Sieves for Nonparametric Estimation of Densities and Regressions

by

Stuart Geman

Reports in Pattern Analysis No. 99

Division of Applied Mathematics
Brown University
Providence, Rhode Island 02912

January, 1981

Research supported by the Department of the Army under contract DAAG29-80-K-0006.

Sieves for Nonparametric Estimation of Densities and Regressions

Stuart Geman
Division of Applied Mathematics
Brown University
Providence, Rhode Island 09212

I. Introduction

This report is about the use of least squares for nonparametric regression, and the use of maximum likelihood for nonparametric density estimation. Typically, these classical techniques will fail when applied to infinite dimensional problems. Grenander's method of sieves is a method for modifying classical estimators so as to make them appropriate for nonclassical problems (see Grenander [10], Geman and Hwang [8], Geman [9]). Examples will be given here of the application of this method to problems of regression and density estimation.

The difficulties encountered in moving from finite to infinite dimensional estimation are well illustrated by the failure of maximum likelihood in nonparametric density estimation. Let $x_1,\ldots x_n$ be an i.i.d. sample from an absolutely continuous distribution with unknown density function, $\alpha_0(x)$. The maximum likelihood estimator for α_0 maximizes

$$\prod_{i=1}^{n} \alpha(x_i) \qquad \text{(I.1)}$$

over some specified set of candidates: a set of "allowable" densities. But if this set is too large, then the method will fail to produce a meaningful estimator. For instance, in the extreme case, nothing is known about α_0 and the maximum of (I.1) is not achieved. Roughly speaking, we move out of the parameter space (the space of all densities), approaching a discrete distribution with jumps at the sample points.

Another example of the failure of classical methods to solve infinite dimensional problems is the breakdown of least squares in the nonparametric estimation of a regression. Let X and Y be random variables and let $(x_1,y_1),\ldots(x_n,y_n)$ be an i.i.d. sample from the bivariate distribution of (X,Y). The least squares estimator of the regression function, $E[Y|X=x]$, minimizes

$$\sum_{i=1}^{n}(y_i-\alpha(x_i))^2.$$

Observe: the minimum is 0, and is achieved by any (allowable) function which passes through all of the points of observation, $(x_1,y_1),\ldots(x_n,y_n)$. Excepting some very special cases, this set will not converge to the true regression.

Grenander (see [10]) suggests that we attempt our optimization (maximization of the likelihood, minimization of the sum of square errors) within a subset of the parameter space, and then allow this subset to "grow" with the sample size. This sequence of subsets from which the estimator is drawn is called a "sieve", and the resulting estimation procedure is the "method of sieves". The method leads easily to consistent nonparametric estimators in even the most general settings, with different sieves giving rise to different estimators. Often, the sieve estimator is closely related to an already well-studied estimator, and may suggest an improvement, or a new point of view and a new motivation. I believe that this report (taken together with [10] and [8]) gives good evidence for the very broad application of the method of sieves. I hope that it also indicates the range of interesting and mostly unanswered questions raised by our as yet very preliminary study.

Section II is about sieves that make least squares work for nonparametric regression, and section III about sieves that make maximum likelihood work for nonparametric density estimation. In most cases, an explicit asymptotic growth rate for the sieve will be derived, which rate guarantees consistent estimation. But the practical problem of choosing an appropriate sieve size for a given finite collection of observations is still largely unresolved. Section IV discusses one possible solution: the method of "cross-validation".

(This is a "progress report" on the method of sieves; it is not a final manuscript intended for publication. In particular, no attempt is made to meaningfully relate this work to the large body of relevant work by other authors.)

II. Least Squares Nonparametric Regression

A. Hermite Functions

Let us suppose that we have observed n pairs of numbers, $(x_1,y_1),\ldots(x_n,y_n)$. Think of $x_1,\ldots x_n$ as observations of an "independent" variable X, and $y_1,\ldots y_n$ as observations of a "dependent" random variable Y. The regression problem is to estimate the mean value of Y, which is assumed to be a function of X defined on a prescribed interval (possibly the entire line). Depending on whether or not X is a random variable, we distinguish two classes of regression problems. Observe that if $x_1,\ldots x_n$ is a nonrandom ("design") sequence, then without smoothness assumptions, the regression

$$\alpha_0(x) \equiv E_x[Y]$$

is not identifiable. Because, in the absence of a continuity condition for α_0, observations of Y at a predetermined countable set of X values do not give information about α_0 at X values not in the observation set. The problem of estimating α_0 when X is nonrandom and α_0 is sufficiently smooth will be discussed in subsection C below.

This, and the following subsection (B), are about least squares estimators for the regression

$$\alpha_0(x) \equiv E[Y|X=x]$$

when (X,Y) is a bivariate random variable. We will see that if $(x_1,y_1),\ldots(x_n,y_n)$ are i.i.d. observations from the bivariate distribution of (X,Y), then we can construct least squares estimators, $\hat{\alpha}_n$, which are consistent in the sense that

$$\|\alpha_n-\alpha_0\|^2 \equiv E_X|\hat{\alpha}_n(X)-\alpha_0(X)|^2 \qquad \text{(II.A.1)}$$

$$= \int_{-\infty}^{\infty}|\hat{\alpha}_n(x)-\alpha_0(x)|^2F_X(dx) \to 0 \quad \text{a.s.,}$$

where F_X is the marginal distribution of X, and "a.s." is with respect to the distribution on the observations $(x_1,y_1),(x_2,y_2),\ldots$. In words, with probability one our estimator will converge, in the $L_2(F_X)$ metric, to the true regression α_0. No smoothness assumption for α_0 is necessary, and only the (obviously necessary) regularity condition

$$E_Y|Y|^2 = \int_{-\infty}^{\infty} |y|^2F_Y(dy) < \infty$$

needs to be assumed. The conclusion, then, is quite similar to Stone's [16], who constructed a class of estimators which, under the same conditions, are consistent in the sense that

$$E\ E_X|\hat{\alpha}_n(X)-\alpha_0(X)|^2 \to 0$$

where E is with respect to the observations $(x_1,y_1),(x_2,y_2),\ldots$. Although the discussion will be by example, it should be clear that the approach is a general one, and that by merely substituting one sieve for another, we have available an unlimited variety of estimators.

Let us now consider a particular solution, by the method of sieves, to the nonparametric regression problem with random X variables. We wish to estimate

$$\alpha_0(x) = E[Y|X=x]$$

from a random sample $(x_1,y_1),\dots(x_n,y_n)$ of the bivariate distribution of (X,Y). If we use a least squares estimate[1], then some modification of the classical procedure must be introduced, as was demonstrated in the Introduction. A sieve is a sequence of sets of functions, indexed by the sample size, from which the estimator is drawn. As a simple example, consider the following sieve consisting of truncated Fourier expansions in Hermite functions:

$$S_n = \{\alpha(x):\alpha(x) = \sum_{k=0}^{m_n} a_k f_k(x),\ \sum_{k=0}^{m_n} |a_k| \le \lambda_n\} \qquad \text{(II.A.2)}$$

where

$$f_n(x) = x^n e^{-x^2/2}/\{\int_{-\infty}^{\infty} (y^n e^{-y^2/2})^2 dy\}^{1/2} \qquad n=0,1,2,\dots$$

and m_n and λ_n are increasing sequences, to be specified more precisely later. (The Hermite functions themselves are the functions obtained from Gram-Schmidt orthonormalization of $\{f_n\}$ $n=0,1,\dots$.) Given this sieve, the (least squares) method of sieves estimator, $\hat{\alpha}_n$, is defined by

$$\text{minimize} \sum_{i=1}^{n} (y_i-\alpha(x_i))^2 \text{ subject to } \alpha \in S_n.$$

Since the solution may not be unique, it is convenient to work with, instead, the set of least squares estimators:

$$A_n = \{\alpha \in S_n: \sum_{i=1}^{n} (y_i-\alpha(x_i))^2 = \inf_{\beta \in S_n} (\sum_{i=1}^{n} (y_i-\beta(x_i))^2\}.$$

[1] The method of sieves will apply as well to other approaches. For example, the analysis is essentially the same if the square error is replaced by some other, perhaps more "robust", criterion.

We wish to show that when m_n and λ_n increase to infinity sufficiently slowly, the set A_n converges to the target parameter, α_0, in the sense that

$$\sup_{\alpha \in A_n} \|\alpha-\alpha_0\| \to 0 \quad \text{a.s.}$$

($\|\ \|$ as defined in (II.A.1)). For now, let us assume only that $E|Y|^2 < \infty$. Equivalently, we assume

$$\alpha_0 \in L_2(R^1,B^1,F_X)$$

(B^1 is the collection of Borel sets in R^1). It is evident that in the absence of further assumptions, the sieve S_n must have the following property: for an arbitrary distribution function F_X, and an arbitrary $\alpha \in L_2(R^1,B^1,F_X)$, there exists a sequence $\{\beta_n\}$ such that

1. $\beta_n \in S_n$ $n=1,2,\dots,$ and
2. $\|\beta_n-\alpha\| \to 0$ as $n \to \infty$.

Otherwise, there would exist an α_0 such that no estimator drawn from S_n could be consistent. For the particular sieve (II.A.2), it is evidently sufficient that the Hermite-type functions $f_0,f_1,\dots$ span a dense set in $L_2(R^1,B^1,F_X)$ (for arbitrary distribution function, F_X):

Lemma 1. The linear span of $\{f_n\}_{n=0}^{\infty}$ is dense in $L_2(R^1,B^1,F_X)$.

(Proofs are deferred to the Appendix.)

Hence, we know that no matter how m_n and λ_n increase to infinity, S_n gets arbitrarily close (and may eventually contain) α_0. For now, let us imagine that we have fixed sequences m_n and λ_n increasing to infinity, and let $\{\beta_n\}$ be a sequence of functions satisfying

1. $\beta_n \in S_n$ $n=1,2,\ldots,$ and
2. $\|\beta_n-\alpha_0\| \to 0$ as $n \to \infty$.

Why should we expect the set A_n to converge to α_0? For the intuitive reason, observe first that

$$E[(Y-\alpha(X))^2]$$

is minimized when $\alpha = \alpha_0$. Since $\beta_n \to \alpha_0$:

$$E[(Y-\beta_n(X))^2] \to E[(Y-\alpha_0(X))^2],$$

and we can expect that

$$\sup_{\alpha\in A_n} E[(Y-\alpha(X))^2]$$

$$\simeq \text{ (with luck, by the LLN) } \sup_{\alpha\in A_n} \frac{1}{n}\sum_{i=1}^{n}(y_i-\alpha(x_i))^2$$

$$\leq \frac{1}{n}\sum_{i=1}^{n}(y_i-\beta_n(x_i))^2$$

$$\approx \text{ (again, by the LLN) } E[(Y-\beta_n(X))^2]$$

$$\to E[(Y-\alpha_0(X))^2]. \qquad \text{(II.A.3)}$$

In other words, we expect

$$\sup_{\alpha\in A_n} E[(Y-\alpha(X))^2] - E[(Y-\alpha_0(X))^2] \to 0.$$

But observe that

$$\sup_{\alpha\in A_n} E[(Y-\alpha(X))^2]-E[(Y-\alpha_0(X))^2]$$

$$= \sup_{\alpha\in A_n} E|\alpha(X)-\alpha_0(X)|^2 = \sup_{\alpha\in A_n} \|\alpha-\alpha_0\|^2,$$

and so, it should be that

$$\sup_{\alpha\in A_n} \|\alpha-\alpha_0\| \to 0$$

as $n \to \infty$.

The only gap in the argument is the use, twice, of an "LLN". It should be clear that with S_n growing slowly enough (i.e. with m_n and λ_n increasing slowly enough) (II.A.3) can be made precise. In fact:

Theorem 1. If $E[e^{t_0|Y|}] < \infty$ for some $t_0 > 0$, and if $m_n\uparrow\infty$ and $\lambda_n\uparrow\infty$ with $m_n = O(n^{1-\epsilon})$ and $\lambda_n = O(n^{\delta})$ for some $\epsilon\in(0,1)$ and $\delta\in(0,\epsilon/4)$, then

$$\sup_{\alpha\in A_n} \|\alpha-\alpha_0\| \to 0 \quad \text{a.s.}$$

as $n \to \infty$.

