 1969. Efficient estimation of a probability density function. The Annals of Mathematical Statistics 40:854-864.

Rao, C. R. 1965. Linear statistical inference and its applications. New York, Wiley. 522 p.

Rockafellar, R. Tyrrell. 1970. Convex analysis. Princeton, Princeton University Press. 451 p.

Rosenblatt, Murray. 1956. Remarks on some nonparametric estimates of a density function. The Annals of Mathematical Statistics 27:832-837.

Schwartz, Stuart C. 1967. Estimation of probability density by an orthogonal series. The Annals of Mathematical Statistics 38:1262-1265.

Van Ryzin, J. 1969. On strong consistency of density estimates. The Annals of Mathematical Statistics 40:1765-1772.

Walker, A.M. 1969. On the asymptotic behavior of posterior distributions. Journal of the Royal Statistical Society, ser. B, 31:80-88.

Watson, G.S. 1969. Density estimation by orthogonal series. The Annals of Mathematical Statistics 40:1496-1498.

Watson, G.S. and M.R. Leadbetter. 1963. On the estimation of the probability density, I. The Annals of Mathematical Statistics 34:480-491.

Wegman, Edward J. 1969. Nonparametric probability density estimation. Univ. of N. Carolina Institute of Statistics, Mimeo Series No. 638.

Wilks, Samuel S. 1962. Mathematical statistics. New York, Wiley. 644 p.

Zangwill, Willard I. 1969. Nonlinear programming: a unified approach. Englewood Cliffs, N.J., Prentice-Hall. 356 p.