The moment condition, $E[e^{t_0|Y|}] < \infty$, is stronger than necessary. As I have said, all that is necessary is that $E|Y|^2 < \infty$, although this weaker condition leads to a slower growth for the sieve parameters m_n and λ_n. The proof is the same, except that the Chebyshev bound on large deviations (Chebyshev's inequality) must be substituted for the stronger exponential bound now used to prove theorem 1 (see lemma 3, appendix), the latter being in force only when $E[e^{t_0|Y|}] < \infty$ for some $t_0 > 0$.

B. Spline Functions

Different sieves can lead to very different estimators. Return to the problem discussed in subsection A, but with the added assumption that F_X concentrates on $[0,1]$. Consider the sieve

$$S_n = \{\alpha(x):\ \alpha(x),\ \frac{d}{dx}\alpha(x),\ldots \frac{d^{m-1}}{dx^{m-1}}\alpha(x) \text{ continuous},$$
$$\frac{d^m}{dx^m}\alpha(x) \text{ piecewise continuous},\ \int_0^1 |\frac{d^m}{dx^m}\alpha(x)|^2 dx \le \lambda_n\}$$

where $m \ge 1$ is a fixed integer and λ_n increases to infinity with n. Let us again use least squares, defining the estimator $\hat{\alpha}_n$ (for $n \ge m$) by

$$\text{minimize} \sum_{i=1}^{n}(y_i-\alpha(x_i))^2 \quad \text{subject to} \quad \alpha \in S_n. \qquad \text{(II.B.1)}$$

(If $\inf_{\alpha\in S_n} \sum_{i=1}^{n}(y_i-\alpha(x_i))^2 = 0$, then $\hat{\alpha}_n$ may not be uniquely defined. In this case, define $\hat{\alpha}_n$ to be the (unique) element of S_n which minimizes

$$\int_0^1 |\frac{d^m}{dx^m}\alpha(x)|^2 dx$$

subject to $\sum_{i=1}^{n}(y_i-\alpha(x_i))^2 = 0$.)

The solution to (II.B.1) is the well-studied 2m-1 degree polynomial spline, with knots at the observation points $x_1,\ldots x_n$.

That is, $\hat{\alpha}_n$ has 2m-2 continuous derivatives on $[0,1]$, $\frac{d^{2m-1}}{dx^{2m-1}}\hat{\alpha}_n(x)$ is piecewise continuous with discontinuities at $x_1,\ldots x_n$, and $\hat{\alpha}_n$ is represented by a 2m-1 degree polynomial in each interval (x_i,x_{i+1}) $i=0,1,\ldots n$ (defining $x_0=0$ and $x_{n+1}=1$). If $\lambda_n\uparrow\infty$, then $S_1 \subseteq S_2 \subseteq \ldots$ etc., and $\bigcup_{n=1}^{\infty} S_n$ is dense in $L_2([0,1],B^1,F_X)$. Hence, for the same reasons discussed in the previous section, we should expect that $\hat{\alpha}_n \to \alpha_0$, provided that λ_n increases sufficiently slowly. Indeed, for m=1 (for example):

Theorem 2. If $E[e^{t_0|Y|}] < \infty$ for some $t_0 > 0$, and if $\lambda_n\uparrow\infty$ with $\lambda_n = 0(n^{1/4-\varepsilon})$ for some $\varepsilon > 0$, then

$$\|\hat{\alpha}_n-\alpha_0\| \to 0 \quad \text{a.s.}$$

as $n \to \infty$.

Again, only $E|Y|^2 < \infty$ is necessary, but this entails a slower rate of growth for the sieve S_n, i.e. a slower rate of increase for the sieve parameter, λ_n.

C. Dirichlet Kernel

The method is just as easily applied when the X, or "independent", variable is deterministic. In the nonparametric problem, we think of the distribution on Y as being an unknown function of x, $F_x(\cdot)$, with x taking values in some prescribed interval. Let us take, for example, $x\in[0,1]$. The problem then is to estimate

$$\alpha_0(x) = E_x[Y] \equiv \int_{-\infty}^{\infty} yF_x(dy) \qquad x\in[0,1]$$

from independent observations $y_1,\ldots y_n$, where $y_i \sim F_{x_i}$, and $x_1,\ldots x_n$ is a deterministic, so-called design, sequence. In other words, for each $i=1,2,\ldots n$ we make an observation, y_i, from the distribution F_{x_i}, and from these observations we wish to estimate the mean of Y as a function of x. For a specific example, let us assume that the design sequence, for fixed n, is equally spaced on the interval [0,1]:

$$x_j = \frac{j}{n} \qquad j=1,2,\ldots n.$$

Here again, an unconstrained minimization of the sum of square errors

$$\sum_{i=1}^{n} (y_i-\alpha(x_i))^2$$

fails to produce a useful estimator. Introduce a sieve: the "Fourier sieve"

$$S_n = \{\alpha(x):\alpha(x) = \sum_{k=-m_n}^{m_n} a_k e^{2\pi ikx}\}$$

is particularly tractable, and makes for a good illustration of the method in this setting. The sieve size is governed by the parameter m_n, which will be allowed to increase to infinity with n. If we restrict m_n so that $m_n \leq n$ for all n, then $\hat{\alpha}_n$ is uniquely defined by requiring that it

$$\text{minimize} \sum_{i=1}^{n} (y_i-\alpha(x_i))^2 \quad \text{subject to } \alpha\in S_n.$$

The $L_2(dx)$ norm provides a natural metric for this problem:

$$\|\alpha-\beta\|^2 = \int_0^1 |\alpha(x)-\beta(x)|^2 dx,$$

and using this we get

Theorem 3. If

A1. there exists a constant K such that

$$|\alpha_0(x)-\alpha_0(y)| \leq K|x-y| \qquad \forall\ x,y\in[0,1]$$

and

A2. $$\sup_{x\in[0,1]} \int_{-\infty}^{\infty} y^2F_x(dy) < \infty ,$$

then for any sequence $m_n\uparrow\infty$ such that $m_n/n \to 0$ and $m_n \leq n$,

$$E\int_0^1 |\hat{\alpha}_n(x)-\alpha_0(x)|^2dx = O(\frac{1}{m_n} + \frac{m_n}{n} + \frac{1}{\sqrt{n}})$$

as $n\to\infty$. In particular, if $m_n \sim \sqrt{n}$, then

$$E\int_0^1 |\hat{\alpha}_n(x)-\alpha_0(x)|^2dx = O(\frac{1}{\sqrt{n}})$$

as $n\to\infty$.

In analogy to subsections A and B, one might expect, instead, the conclusion

$$\int_0^1 |\hat{\alpha}_n(x)-\alpha_0(x)|^2dx \to 0 \quad \text{a.s.}$$

Indeed, if we follow the analysis of the previous subsections, but with "$F_X(dx)$" replaced by "dx", then we are led to exactly this conclusion, but under a different growth condition for m_n (which condition depends on the moment assumptions we are willing to make for Y).

What does the least squares estimator, $\hat{\alpha}_n$, look like? A simple calculation gives the explicit form:

$$\hat{\alpha}_n(x) = \frac{1}{n}\sum_{i=1}^{n} y_i D_{m_n}(x-x_i)$$

where D_m is the Dirichlet kernel

$$D_m(x) = \sum_{|k|\leq m} e^{-2\pi ikx} = \frac{\sin \pi(2m+1)x}{\sin \pi x} .$$

Here, then, the least squares (sieve) estimator turns out to be a kernel estimator. Kernel estimators for nonparametric regression have been widely studied, although from a somewhat different point of view. See [1],[5],[7],[13] and [14] for some recent examples. It is not too difficult to now exploit this simple form for $\hat{\alpha}_n$, and say a good deal more about its behavior. Let

$$V(x) = \int_{-\infty}^{\infty} (y-\alpha_0(x))^2 F_x(dy),$$

the variance of Y at x. Then:

Theorem 4. If

A1. $\alpha_0(0) = \alpha_0(1)$,

A2. $\alpha_0(x)$ has a continuous derivative, $\alpha_0(x)'$, and for some constant k_1

$$|\alpha_0(x)'-\alpha_0(y)'| \leq k_1|x-y| \quad \forall\ x,y\in[0,1],$$

A3. there exists a constant, k_2, such that

$$|V(x)-V(y)| \leq k_2|x-y| \quad \forall\ x,y\in[0,1],$$

and

A4.

$$\sup_{x\in[0,1]} \int_{-\infty}^{\infty} y^4\, F_x(dy) < \infty ,$$

then for any sequence $m_n\uparrow\infty$, such that $m_n = 0(n^{\beta})$ for some $\frac{1}{4} < \beta < \frac{1}{2}$, the process

$$p_n(t) \equiv \sqrt{n}\int_0^t (\hat{\alpha}_n(x)-\alpha_0(x))dx$$

converges weakly on [0,1] to the diffusion, p(t), defined by

$$dp(t) = \sqrt{V(t)}\, dW_t, \quad p(0) = 0$$

where W_t is standard Brownian motion.

The condition $\alpha_0(0) = \alpha_0(1)$ is awkward, but unfortunately can not be removed. It is a consequence of the sieve, S_n, which admits only functions which are continuous on the unit torus. A sieve closely related to S_n, but perhaps more natural (in the absence of the assumption $\alpha_0(0) = \alpha_0(1)$), is

$$S_n' = \{\alpha(x):\alpha(x) = \sum_{k=-m_n}^{m_n} a_k \cos[k \text{ arc } \cos(2x-1)]\}$$

i.e. replace the trigonometric polynomials by the Chebyshev polynomials. Here we would want to choose a design sequence which preserves the orthogonality of the basis sequence:

$$x_j = \frac{1}{2} + \frac{1}{2}\cos[2j-1)\pi/2n], \quad j=1,2,\ldots n.$$

Theorems 3 and 4 undoubtedly have their analogues for S_n' as well.

Still a good deal more can be said about the estimator $\hat{\alpha}_n$. With suitable restrictions on the growth of m_n, we can establish: pointwise convergence ($E|\hat{\alpha}_n(x)-\alpha_0(x)|^2 \to 0$ for each $x\in(0,1)$); pointwise asymptotic normality; and a relation between the smoothness of α_0 and the rate at which $E\int_0^1 |\hat{\alpha}_n(x)-\alpha_0(x)|^2dx$ converges to zero.

III. Nonparametric Density Estimation by Maximum Likelihood

In the Introduction, I discussed the difficulties which arise when one attempts nonparametric density estimation by direct application of maximum likelihood. One solution is to introduce a sieve, a particularly interesting example of which is the "convolution sieve", suggested by Chii-Ruey Hwang:

$$S_n = \{\alpha(x):\ \alpha(x) = \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi\sigma_n^2}}\, e^{-(x-y)^2/2\sigma_n^2}\, F(dy)$$

$$F \text{ a probability distribution function}\}$$

where σ_n is a sequence decreasing to zero as n increases to infinity. The maximum likelihood sieve estimator solves the problem

$$\text{maximize} \quad \prod_{i=1}^{n} \alpha(x_i) \quad \text{subject to} \quad \alpha \in S_n.$$

We do not yet know whether the solution to this problem is unique, so let us define A_n to be the set of maximum likelihood solutions:

$$A_n = \{\alpha \in S_n:\ \prod_{i=1}^{n} \alpha(x_i) = \sup_{\beta\in S_n} \prod_{i=1}^{n} \beta(x_i)\}.$$

As it turns out, the elements of A_n have a simple and familiar form:

Proposition 1 (Geman and McClure). A_n is not empty, and $\alpha\in A_n$ $\Rightarrow$

$$\alpha(x) = \sum_{j=1}^{n} p_j \frac{1}{\sqrt{2\pi\sigma_n^2}}\, e^{-(x-y_j)^2/2\sigma_n^2}$$

for some $y_1,\ldots y_n$ and $p_1,\ldots p_n$ satisfying $p_i \geq 0 \quad 1 \leq i \leq n$, $\sum_{i=1}^{n} p_i = 1$. Furthermore, if $\min(x_1,\ldots x_n) < \max(x_1,\ldots x_n)$, then $\min(x_1,\ldots x_n) < \min(y_1,\ldots y_n) \leq \max(y_1,\ldots y_n) < \max(x_1,\ldots x_n)$.

Recall the Parzen-Rosenblatt kernel estimator (with Gaussian kernel):

$$\hat{\beta}_n(x) = \frac{1}{n}\sum_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma_n^2}} e^{-(x-x_i)^2/2\sigma_n^2}. \qquad \text{(III.1)}$$

Observe that $\hat{\beta}_n \in S_n$, and is therefore potentially in the maximum likelihood set, A_n. However, the last statement in the proposition indicates that $\hat{\beta}_n$ is *not* in fact among the maximum likelihood solutions. This observation suggests that one may be able to improve on the performance of the kernel estimator by allowing the locations, and possibly the weights, of the kernels to move in such a way as to increase the likelihood. Although some preliminary experiments have been promising, we have not yet fully explored this possibility.

Although we have characterized the maximum likelihood set A_n up to the 2n parameters $y_1,\ldots y_n$, $p_1,\ldots p_n$, its actual computation is difficult. Proposition 1 suggests a smaller and computationally more attractive sieve:

$$S_n^2 = \{\alpha(x)\colon \alpha(x) = \frac{1}{n}\sum_{j=1}^{n} \frac{1}{\sqrt{2\pi\sigma_n^2}} e^{-(x-y_j)^2/2\sigma_n^2}\}$$

i.e. we give equal mass to each kernel, but allow the locations to move in such a way as to maximize the likelihood. Here again, it is easy to show that for $\alpha \in A_n^2$ (the maximum likelihood set)

$$\min(x_1,\ldots x_n) < \min(y_1,\ldots y_n) \leq \max(y_1,\ldots y_n) < \max(x_1,\ldots x_n)$$

provided $\min(x_1,\ldots x_n) < \max(x_1,\ldots x_n)$, and so, again, the kernel estimator is not among the maximum likelihood solutions. We have not yet systematically experimented with S_n^2, but the simulations have so far been quite interesting. As a rule, we have found that the number of *distinct* y's in a maximum likelihood solution is considerably smaller than n. In other words, the kernels will often coalesce to achieve an increased likelihood. Sometimes this results in strikingly accurate density estimators, while at other times this "maximum likelihood" solution is a very poor second to the corresponding (same σ) kernel estimator. The estimator suffers the very same stability problem as the kernel estimator: the results are critically dependent on the choice of σ , the kernel width. Whereas we will be able to specify asymptotic rates of decrease for σ_n which guarantee consistent estimation (more on this below), these rates tell us nothing about an appropriate choice for σ when faced with a particular finite sample.

One promising solution to the problem of choosing an appropriate finite sample σ (or, more generally, an appropriate sieve size) is the method of cross-validation, discussed in section IV, below. Another approach is to include σ as a free parameter within the sieve, and thus allow it to be chosen by maximum likelihood. However, we can not simply define the sieve to be

$$\hat{S}_n^2 = \{\alpha(x):\ \alpha(x) = \frac{1}{n}\sum_{j=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}}\ e^{-(x-y_j)^2/2\sigma^2}\}$$

and leave σ arbitrary, since then the maximum of the likelihood is achieved with $\sigma = 0$ and the kernels centered at the sample points. Let us instead define a sieve parameter $m_n < n$ to be the number of kernels, and consider

$$S_n^3 = \{\alpha(x):\ \alpha(x) = \frac{1}{m_n}\sum_{j=1}^{m_n} \frac{1}{\sqrt{2\pi\sigma^2}}\ e^{-(x-y_j)^2/2\sigma^2}\}\ ,$$

choosing $y_1,\ldots y_{m_n}$ and σ by maximum likelihood, and letting $m_n \uparrow \infty$ with the sample size. See section IV for some preliminary simulation results on this sieve.

Let us now discuss the question of whether there exist rates of growth for these sieves which guarantee consistent estimation. That such rates do in fact exist can be established by arguments somewhat similar to those used for the regression estimators discusses in section II (subsections A and B). First, look at the general case: For each value of some parameter λ define a collection of density functions S_λ. A_λ is then defined to be the set of maximum likelihood solutions within S_λ:

$$A_\lambda = \{\alpha \in S_\lambda:\ \prod_{i=1}^{n} \alpha(x_i) = \sup_{\beta \in S_\lambda} \prod_{i=1}^{n} \beta(x_i)\}.$$

We wish to show that there exists a sequence $\lambda_n \to 0$ such that, for any α_0 ("true" density), $A_{\lambda_n} \to \alpha_0$ in some suitable metric. (Think of λ as σ in S_n and S_n^2, and as $\frac{1}{m}$ in S_n^3.) Loosely, what follows is a general recipe for identifying such a sequence, λ_n.

What we will wind up showing is that the "conditional" entropies of the maximum likelihood estimators $(\hat{\alpha}_n)$,

$$-\int_{-\infty}^{\infty} \alpha_0(x) \log \hat{\alpha}_n(x)dx,$$

approach the formal entropy

$$-\int_{-\infty}^{\infty} \alpha_0(x) \log \alpha_0(x)dx,$$

as $n \to \infty$. The following proposition tells us that this is indeed a meaningful notion of convergence:

<u>Proposition 2</u>. Let $\alpha_0(x)$ be a density function satisfying $\int_{-\infty}^{\infty}\alpha_0(x)\log \alpha_0(x)dx < \infty$. If, for each n, A_n is a collection of density functions, and if

$$\lim_{n\to\infty} \sup_{\alpha \in A_n} \int_{-\infty}^{\infty} \alpha_0(x) \log \frac{\alpha_0(x)}{\alpha(x)}\, dx = 0,$$

then also

$$\lim_{n\to\infty} \sup_{\alpha \in A_n} \int_{-\infty}^{\infty} |\alpha(x)-\alpha_0(x)|dx = 0.$$

Now suppose that we define a sequence $\lambda_n \to 0$ such that, for any α_0,

1. there exists a sequence $\{\beta_n\}$ with

   a. $\beta_n \in S_{\lambda_n}$ $\quad n=1,2,\ldots$

   b. $\int_{-\infty}^{\infty} \alpha_0(x) \log \frac{\alpha_0(x)}{\beta_n(x)}\, dx \to 0$ as $n \to \infty$,

and

2. $\sup_{\alpha\in S_{\lambda_n}} \left|\frac{1}{n}\sum_{i=1}^{n} \log \alpha(x_i) - \int_{-\infty}^{\infty} \alpha_0(x)\log \alpha(x)dx\right| \to 0$ a.s. (III.2)

as $n \to \infty$.

Then we can reason that

$$\overline{\lim_{n\to\infty}} \sup_{\alpha\in A_{\lambda_n}} \int_{-\infty}^{\infty} \alpha_0(x) \log \frac{\alpha_0(x)}{\alpha(x)} dx$$

$$= \text{(a.s.) } \overline{\lim_{n\to\infty}} \sup_{\alpha\in A_n} \left\{ \int_{-\infty}^{\infty} \alpha_0(x)\log \alpha_0(x)dx - \frac{1}{n}\sum_{i=1}^{n} \log \alpha(x_i)\right\}$$

$$\leq \overline{\lim_{n\to\infty}} \left\{ \int_{-\infty}^{\infty} \alpha_0(x)\log \alpha_0(x)dx - \frac{1}{n}\sum_{i=1}^{n} \log \beta_n(x_i)\right\}$$

$$= \text{(a.s.) } \overline{\lim_{n\to\infty}} \int_{-\infty}^{\infty} \alpha_0(x)\log \frac{\alpha_0(x)}{\beta_n(x)} dx = 0.$$

Since $\int_{-\infty}^{\infty} \alpha_0(x)\log \frac{\alpha_0(x)}{\alpha(x)} dx$ is never negative (Jensen's inequality), we conclude, using proposition 2, that

$$\lim_{n\to\infty} \sup_{\alpha\in A_{\lambda_n}} \int_{-\infty}^{\infty} |\alpha(x)-\alpha_0(x)|dx = 0 \quad \text{a.s.}$$

For a specific example, return to S_n^2:

Lemma 5. Let $\alpha_0(x)$ be a density function with support on $[0,1]$.

(a) For any sequence $\alpha_n \downarrow 0$ such that $n\sigma_n \to \infty$, there exists a sequence of density functions, β_n, such that $\beta_n \in S_n^2$ for all n, and

$$\int_0^1 \alpha_0(x)\log \beta_n(x)dx \to \int_0^1 \alpha_0(x)\log \alpha_0(x)dx$$

(finite or infinite),

and (b) for any sequence $\sigma_n \downarrow 0$ such that for some $\varepsilon > 0$ $n^{1/4-\varepsilon}\sigma_n \to \infty$,

$$\lim_{n\to\infty} \sup_{\alpha\in A_n^2} \left|\frac{1}{n}\sum_{i=1}^{n} \log \alpha(x_i) - \int_0^1 \alpha_0(x)\log \alpha(x)dx\right| = 0 \quad \text{a.s.}$$

And hence:

Theorem 5. If $\alpha_0(x)$ has support on $[0,1]$, and if

$$\int_0^1 \alpha_0(x)\log \alpha_0(x)dx < \infty ,$$

then for any sequence $\sigma_n \downarrow 0$ such that for some $\varepsilon > 0$ $n^{1/4-\varepsilon}\sigma_n \to \infty$

$$\lim_{n\to\infty} \sup_{\alpha\in A_n^2} \int_{-\infty}^{\infty} |\alpha(x)-\alpha_0(x)|dx = 0 \quad \text{a.s.}$$

A very similar approach can be used to calculate suitable growth rates for the sieves S_n and S_n^3 as well. (Although, for S_n^3 one must introduce a slight modification into the argument, since (III.2) is not true for any sequence m_n. One approach is to first lower bound the asymptotic rate of decrease of σ in the maximum likelihood set A_n^3. Then define a new, equivalent, sieve $\hat{S}_n^3$, such that eventually $A_n^3 \subseteq \hat{S}_n^3$ but within which σ must respect this lower bound. After this, everything can proceed as before, but applied instead to $\hat{S}_n^3$.)

IV. Cross-validated Sieves

As with the kernel estimator, $\hat{\beta}_n$ (see III.1), the maximum likelihood set A_n^2, drawn from the sieve S_n^2,is extremely sensitive to the choice of σ. Recall that in order to avoid an arbitrary choice for σ , a new sieve (S_n^3) was introduced in which σ is chosen by maximum likelihood. Although a suitable asymptotic rate of growth for m_n, the sieve parameter governing the size of S_n^3, can be identified, there is still the problem of choosing good values of m for finite samples. Indeed, in one form or another, this problem faces all nonparametric estimators of densities and regressions. For the method of sieves, it is the problem of choosing a proper sieve size. For the kernel estimators (of regressions and densities), it is the problem of choosing the right kernel width. For the so-called penalized maximum likelihood estimators, it is the problem of choosing an appropriate weight to be given the penalty function. In these, and in all cases, the problem is one of choosing the right degree of smoothing, when given finite data for an infinite dimensional estimation.

In numerous settings, the method of cross-validation (and so-called, "generalized cross-validation") has proven to be an effective data driven technique for choosing an appropriate degree of smoothing for nonparametric estimators. Many authors have demonstrated the utility of this approach, mostly through simulations and applications to real data (see, especially the work of Wahba and coworkers). The method has its natural application to sieves, as I will describe below. But before entering a discussion of cross-validated sieves, it should be pointed out that although there has been a good deal of theoretical work on cross-validation (again, mostly due to Wahba and coworkers), many of the most basic questions, such as consistency, remain unanswered. This would appear to be a particularly exciting and promising area for future theoretical research.

For the purpose of introducing the method, let us consider again the (Gaussian) kernel estimator for nonparametric density estimation:

$$\hat{\beta}_n(x) = \frac{1}{n}\sum_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-(x-x_i)^2/2\sigma^2} . \qquad \text{(IV.1)}$$

If we observe a particular sample $x_1,\ldots x_n$, how are we to choose σ in forming this estimate? Theoretical results giving optimal asymptotic values for σ are not very helpful; they invariably require a knowledge of the true underlying density α_0. Consider the kernel estimator formed from the data after removing one point, say the j^{th} point:

$$\hat{\beta}_n^j(x) = \frac{1}{n-1}\sum_{i\neq j} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-(x-x_i)^2/2\sigma^2} .$$

One measure of the quality of this estimator is its ability to predict the excluded data point, x_j. And, one measure of this ability is the likelihood of x_j under $\hat{\beta}_n^j$, i.e. $\hat{\beta}_n^j(x_j)$. Consider now the "pseudolikelihood"

$$L_\sigma = \prod_{i=1}^{n} \hat{\beta}_n^i(x_i)$$

which, given the observations $x_1,\ldots x_n$, depends only on σ . For each σ , L_σ is a natural measure of the appropriateness of σ for the observed data. The method of cross-validation chooses σ to maximize L_σ, and then uses this σ in the kernel estimator, (IV.1). Our simulations strongly support this method for choosing a window width, but we have been unable to answer even the most basic question: is the resulting estimator consistent?

I will use two specific examples to illustrate the application of cross-validation to the choice of sieve size. First, look again at the spline sieve defined in II.B (for some fixed $m \geq 1$):

$$S = \{\alpha(x):\alpha(x),\ \frac{d}{dx}\ (x),\ldots \frac{d^{m-1}}{dx^{m-1}}\alpha(x) \text{ continuous},$$

$$\frac{d^m}{dx^m}\alpha(x) \text{ piecewise continuous},\ \int_0^1 \left|\frac{d^m}{dx^m}\alpha(x)\right|^2 dx \leq \lambda\}.$$

Here, the "smoothing parameter" is λ, and we seek a rational method for choosing λ when given a particular finite sample $(x_1,y_1),\ldots(x_n,y_n)$. Fix λ and define $\hat{\alpha}_n^j(x)$ to be the solution to

$$\text{minimize} \quad \sum_{i\neq j}(y_i-\alpha(x_i))^2 \quad \text{subject to} \quad \alpha\in S.$$

The expression

$$E_\lambda \equiv \sum_{i=1}^{n}(y_i-\hat{\alpha}_n^i(x_i))^2$$

depends only on λ . Choose λ^* to minimize E_λ, and then define $\hat{\alpha}_n$ as the least squares estimator in S, using $\lambda = \lambda^*$. λ^* is the "cross-validated smoothing parameter", and $\hat{\alpha}_n$ is the cross-validated estimator.

For a second example, consider again the sieve S^3 for nonparametric density estimation:

$$S^3 = \{\alpha(x):\ \alpha(x) = \frac{1}{m}\sum_{j=1}^{m}\frac{1}{\sqrt{2\pi\sigma^2}}\ e^{-(x-y_j)^2/2\sigma^2}\},$$

for which m must be chosen before the estimator can be constructed. m can be cross-validated in much the same way as σ was for the kernel estimator. For fixed m, define $\hat{\alpha}_n^j(x)$ as the solution to

$$\text{maximize} \quad \prod_{i\neq j}\alpha(x_i) \quad \text{subject to } \alpha\in S^3,$$

and then define m^* $(1 \leq m^* \leq n-1)$ to maximize

$$\prod_{i=1}^{n}\hat{\alpha}_n^i(x_i).$$

Some very preliminary experiments, with extremely small sample sizes, have been encouraging. Random samples of 20, 15, and 10 observations were drawn from three different ("true") densities, α_0. In each case the cross-validated estimator $\hat{\alpha}_n$ was computed using the S^3 sieve. In the first experiment

$$\alpha_0(x) = \frac{1}{\sqrt{2\pi}}\ e^{-2x^2} + \frac{1}{\sqrt{2\pi}}\ e^{-2(x-1)^2}$$

and n=20. Figure 1 shows the true density (solid line) together with the corresponding estimator (broken line). The cross-validated value for m was 1, and this reflects the unimodal character of the underlying density. Figure 2 shows the result of cross-validating m on a sample of size 15 from the density

$$\alpha_0(x) = \frac{1.25}{\sqrt{2\pi}} e^{-x^2/.32} + \frac{2.5}{\sqrt{2\pi}} e^{-(x-1)^2/.08}.$$

The bimodal character of the data is reflected in the cross-validated value m=2. (The x's at the bottom of the figure locate the actual observations.) Finally, figure 3 shows the results of an experiment with n=10 sample points drawn from the density

$$\alpha_0(x) = \frac{1}{1.2\sqrt{2\pi}} e^{-x^2/.32} + \frac{1}{.3\sqrt{2\pi}} e^{-(x-1)^2/.08}.$$

The cross-validated m was 1, which is not consistent with the bimodal density α_0. Observe, however, that this small sample did not reflect the bimodal distribution.

It is obvious that many more experiments need to be run, particularly with larger samples and less regular densities.

References

1. Ahmad, I.A. and Lin, P.E. Nonparametric sequential estimation of a multiple regression function. Bulletin of Mathematical Statistics, Vol. 17, 63-75, 1976.

2. Billingsley, P. Convergence of Probability Measures. John Wiley & Sons, New York, 1968.

3. Chernoff, H. A measure of asymptotic efficiency for tests of a hypothesis based on the sum of observations. Annals of Mathematical Statistics, Vol. 23, 493-507, 1952.

4. Chung, K.L. A Course in Probability Theory. Academic Press, New York, 1974.

5. Devroye, L.P. and Wagner, T.J. Distribution-free consistency results in nonparametric discrimination and regression function estimation. Annals of Statistics, Vol. 8, 231-239, 1980.

6. Dym, H. and McKean, H.P. Fourier Series and Integrals. Academic Press, New York, 1972.

7. Gasser, T. and Muller, H.G. Kernel estimation of regression functions. In: Smoothing Techniques for Curve Estimation. Ed. Gasser and Rosenblatt, Lecture Notes in Mathematics, Springer-Verlag, 1979.

8. Geman, S. and Hwang, C.R. Nonparametric maximum likelihood estimation by the method of sieves. Annals of Statistics (to appear).

9. Geman, S. An application of the method of sieves: functional estimator for the drift of a diffusion. In: Proceedings of the Janos Bolyai Mathematical Society, 1980 Colloquium on Nonparametric Statistical Inference. North-Holland Publishing Co. (to appear).

10. Grenander, U. Abstract Inference. John Wiley & Sons, New York, 1981.

11. Helmberg, G. Introduction to Spectral Theory in Hilbert Space. North-Holland, Amsterdam, 1969.

12. Karlin, S.J. and Studden, W.J. Tchebycheff Systems: With Applications in Analysis and Statistics. John Wiley & Sons, New York, 1966.

13. Schuster, E. and Yakowitz, S. Contributions to the theory of nonparametric regression, with application to system identification. Annals of Statistics, Vol. 7, 139-149, 1979.

14. Spiegelman, C. and Sacks, J. Consistent window estimation in nonparametric regression. Annals of Statistics, Vol. 8, 240-246, 1980.

15. Stein, E.M. Singular Integrals and Differentiability Properties of Functions. Princeton University Press, 1970.

16. Stone, C.J. Consistent nonparametric regression. Annals of Statistics, Vol. 5, 595-645, 1977.

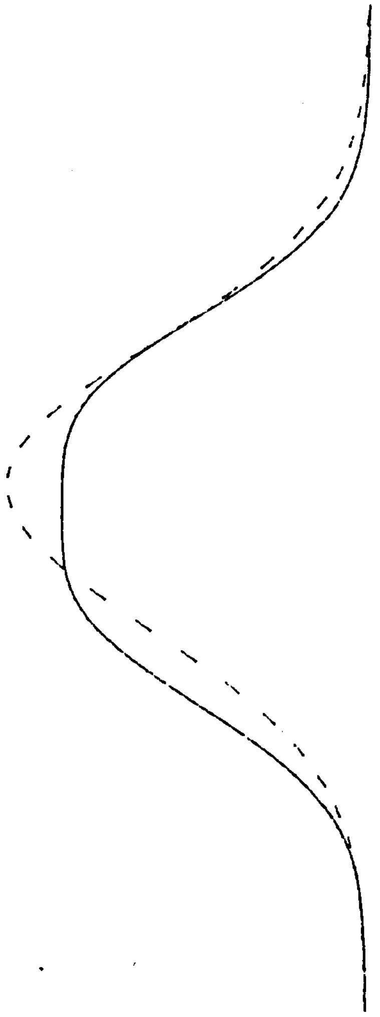


Figure 1

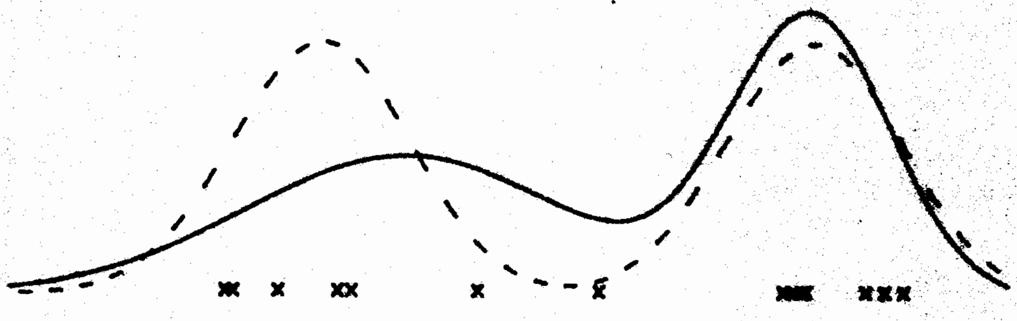
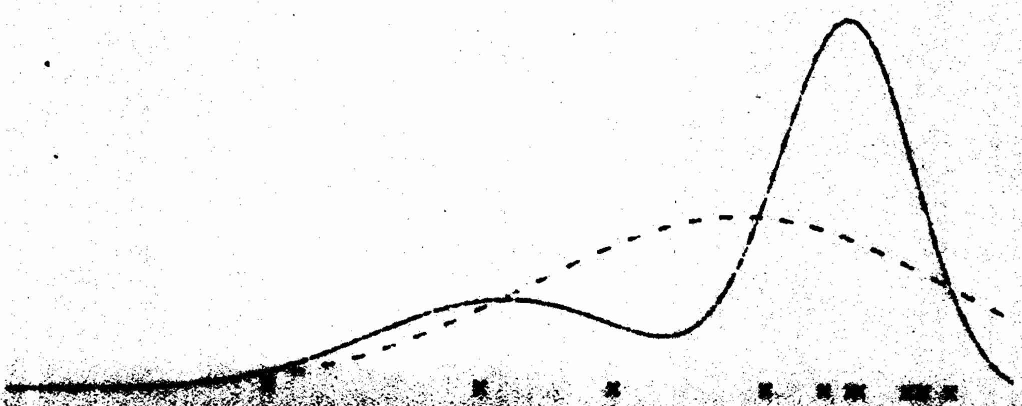


Figure 2



Appendix

II. Least Squares Nonparametric Regression

A. Hermite Functions

Notation:

1. X, Y R^1-valued random variables
2. $(x_1,y_1),(x_2,y_2),\ldots$ i.i.d. observations of (X,Y)
3. $\alpha_0(x) = E[Y|X=x]$
4. $f_n(x) = x^n e^{-x^2/2}/\{\int_{-\infty}^{\infty} (y^n e^{-y^2/2})^2 dy\}^{1/2}$ $n=0,1,2,\ldots$
5. $S_n = \{\alpha(x): \alpha(x) = \sum_{k=0}^{m_n} a_k f_k(x),\ \sum_{k=0}^{m_n} |a_k| \le \lambda_n\}$
6. $A_n = \{\alpha \in S_n: \sum_{i=1}^{n} (y_i-\alpha(x_i))^2 = \inf_{\beta\in S_n} \sum_{i=1}^{n} (y_i-\beta(x_i))^2\}$
7. F_X Distribution function of X
8. $\|\alpha-\beta\| = \sqrt{E|\alpha(X)-\beta(X)|^2} = \{\int_{-\infty}^{\infty} |\alpha(x)-\beta(x)|^2 F_X(dx)\}^{1/2}$

Lemma 1. The linear span of $\{f_n\}_{n=0}^{\infty}$ is dense in $L_2(R^1,B^1,F_X)$.

Proof: (Similar to the argument showing that the linear span of $\{f_n\}_{n=0}^{\infty}$ is dense in $L_2(R^1,B^1,\lambda)$ when λ is Lebesgue measure, c.f. Helmburg [11], section 9, Theorem 5.)

Let $f \in L_2(R^1,B^1,F_X)$. It is enough to show that if

$$\int_{-\infty}^{\infty} f(x)x^n e^{-\frac{1}{2}x^2} dx = 0 \qquad (A.1)$$

$\forall n = 0,1,2,\ldots$ then $f=0$. Suppose (A.1) is true. Then

$$\int_{-\infty}^{\infty} f(x)e^{-\frac{1}{2}x^2} e^{inx} F_X(dx) \qquad (A.2)$$

$$= \int_{-\infty}^{\infty} f(x)e^{-\frac{1}{2}x^2} \sum_{k=0}^{\infty} \frac{(inx)^k}{k!} F_X(dx)$$

$$= \sum_{k=0}^{\infty} \frac{(in)^k}{k!} \int_{-\infty}^{\infty} f(x)e^{-\frac{1}{2}x^2} x^k F_X(dx) = 0.$$

Fix $\varepsilon > 0$ and choose r such that

$$\int_{|x|\ge r} |f(x)|e^{-\frac{1}{2}x^2} F_X(dx) < \varepsilon .$$

Define

$$g(x) = \begin{cases} 1 & \text{when } f(x) > 0 \\ 0 & \text{when } f(x) = 0 \\ -1 & \text{when } f(x) < 0, \end{cases}$$

and choose

$$t(x) = \sum_{k=-m}^{m} \alpha_k e^{i\pi kx/r}$$

such that

1. $\sup_{x\in R^1} |t(x)| \le 2$, and
2. $\int_{|x|<r} |g(x)-t(x)|^2 F_X(dx) < \varepsilon$.

(This can be done by first approximating $g(x)$ by a continuous function, $\hat{g}(x)$, such that $|\hat{g}(x)| \le 1$ $\forall x \in R^1$, using Lusin's Theorem. Then, $\hat{g}(x)$ can be uniformly approximated, on $[-r,r]$, by a trigonometric polynomial, $t(x)$, using the Stone-Weierstrass

Theorem. Since t(x) is periodic, the bound demanded in 1 above will be achieved when t(x) is sufficiently close to $\hat{g}(x)$ on [-r,r].) Finally,

$$\int_{-\infty}^{\infty}|f(x)|e^{-\frac{1}{2}x^2}F_X(dx)=\int_{-\infty}^{\infty}f(x)g(x)e^{-\frac{1}{2}x^2}F_X(dx)$$

$$=\int_{-\infty}^{\infty}f(x)t(x)e^{-\frac{1}{2}x^2}F_X(dx)+\int_{-\infty}^{\infty}f(x)e^{-\frac{1}{2}x^2}(g(x)-t(x))F_X(dx)$$

= (apply (A.2))

$$\int_{|x|<r}f(x)e^{-\frac{1}{2}x^2}(g(x)-t(x))F_X(dx)$$

$$+\int_{|x|\geq r}f(x)e^{-\frac{1}{2}x^2}(g(x)-t(x))F_X(dx)$$

$$\leq\{\int_{|x|<r}|f(x)|^2e^{-x^2}F_X(dx)\}^{1/2}\{\int_{|x|<r}|g(x)-t(x)|^2F_X(dx)\}^{1/2}$$

$$+\ 3\int_{|x|\geq r}|f(x)|e^{-\frac{1}{2}x^2}F_X(dx)$$

$$\leq\|f\|\sqrt{\varepsilon}+3\varepsilon\Rightarrow\int_{-\infty}^{\infty}|f(x)|e^{-\frac{1}{2}x^2}F_X(dx)=0$$

$\Rightarrow\ f=0.$ □

Theorem 1. If $E[e^{t_o|Y|}]<\infty$ for some $t_o>0$, and if $m_n\uparrow\infty$ and $\lambda_n\uparrow\infty$ with $m_n=0(n^{1-\varepsilon})$ and $\lambda_n=0(n^{\delta})$ for some $\varepsilon\in(0,1)$ and $\delta\in(0,\varepsilon/4)$, then

$$\sup_{\alpha\in A_n}\|\alpha-\alpha_0\|\to 0\quad\text{a.s.}$$

as n → ∞ .

Proof: Based on the following lemma:

Lemma 2.

$$\sup_{\alpha\in S_n}\left|\frac{1}{n}\sum_{i=1}^{n}(\alpha(x_i)^2-2\alpha(x_i)y_i)-E[\alpha(X)^2-2\alpha(X)Y]\right|\to 0\quad\text{a.s.}$$

Defer the proof of this, for now. Choose $\{\beta_n\}$ n=1,2,...,is a sequence of functions satisfying $\beta_n\to\alpha_0$ and $\beta_n\in S_n$ ∀n (existence of such a sequence is guaranteed by lemma 1). Observe, then, that

$$E[\beta_n(X)^2-2\beta_n(X)Y]\to E[\alpha_0(X)^2-2\alpha_0(X)Y].$$

Fix $\xi>0$. Lemma 2 implies that (with probability one) for all n sufficiently large:

$$\sup_{\alpha\in A_n}E[\alpha(X)^2-2\alpha(X)Y]$$

$$\leq\sup_{\alpha\in A_n}\frac{1}{n}\sum_{i=1}^{n}[\alpha(x_i)^2-2\alpha(x_i)y_i]+\xi$$

$\leq$ (by the definition of A_n)

$$\frac{1}{n}\sum_{i=1}^{n}[\beta_n(x_i)^2-2\beta_n(x_i)y_i]+\xi$$

$$\leq E[\beta_n(X)^2-2\beta_n(X)Y]+2\xi$$

$$\leq E[\alpha_0(X)^2-2\alpha_0(X)Y]+3\xi.$$

But $E[\alpha(X)^2-2\alpha(X)Y] - E[\alpha_0(X)^2-2\alpha_0(X)Y]$

$$= E[\alpha(X)^2-2\alpha(X)\alpha_0(X)+\alpha_0(X)^2] = \|\alpha-\alpha_0\|^2.$$

Hence $\overline{\lim_{n\to\infty}} \sup_{\alpha\in A_n} \|\alpha-\alpha_0\|^2 \leq 3\xi$ a.s.,

and hence the Theorem.

It remains to prove Lemma 2. I will show that, for arbitrary $\xi > 0$,

$$\sum_{n=1}^{\infty} P\{\sup_{\alpha\in S_n} |\tfrac{1}{n} \sum_{i=1}^{n} (\alpha(x_i)^2-2\alpha(x_i)y_i)-E[\alpha(X)^2-2\alpha(X)Y]| > \xi\} \tag{A.3}$$

$$< \infty ,$$

which is sufficient, because of the Borel-Cantelli Lemma.

Let $\alpha_{n,1},\alpha_{n,2},\ldots\alpha_{n,\ell_n}$ denote the functions $\alpha \in S_n$ of the form

$$\alpha(x) = \sum_{k=0}^{m_n} \frac{p_k}{n^2} f_k(x)$$

where $p_0,p_1,\ldots,p_{m_n}$ are (positive or negative) integers. Observe that

$$\ell_n = 0(n^{3m_n})$$

(because $\alpha \in S_n$, $\alpha(x) = \sum_{k=0}^{m_n} a_k f_k(x) \Longrightarrow$

$\sup_k |a_k| \leq \lambda_n$ where $\lambda_n = 0(n^{\delta})$ with $\delta < 1$).

One can easily check that for some constant $c_1 > 0$

$$\sup_n \sup_x |f_n(x)| \leq c_1.$$

If we define

$$B_{n,j} = \{\alpha\in S_n: \sup_x |\alpha(x)-\alpha_{n,j}(x)| \leq \frac{c_1}{n}\} , \quad j=1,2,\ldots\ell_n,$$

then, since $\sup_x |\sum_{k=0}^{m_n} \frac{1}{n^2} f_n(x)| \leq \frac{c_1}{n}$, $S_n \subseteq \bigcup_{j=1}^{\ell_n} B_{n,j}$.

Now return to (A.3):

$$P\{\sup_{\alpha\in S_n} |\tfrac{1}{n} \sum_{i=1}^{n} (\alpha(x_i)^2-2\alpha(x_i)y_i)-E[\alpha(X)^2-2\alpha(X)Y]| > \xi\}$$

$$\leq \sum_{j=1}^{\ell_n} P\{\sup_{\alpha\in B_{n,j}} |\tfrac{1}{n} \sum_{i=1}^{n} (\alpha(x_i)^2 - 2\alpha(x_i)y_i)-E[\alpha(X)^2-2\alpha(X)Y]| > \xi\}$$

$$\leq \sum_{j=1}^{\ell_n} P\{|\tfrac{1}{n} \sum_{i=1}^{n} (\alpha_{n,j}(x_i)^2-2\alpha_{n,j}(x_i)y_i)-E[\alpha(X)^2-2\alpha(X)Y]| > \xi\} \tag{A.4}$$

$$+ \sum_{j=1}^{\ell_n} P\{\sup_{\alpha\in B_{n,j}} |\tfrac{1}{n} \sum_{i=1}^{n} (\alpha(x_i)^2-\alpha_{n,j}(x_i)^2 \tag{A.5}$$

$$+ 2\alpha_{n,j}(x_i)y_i-2\alpha(x_i)y_i)$$

$$-E[\alpha(X)^2-\alpha_{n,j}(X)^2+2\alpha_{n,j}(X)Y-2\alpha(X)Y]| > \xi/2\} .$$

In (A.5), observe that

$$\sup_x \sup_{\alpha\in B_{n,j}} |\alpha(x)^2-\alpha_{n,j}(x)^2|$$

$$\leq \sup_x \sup_{\alpha\in B_{n,j}} |\alpha(x)-\alpha_{n,j}(x)|\cdot 2 \sup_x \sup_{\alpha\in S_n} |\alpha(x)|$$

$$\leq \frac{c_1}{n} 2c_1\lambda_n = 0(\frac{1}{n^{1-\delta}})$$

and

$$\sup_{x}\ \sup_{\alpha\in B_{n,j}} |2\alpha_{n,j}(x)y-2\alpha(x)y| \le \frac{2c_1|y|}{n} .$$

With these bounds, and the moment condition $E[e^{t_o|Y|}] < \infty$, it is easily verified that the expression in (A.5), as a sequence indexed by n, is summable. (The analysis of (A.4) is similar, but more delicate. See below.) The same must now be shown for the expression in (A.4). This, in turn, is a consequence of the following Lemma.

Lemma 3. Fix $\varepsilon > 0$. Let $Z_1,Z_2,\ldots Z_n$ be a sequence of i.i.d. random variables satisfying

a. $E[Z_1] = 0$,

b. $E[(Z_1-\varepsilon)^4] \le K_1^2$, and

c. $E[e^{s|Z_1|}] \le K_2^2 < \infty$

for some positive constants K_1,K_2, and $s > \frac{2\varepsilon}{K_1K_2}$. Then

$$P(|\frac{1}{n}\sum_{i=1}^{n} Z_i| > \varepsilon) \le 2(1 - \frac{\varepsilon^2}{2K_1K_2})^n .$$

Proof of Lemma 3. (Typical use of "large deviation" techniques, c.f. Chernoff [3].) For any $t\in(0,s)$:

$P(\frac{1}{n}\sum_{i=1}^{n} Z_i > \varepsilon) \le E[e^{t(Z_1-\varepsilon)}]^n$. Let $\phi(t) = E[e^{t(Z_1-\varepsilon)}]$. Then

$\phi(0) = 1$, $\frac{d}{dt}\phi(t)|_{t=0} = -\varepsilon$, and

$$\frac{d^2}{dt^2}\phi(t) = E[(Z_1-\varepsilon)^2 e^{t(Z_1-\varepsilon)}]$$

$$\le \sqrt{E[(Z_1-\varepsilon)^4]}\ \sqrt{E[e^{2tZ_1}]}\ \le K_1K_2 \quad \text{for} \quad t\in(0,s/2).$$

Integrating $\frac{d^2}{dt^2}\phi(t)$:

$$\frac{d}{dt}\phi(t) \le -\varepsilon + K_1K_2t \quad \text{for} \quad t\in(0,s/2).$$

And, integrating again:

$$\phi(t) \le 1-\varepsilon t + \frac{1}{2}K_1K_2t^2 \quad \text{for} \quad t\in(0,s/2).$$

$1-\varepsilon t + \frac{1}{2}K_1K_2t^2$ is minimized at $t = \varepsilon/K_1K_2$, which is, by assumption, smaller than $s/2$. Since

$$\phi(\frac{\varepsilon}{K_1K_2}) \le 1 - \frac{\varepsilon^2}{2K_1K_2} ,$$

$$P(\frac{1}{n}\sum_{i=1}^{n} Z_i > \varepsilon) \le (1 - \frac{\varepsilon^2}{2K_1K_2})^n .$$

Now do the same for $P(\frac{1}{n}\sum_{i=1}^{n} Z_i < -\varepsilon)$. □

Return to (A.4). Fix n and j and let

$$Z_i = \alpha_{n,j}(x_i)^2-2\alpha_{n,j}(x_i)y_i-E[\alpha_{n,j}(X)^2-2\alpha_{n,j}(X)Y].$$

Observe that $Z_1,\ldots,Z_n$ are i.i.d., with $E[Z] = 0$,

$$\sqrt{E[(Z_1-\xi/2)^4]} \le c_2n^{4\delta}, \quad \text{and}$$

$$\sqrt{E[e^{t|Z_1|}]} \le c_3\ \forall t \le \frac{c_4}{n^{2\delta}}$$

where c_2 and c_3 are sufficiently large constants and c_4 is a sufficiently small constant (c_2,c_3 and c_4 are independent of n and j). Now apply Lemma 3:

$$P(|\frac{1}{n}\sum_{i=1}^{n} Z_i| > \frac{\xi}{2}) = 0((1 - \frac{\xi^2}{c_5n^{4\delta}})^n)$$

for some sufficiently large c_5 (also independent of n and j). Finally, use this in (A.4):

$$\text{"A.4"} = 0\left(\ell_n\left(1 - \frac{\xi^2}{c_5 n^{4\delta}}\right)^n\right)$$

$$= 0\left(n^{3m_n}\left(1 - \frac{\xi^2}{c_5 n^{4\delta}}\right)^n\right) = 0\left(e^{-n^{1-\varepsilon}}\right)$$

which is summable. □

B. Spline Functions

Notation:

1. X,Y R^1-valued random variables, with the distribution of X concentrated on [0,1]
2. $(x_1,y_1),(x_2,y_2),\ldots$ i.i.d. observations of (X,Y)
3. $\alpha_0(x) = E[Y|X=x]$
4. $S_n = \{\alpha(x)$: $\alpha(x)$ continuous, $\frac{d}{dx}\alpha(x)$ piecewise continuous, $\int_0^1 |\frac{d}{dx}\alpha(x)|^2 dx \leq \lambda_n\}$
5. $\hat{\alpha}_n(x)$ is the solution to:

   minimize $\sum_{i=1}^{n} (y_i - \alpha(x_i))^2$

   subject to $\alpha \in S_n$. (If, for some $\alpha \in S_n$, $\sum_{i=1}^{n}(y_i-\alpha(x_i))^2 = 0$, then the solution may not be unique. In this case, define $\hat{\alpha}_n$ to be the linear interpolation of $(x_1,y_1),\ldots(x_n,y_n)$, with zero slope on the intervals between 0 and the first knot and between the last knot and 1. This function minimizes $\int_0^1 |\frac{d}{dx}\alpha(x)|^2 dx$ subject to $\sum_{i=1}^{n}(y_i-\alpha(x_i))^2 = 0$.)

6. F_X Distribution function of X
7. $\|\alpha-\beta\| = \sqrt{E|\alpha(X)-\beta(X)|^2} = \{\int_0^1 |\alpha(x)-\beta(x)|^2 F_X(dx)\}^{1/2}$

Theorem 2. If $E[e^{t_0|Y|}] < \infty$ for some $t_0 > 0$, and if $\lambda_n \uparrow \infty$ with $\lambda_n = 0(n^{1/4-\varepsilon})$ for some $\varepsilon > 0$, then

$$\|\hat{\alpha}_n - \alpha_0\| \to 0 \quad \text{a.s.}$$

as $n \to \infty$.

Proof: For any $\alpha \in S_n$, we can write

$$\alpha(x) = \sum_{k=0}^{\infty} a_k \cos k\pi x.$$

Since $\frac{d}{dx}\alpha(x) \in L_2([0,1],B,F_X)$, $\frac{d}{dx}\alpha(x)$ can be written as

$$\frac{d}{dx}\alpha(x) = \sum_{k=1}^{\infty} b_k \sin k\pi x$$

where

$$b_k = 2\int_0^1 (\frac{d}{dx}\alpha(x)) \sin k\pi x \, dx$$

$$= -2\pi k \int_0^1 \alpha(x) \cos k\pi x \, dx = -\pi k a_k.$$

Hence,

$$\lambda_n \geq \int_0^1 |\frac{d}{dx}\alpha(x)|^2 dx = \frac{1}{2}\sum_{k=1}^{\infty} b_k^2$$

$$= \frac{1}{2}\pi^2 \sum_{k=1}^{\infty} k^2 a_k^2, \quad \text{i.e.}$$

$$\sum_{k=1}^{\infty} k^2 a_k^2 \leq \frac{2\lambda_n}{\pi^2}. \qquad \text{(A.6)}$$

For $\hat{\alpha}_n(x)$, write

$$\hat{\alpha}_n(x) = \sum_{k=0}^{\infty} \hat{a}_k \cos k\pi x.$$

Then, since $\hat{a}_0$ is not restricted by the inequality

$\int_0^1 |\frac{d}{dx} \hat{\alpha}_n(x)|^2 dx \leq \lambda_n$, $\hat{a}_0$ must minimize

$$\sum_{i=1}^{n} (y_i - \sum_{k=1}^{\infty} \hat{a}_k \cos k\pi x_i - a_0)^2,$$

i.e. $$\hat{a}_0 = \frac{1}{n} \sum_{i=1}^{n} (y_i - \sum_{k=1}^{\infty} \hat{a}_k \cos k\pi x).$$

Consequently,

$$|\hat{a}_0| \leq |\frac{1}{n} \sum_{i=1}^{n} y_i| + \sum_{k=1}^{\infty} |\hat{a}_k|$$

$$\leq |\frac{1}{n} \sum_{i=1}^{n} y_i| + \sqrt{\sum_{k=1}^{\infty} k^2 \hat{a}_k^2} \sqrt{\sum_{k=1}^{\infty} 1/k^2}$$

$$\leq \text{(use (A.6))} \; |\frac{1}{n} \sum_{i=1}^{n} y_i| + \frac{\sqrt{2\lambda_n}}{\pi} \sqrt{\sum_{k=1}^{\infty} 1/k^2} \; .$$

Hence, there is a constant $c_1 > 0$ such that, with probability one,

$$|\int_0^1 \hat{\alpha}_n(x)dx| = |\hat{a}_0| \leq 2c_1\sqrt{\lambda_n}$$

for all n sufficiently large.

Define

$$S_n' = \{\alpha \in S_n : \; |\int_0^1 \alpha(x)dx| \leq 2c_1 \sqrt{\lambda_n}\},$$

and

$$A_n' = \{\alpha \in S_n' : \; \sum_{i=1}^{n} (y_i - \alpha(x_i))^2 = \inf_{\beta \in S_n'} \sum_{i=1}^{n} (y_i - \beta(x_i))^2\} \; .$$

Then, eventually (i.e. for all n large enough) $\hat{\alpha}_n(x) \in A_n'$, and it is therefore sufficient to prove that

$$\sup_{\alpha \in A_n'} \|\alpha - \alpha_0\| \to 0 \quad \text{a.s.} \qquad \text{(A.7)}$$

as $n \to \infty$.

The continuously differentiable functions are dense in $L_2([0,1],B,F_X)$. Consequently, $\bigcup_{n=1}^{\infty} S_n'$ is also dense in $L_2([0,1],B,F_X)$, and since $S_1' \subset S_2' \subset \ldots$, we can find a sequence $\beta_1, \beta_2, \ldots$ such that $\beta_n \in S_n'$ for each n, and $\beta_n \to \alpha_0$ as $n \to \infty$. Recall now the proof of Theorem 1; once an analogue for Lemma 2 is established, that proof can be used without modification for (A.7). In other words, it remains only to prove

Lemma 4

$$\sup_{\alpha \in S_n'} |\frac{1}{n} \sum_{i=1}^{n} (\alpha(x_i)^2 - 2\alpha(x_i)y_i) - E[\alpha(X)^2 - 2\alpha(X)Y]| \to 0 \quad \text{a.s.}$$

Proof of Lemma 4. (Similar to proof of Lemma 2).

Because of the Borel-Cantelli Lemma, it will be sufficient to show that, for arbitrary $\xi > 0$,

$$\sum_{n=1}^{\infty} P\{\sup_{\alpha \in S_n'} |\frac{1}{n} \sum_{i=1}^{n} (\alpha(x_i)^2 - 2\alpha(x_i)y_i) - E[\alpha(X)^2 - 2\alpha(X)Y]| > \xi\}$$

$$< \infty. \qquad \text{(A.8)}$$

Use (A.6), and the observation

$$|a_0| = |\int_0^1 \alpha(x)dx| \leq 2c_1\sqrt{\lambda_n}$$

for $\alpha \in S_n'$, to rewrite S_n':

$$S_n' = \{\alpha(x) = \sum_{k=0}^{\infty} a_k \cos k\pi x: \ |a_0| \leq 2c_1\sqrt{\lambda_n}\ ,$$
$$\sum_{k=1}^{\infty} k^2 a_k^2 \leq \frac{2\lambda_n}{\pi^2}\} \ .$$

Let [x] denote the greatest integer less than or equal to x, and let $\alpha_{n,1}, \alpha_{n,2}, \ldots \alpha_{n,\ell_n}$ denote the functions $\alpha \in S_n'$ of the form

$$\alpha(x) = \sum_{k=0}^{[n^{1/2}]} \frac{p_k}{n} \cos k\pi x,$$

where $p_0, p_1, \ldots p_{[n^{1/2}]}$ are (positive or negative) integers.

If $\alpha \in S_n'$ and $\alpha = \sum_{k=0}^{\infty} a_k \cos k\pi x$ then

$$|a_k| \leq \sqrt{\lambda_n} \max(2c_1, \frac{\sqrt{2}}{\pi}) = 0(n^{1/8}).$$

Hence $\ell_n = 0(n^{2\sqrt{n}})$. Fix j $1 \leq j \leq \ell_n$. For some set of integers $i_0, i_1, \ldots, i_{[n^{1/2}]}$, we can write

$$\alpha_{n,j}(x) = \sum_{k=0}^{[n^{1/2}]} \frac{i_k}{n} \cos k\pi x.$$

Define

$$B_{n,j} = \{\alpha = \sum_{k=0}^{\infty} a_k \cos k\pi x: \ \alpha \in S_n', \ |a_k - i_k| \leq 1$$
$$\text{for } k=0,1,\ldots [n^{1/2}] \ ,$$

and define, similarly, $B_{n,\ell}$ $\ell \neq j$.

Obviously $S_n' \subseteq \bigcup_{j=1}^{\ell_n} B_{n,j}$. Furthermore, $\alpha, \beta \in B_{n,j} \Rightarrow$ (writing $\alpha(x) = \sum_{k=0}^{\infty} a_k \cos k\pi x$ and $\beta(x) = \sum_{k=0}^{\infty} b_k \cos k\pi x$)

$$\sup_x |\alpha(x) - \beta(x)| \leq \sum_{k=0}^{\infty} |a_k - b_k|$$
$$= \sum_{k=0}^{[n^{1/2}]} |a_k - b_k| + \sum_{k=[n^{1/2}]+1}^{\infty} |a_k - b_k|$$
$$\leq \sum_{k=0}^{[n^{1/2}]} \frac{2}{n} + \sum_{k=[n^{1/2}]+1}^{\infty} |a_k| + \sum_{k=[n^{1/2}]+1}^{\infty} |b_k|$$
$$\leq \frac{2}{n}(n^{1/2}+1) + \{(\sum_{k=0}^{\infty} k^2 a_k^2)(\sum_{k=[n^{1/2}]+1}^{\infty} 1/k^2)\}^{1/2}$$
$$+ \{(\sum_{k=0}^{\infty} k^2 b_k^2)(\sum_{k=[n^{1/2}]+1}^{\infty} 1/k^2)\}^{1/2}$$
$$\leq c_2(n^{-1/2} + \lambda_n^{1/2} n^{-1/4}) \quad \text{for some } c_2 \text{ sufficiently large}$$

(in the last step, use (A.6) and the definition of S_n'). Then, since $\lambda_n = 0(n^{1/4})$,

$$\sup_x |\alpha(x) - \beta(x)| = 0(n^{-1/8}). \qquad \text{(A.9)}$$

Also, for any $\alpha \in S_n'$ (say $\alpha(x) = \sum_{k=0}^{\infty} a_k \cos k\pi x$):

$$\sup_x |\alpha(x)| \leq \sum_{k=0}^{\infty} |a_k| \leq \{(\sum_{k=0}^{\infty} k^2 a_k^2)(\sum_{k=0}^{\infty} 1/k^2)\}^{1/2}$$
$$= 0(\lambda_n^{1/2}) = 0(n^{1/8-\varepsilon/2}). \qquad \text{(A.10)}$$

Return now to (A.8), and follow exactly the reasoning used in the proof of Lemma 2:

$$P\{\sup_{\alpha \in S_n}, |\frac{1}{n} \sum_{i=1}^{n} (\alpha(x_i)^2 - 2\alpha(x_i)y_i) - E[\alpha(X)^2 - 2\alpha(X)Y]| > \xi\}$$

$$\leq \sum_{j=1}^{\ell_n} P\{|\frac{1}{n} \sum_{i=1}^{n} (\alpha_{n,j}(x_i)^2 - 2\alpha_{n,j}(x_i)y_i) - E[\alpha_{n,j}(X)^2 - 2\alpha_{n,j}(X)Y]| > \xi/2\} \qquad (A.11)$$

$$+ \sum_{j=1}^{\ell_n} P\{\sup_{\alpha \in B_{n,j}} |\frac{1}{n} \sum_{i=1}^{n} (\alpha(x_i)^2 - \alpha_{n,j}(x_i)^2 \qquad (A.12)$$

$$+ 2\alpha_{n,j}(x_i)y_i - 2\alpha(x_i)y_i)$$

$$- E[\alpha(X)^2 - \alpha_{n,j}(X)^2 + 2\alpha_{n,j}(X)Y - 2\alpha(X)Y]| > \xi/2\}.$$

In (A.12) apply the bounds

$$\sup_{x} \sup_{\alpha \in B_{n,j}} |\alpha(x)^2 - \alpha_{n,j}(x)^2|$$

$$\leq \sup_{x} \sup_{\alpha \in B_{n,j}} |\alpha(x) - \alpha_{n,j}(x)| \cdot 2 \sup_{x} \sup_{\alpha \in S_n}, |\alpha(x)|$$

$$= (\text{use (A.9) and (A.10)}) \; 0(n^{-\varepsilon/2}),$$

and

$$\sup_{x} \sup_{\alpha \in B_{n,j}} |2\alpha_{n,j}(x)y - 2\alpha(x)y| = (\text{use (A.9)}) \quad |y| 0(n^{-1/8}).$$

Using $E[e^{t_o|Y|}] < \infty$, conclude that (A.12) is a summable sequence in n. In (A.11), define (for fixed n and j)

$$z_i = \alpha_{n,j}(x_i)^2 - 2\alpha_{n,j}(x_i)y_i - E[\alpha_{n,j}(X)^2 - 2\alpha_{n,j}(X)Y].$$

Because of (A.10), Lemma 3 applies with $\varepsilon = \xi/2$, $k_1 = c_3 n^{1/2-2\varepsilon}$, $k_2 = c_4$, and $s = c_5/n^{1/4}$, provided that c_3 and c_4 are sufficiently large constants, and c_5 is sufficiently small. Since, for all n sufficiently large,

$$s = \frac{c_5}{n^{1/4}} \geq \frac{\xi}{c_3 c_4 n^{1/2-2\varepsilon}} \doteq \frac{2\varepsilon}{k_1 k_2},$$

Lemma 3 implies that

$$P(|\frac{1}{n} \sum_{i=1}^{n} z_i| > \varepsilon) \leq 2(1 - \frac{\xi^2}{8c_3c_4 n^{1/2-2\varepsilon}})^n$$

for all n sufficiently large. Finally, put this into (A.11):

$$\text{"A.11"} = 0(\ell_n(1 - \frac{\xi^2}{8c_3c_4 n^{1/2-2\varepsilon}})^n)$$

$$= 0(n^{2\sqrt{n}}(1 - \frac{\xi^2}{8c_3c_4 n^{1/2-2\varepsilon}})^n)$$

$$= 0(e^{-\sqrt{n}}).$$

Since this too is summable, (A.8) is established, and the proof is complete. □

C. Dirichlet Kernel

Notation:

1. $F_x(y)$ for each $x \in [0,1]$, a probability distribution function on R^1
2. $x_j = \frac{j}{n}$ $j=1,2,\ldots n$
3. $y_1, \ldots y_n$ independent observations such that $y_j \sim F_{x_j}$

4. $\alpha_0(x) = \int_{-\infty}^{\infty} y\, F_x(dy) \quad x\in[0,1]$

5. $V(x) = \int_{-\infty}^{\infty} (y-\alpha_0(x))^2 F_x(dy) \quad x\in[0,1]$

6. $S_n = \{\alpha(x): \alpha(x) = \sum_{k=-m_n}^{m_n} a_k e^{2\pi ikx}\} \quad (m_n \le n)$

7. $\hat{\alpha}_n$ the (unique) solution to: minimize $\sum_{j=1}^{n}(y_j-\alpha(x_j))^2$ subject to $\alpha\in S_n$.

Theorem 3. If

A1. there exists a constant K such that $|\alpha_0(x)-\alpha_0(y)| \le K|x-y| \quad \forall\ x,y\in[0,1]$,

and

A2. $\sup_{x\in[0,1]} \int_{-\infty}^{\infty} y^2\, F_x(dy) < \infty$,

then for any sequence $m_n\uparrow\infty$ such that $m_n/n \to 0$ and $m_n \le n$,

$$E\int_0^1 |\hat{\alpha}_n(x)-\alpha_0(x)|^2 dx = 0(\frac{1}{m_n} + \frac{m_n}{n} + \frac{1}{\sqrt{n}})$$

as n → ∞. In particular, if $m_n \sim \sqrt{n}$, then

$$E\int_0^1 |\hat{\alpha}_n(x)-\alpha_0(x)|^2 dx = 0(\frac{1}{\sqrt{n}})$$

as n → ∞.

Proof: By straightforward calculation:

$$\hat{\alpha}_n(x) = \sum_{k=-m_n}^{m_n} (\frac{1}{n} \sum_{j=1}^{n} y_j e^{-2\pi ikx_j}) e^{2\pi ikx}.$$

Let $a_k^0 = \int_0^1 \alpha_0(x) e^{-2\pi ikx} dx$. Then, in the L_2 sense,

$$\alpha_0(x) = \sum_{-\infty}^{\infty} a_k^0 e^{2\pi ikx} .$$

We have:

$$E\int_0^1 |\hat{\alpha}_n(x)-\alpha_0(x)|^2 dx$$

$$= \sum_{|k|>m_n} |a_k^0|^2$$

$$+ \sum_{|k|\le m_n} E|\frac{1}{n} \sum_{j=1}^{n} y_j e^{-2\pi ikx_j} - \int_0^1 \alpha_0(x) e^{-2\pi ikx} dx|^2$$

$$= \sum_{|k|>m_n} |a_k^0|^2$$

$$+ \sum_{|k|\le m_n} E|\frac{1}{n} \sum_{j=1}^{n} e^{-2\pi ikx_j} (y_j-\alpha_0(x_j))$$

$$+ (\frac{1}{n} \sum_{j=1}^{n} \alpha_0(x_j) e^{-2\pi ikx_j} - \int_0^1 \alpha_0(x) e^{-2\pi ikx}\, dx)|^2$$

$$= \sum_{|k|>m_n} |a_k^0|^2 + \sum_{|k|\le m_n} E|\frac{1}{n} \sum_{j=1}^{n} e^{-2\pi ikx_j} (y_j-\alpha_0(x_j))|^2$$

$$+ \sum_{|k|\le m_n} |\frac{1}{n} \sum_{j=1}^{n} \alpha_0(x_j) e^{-2\pi ikx_j} - \int_0^1 \alpha_0(x) e^{-2\pi ikx}\, dx|^2$$

$$= \sum_{|k|>m_n} |a_k^0|^2 + \frac{2m_n+1}{n^2} \sum_{j=1}^{n} V(x_j) + \sum_{|k|\le m_n} |\varepsilon_{k,n}|^2 \qquad (A.13)$$

where $\varepsilon_{k,n} = \frac{1}{n} \sum_{j=1}^{n} \alpha_0(x_j) e^{-2\pi i k x_j} - \int_0^1 \alpha_0(x) e^{-2\pi i k x} dx.$

I will discuss, separately, each of the three terms in (A.13).

First, let us bound the rate at which the Fourier coefficients, a_k^0 $k=0,\pm 1,\pm 2,\ldots,$ approach 0 as $|k| \to \infty$. For this, use the Lipschitz condition A1, and a slight modification of a proof of the Riemann-Lebesgue Lemma (see Dym and McKean [6]):

Define $\alpha_0(-x) = \alpha_0(1-x)$ for $x\in(0,1)$. Then

$$a_k^0 = \int_0^1 \alpha_0(x) e^{-2\pi i k x} dx = - \int_0^1 \alpha_0(x) e^{-2\pi i k (x+\frac{1}{2k})} dx$$

$$= - \int_0^1 \alpha_0(x-\tfrac{1}{2k}) e^{-2\pi i k x} dx.$$

Hence, taking the average of two expressions for a_k^0:

$$|a_k^0| = |\frac{1}{2} \int_0^1 (\alpha_0(x)-\alpha_0(x-\tfrac{1}{2k})) e^{-2\pi i k x} dx|$$

$$\le \frac{1}{2} \int_0^{\frac{1}{2k}} |\alpha_0(x)-\alpha_0(x-\tfrac{1}{2k})| dx + \frac{K}{4K} = 0(\tfrac{1}{k}) \qquad (A.14)$$

Therefore, for the first term in (A13),

$$\sum_{|k|>m_n} |a_k^0|^2 = 0(\tfrac{1}{m_n}) \quad \text{as} \quad n \to \infty . \qquad (A.15)$$

For the second term in (A.13), apply A2:

$$\frac{2m_n+1}{n^2} \sum_{j=1}^{n} V(x_j) = 0(\tfrac{m_n}{n}) \text{ as } n \to \infty . \qquad (A.16)$$

Finally, I will show that, for some c sufficiently large

$$|\varepsilon_{k,n}| \le c \ \min(\tfrac{k}{n},\tfrac{1}{k}) \qquad (A.17)$$

for all n and $|k| \le m_n$. Then

$$\sum_{|k|\le m_n} |\varepsilon_{k,n}|^2 \le \sum_{|k|\le[\sqrt{n}]} \frac{c^2k^2}{n^2} + 2 \sum_{k=[\sqrt{n}]+1}^{m_n} \frac{c^2}{k^2}$$

$$= 0(\tfrac{1}{\sqrt{n}})$$

and this, together with (A.13),(A.15), and (A.16), will complete the proof.

Observe that

$$\frac{1}{n} \sum_{j=1}^{n} \alpha_0(x_j) e^{-2\pi i k x_j}$$

a Riemann approximation to the integral

$$\int_0^1 \alpha_0(x) e^{-2\pi i k x} dx.$$

Since $\alpha_0(x)e^{-2\pi i k x}$ is uniformly Lipschitz continuous, with constant 0(k), and since $x_j - x_{j-1} = \frac{1}{n}$ for $j=2,3,\ldots n,$

$$|\varepsilon_{k,n}| = |\frac{1}{n} \sum_{j=1}^{n} \alpha_0(x_j) e^{-2\pi i k x_j} - \int_0^1 \alpha_0(x) e^{-2\pi i k x} dx|$$

$$= 0(\tfrac{k}{n}),$$

which is "one half" of (A.17). For the other half, write

$$|\varepsilon_{k,n}| \leq |\frac{1}{n} \sum_{j=1}^{n} \alpha_0(x_j) e^{-2\pi ikx_j}| + |\int_0^1 \alpha_0(x) e^{-2\pi ikx} dx|.$$

We already know that $|\int_0^1 \alpha_0(x) e^{-2\pi ikx} dx| = 0(\frac{1}{K})$ (see A.14),

so it remains only to show that

$$|\frac{1}{n} \sum_{j=1}^{n} \alpha_0(x_i) e^{-2\pi ikx_j}| = 0(\frac{1}{K})$$

as well. Summing by parts:

$$\frac{1}{n} \sum_{j=1}^{n} \alpha_0(x_j) e^{-2\pi ikx_j} = - \frac{\alpha_0(x_1)}{n}$$

$$- \frac{1}{n} \sum_{j=1}^{n} [\alpha_0(x_{j+1}) - \alpha_0(x_j)] \frac{1-e^{-2\pi ik(\frac{j+1}{n})}}{1-e^{-2\pi ik(\frac{1}{n})}}$$

where I have defined $\alpha_0(x_{j+1}) = 0$. Therefore for some constant b sufficiently large,

$$|\frac{1}{n} \sum_{j=1}^{n} \alpha_0(x_j) e^{-2\pi ikx_j}|$$

$$\leq \frac{b}{n} + \sqrt{\frac{1}{n} \sum_{j=1}^{n} |\alpha_0(x_{j+1}) - \alpha_0(x_j)|^2} \sqrt{\frac{1}{n} \sum_{j=1}^{n} \left|\frac{1-e^{-2\pi ik(\frac{j+1}{n})}}{1-e^{-2\pi ik(\frac{1}{n})}}\right|^2}$$

$$\leq \frac{b}{n} + \frac{K}{n} \sqrt{\frac{1}{n} \sum_{j=1}^{n} \frac{1-\cos(2\pi k \frac{j+1}{n})}{1-\cos(2\pi k \frac{1}{n})}}$$

$$= \frac{b}{n} + \frac{K}{n} \sqrt{\frac{1}{1-\cos(2\pi k \frac{1}{n})}} = 0(\frac{1}{K}),$$

since $1-\cos(2\pi k \frac{1}{n}) \geq \frac{\pi^2 k^2}{n^2}$ when $\frac{k}{n}$ is small. This, then, establishes (A.17) and completes the proof of Theorem 3.

□

Theorem 4. If

A1. $\alpha_0(0) = \alpha_0(1)$,

A2. $\alpha_0(x)$ has a continuous derivative, $\alpha_0(x)'$, and for some constant k_1

$$|\alpha_0(x)' - \alpha_0(y)'| \leq k_1 |x-y| \quad \forall\ x,y \in [0,1],$$

A3. there exists a constant, k_2, such that

$$|V(x)-V(y)| \leq k_2 |x-y| \quad \forall\ x,y \in [0,1],$$

and

A4.

$$\sup_{x \in [0,1]} \int_{-\infty}^{\infty} y^4 F_x(dy) < \infty,$$

then for any sequence $m_n \uparrow \infty$, such that $m_n = 0(n^{\beta})$ for some $\frac{1}{4} < \beta < \frac{1}{2}$, the process

$$p_n(t) \equiv \sqrt{n} \int_0^t (\hat{\alpha}_n(x) - \alpha_0(x)) dx$$

converges weakly on [0,1] to the diffusion, p(t), defined by

$$dp(t) = \sqrt{V(t)}\, dW_t, \quad p(0) = 0$$

where W_t is standard Brownian motion.

Proof: In two parts: first show appropriate convergence of the "finite dimensional distributions", and then show that the sequence of distributions associated with $p_n(\cdot)$ $n=1,2,\ldots$ is tight (on $C[0,1]$). For the first part, since $p_n(0) = 0$, it is enough to show that for any q, any $\beta_1,\ldots\beta_q$, and any $0 = s_0 \le s_1 \le s_2 \le \ldots \le s_q \le 1$,

$$\sum_{\ell=1}^{q} \beta_\ell(p_n(s_\ell)-p_n(s_{\ell-1})) \xrightarrow{w} N(0, \sum_{\ell=1}^{q} \beta_\ell^2 \int_{s_{\ell-1}}^{s_\ell} V(x)dx). \tag{A.18}$$

Recall that (see proof of Theorem 3):

$$\hat{\alpha}_n(x) = \sum_{|k|\le m_n} (\frac{1}{n} \sum_{j=1}^{n} y_j e^{-2\pi ikx_j}) e^{2\pi ikx} .$$

Let

$$a_k^0 = \int_0^1 \alpha_0(x)e^{-2\pi ikx}dx, \text{ and let}$$

$$t_k = \int_0^1 (\sum_{\ell=1}^{q} \beta_\ell I_{[s_{\ell-1},s_\ell]}(x))e^{-2\pi ikx}dx.$$

Then

$$\sum_{\ell=1}^{q} \beta_\ell(p_n(s_\ell)-p_n(s_{\ell-1}))$$

$$= \sqrt{n} \int_0^1 (\sum_{\ell=1}^{q} \beta_\ell I_{[s_{\ell-1},s_\ell]}(x))(\hat{\alpha}_n(x)-\alpha_0(x))dx$$

$$= - \sqrt{n} \sum_{|k|>m_n} \bar{t}_k a_k^0 \tag{A.19}$$

$$+ \sqrt{n} \sum_{|k|\le m_n} \bar{t}_k(\frac{1}{n} \sum_{j=1}^{n} \alpha_0(x_j)e^{-2\pi ikx_j} - \int_0^1 \alpha_0(x)e^{-2\pi ikx}dx) \tag{A.20}$$

$$+ \frac{1}{\sqrt{n}} \sum_{j=1}^{n} (y_j-\alpha_0(x_j))(\sum_{|k|\le m_n} \bar{t}_k e^{-2\pi ikx_j}). \tag{A.21}$$

I will show that the expressions in (A.19) and (A.20) approach zero, and that the expression in (A.21) converges weakly to

$$N(0, \sum_{\ell=1}^{q} \beta_\ell^2 \int_{s_{\ell-1}}^{s_\ell} V(x)dx),$$

thereby establishing (A.18) and completing the first part of the proof.

Concerning the expression in (A.19), observe that $t_k = 0(\frac{1}{k})$, and that $a_k^0 = 0(\frac{1}{k^2})$; the latter by an argument similar to the one used for (A.14), but preceded by an integration by parts. Therefore

$$\sqrt{n} \sum_{|k|>m_n} \bar{t}_k a_k^0 = 0(\frac{\sqrt{n}}{m_n^2}) \to 0$$

as $n \to \infty$. In (A.20), apply (A.17):

$$\sqrt{n} \sum_{|k|\le m_n} \bar{t}_k(\frac{1}{n} \sum_{j=1}^{n} \alpha_0(x_j)e^{-2\pi ikx_j} - \int_0^1 \alpha_0(x)e^{-2\pi ikx}dx)$$

$$\le \sqrt{n} \sum_{|k|\le m_n} |\bar{t}_k| \frac{ck}{n} = 0(\frac{m_n}{\sqrt{n}}) \to 0$$

as $n \to \infty$.

Now define

$$V_n = \frac{1}{n} \sum_{j=1}^{n} V(x_j) \,|\sum_{|k|\le m_n} \bar{t}_k e^{-2\pi ikx_j}|^2$$

and

$$x_{j,n} = \frac{1}{\sqrt{nV_n}} (y_j - \alpha_0(x_j)) (\sum_{|k| \leq m_n} \bar{t}_k \; e^{-2\pi i k x_j}).$$

Then, the expression in (A.21) can be rewritten as

$$\sqrt{V_n} \sum_{j=1}^{n} x_{j,n}.$$

Suppose, for now, that

$$V_n \to \sum_{\ell=1}^{g} \beta_\ell^2 \int_{s_{\ell-1}}^{s_\ell} V(x)dx \qquad \text{(A.22)}$$

as $n \to \infty$ (to be shown later). Observe that $x_{1,n},\dots x_{n,n}$ are independent with

$$E[x_{j,n}] = 0, \; \sum_{j=1}^{n} E[x_{j,n}^2] = 1, \text{ and } E|x_{j,n}|^3 = O(1/n^{3/2})$$

$(\sum_{|k| \leq m_n} \bar{t}_k \; e^{-2\pi i k x}$ is the truncated Fourier expansion of

$$\sum_{\ell=1}^{g} \beta_\ell I_{[s_{\ell-1},s_\ell]}(x),$$

and it remains uniformly bounded as $n \to \infty$). It is well known that under these conditions (for example, see Chung [4], Theorem 7.1.2)

$$\sum_{j=1}^{n} x_{j,n} \xrightarrow{w} N(0,1),$$

which implies

$$\sqrt{V_n} \sum_{j=1}^{n} x_{j,n} \xrightarrow{w} N(0, \sum_{\ell=1}^{g} \beta_\ell^2 \int_{s_{\ell-1}}^{s_\ell} V(x)dx).$$

Hence, for the first part of the proof, it remains to show (A.22):

$$\begin{aligned} V_n = & \sum_{|k| \leq m_n} \sum_{|r| \leq m_n} \bar{t}_k t_k \{\frac{1}{n} \sum_{j=1}^{n} V(x_j) e^{-2\pi i x_j (k-r)} \\ & - \int_0^1 V(x) e^{-2\pi i x(k-r)} dx\} \qquad \text{(A.23)} \\ & + \int_0^1 V(x) | \sum_{|k| \leq m_n} \bar{t}_k e^{-2\pi i k x}|^2 \; dx. \end{aligned}$$

Because of the assumption A2,

$$\begin{aligned} & |\frac{1}{n} \sum_{j=1}^{n} V(x_j) e^{-2\pi i x_j (k-r)} - \int_0^1 V(x) e^{-2\pi i x(k-r)} dx| \\ & = O(\frac{k-r}{n}) = O(\frac{m_n}{n}). \end{aligned}$$

Put this back into (A.23):

$$\begin{aligned} V_n & = O(\frac{m_n (\log m_n)^2}{n}) + \int_0^1 V(x) | \sum_{|k| \leq m_n} \bar{t}_k e^{-2\pi i k x}|^2 \; dx \\ & \longrightarrow \int_0^1 V(x) \sum_{\ell=1}^{g} \beta_\ell^2 \; I_{[s_{\ell-1},s_\ell]}(x) dx \\ & = \sum_{\ell=1}^{g} \beta_\ell^2 \int_{s_{\ell-1}}^{s_\ell} V(x) dx \end{aligned}$$

(since $\sum_{|k| \leq m_n} \bar{t}_k e^{-2\pi i k x} \longrightarrow \sum_{\ell=1}^{g} \beta_\ell I_{[t_{\ell-1},t_\ell]}(x)$

in the L_2 sense, and $V(x)$ is bounded).

The second part of the proof is to establish tightness for the distributions associated with $p_n(\cdot)$ $n=1,2,\ldots$. I will show that for some sufficiently large constant, c,

$$E|p_n(s_2)-p_n(s_1)|^4 \leq c^2(s_2-s_1)^2 \qquad \text{(A.24)}$$

for all n and all $0 \leq s_1 \leq s_2 \leq 1$. Then, by Theorem 12.3 of Billingsley [2] (with $\gamma=4$, $\alpha=2$, and $F(t) = ct$), the sequence is tight.

Begin as in the first part of the proof:

$$p_n(s_2)-p_n(s_1) = -\sqrt{n} \sum_{|k|>m_n} \bar{t}_k a_k^0$$

$$+ \sqrt{n} \sum_{|k|\leq m_n} \bar{t}_k \left(\frac{1}{n} \sum_{j=1}^{n} \alpha_0(x_j) e^{-2\pi ikx_j} - \int_0^1 \alpha_0(x) e^{-2\pi ikx} dx\right)$$

$$+ 1/\sqrt{n} \sum_{j=1}^{n} (y_j-\alpha_0(x_j)) \left(\sum_{|k|\leq m_n} \bar{t}_k e^{-2\pi ikx_j}\right)$$

where, here,

$$t_k = \int_{s_1}^{s_2} e^{-2\pi ikx} dx.$$

Since, for some c_1 sufficiently large, $|t_k| \leq c_1(s_2-s_1)/k$ $\forall$ $0 \leq s_1 \leq s_2 \leq 1$, and since $a_k^0 = 0(1/k^2)$ (as shown earlier),

$$|-\sqrt{n} \sum_{|k|>m_n} \bar{t}_k a_k^0| \leq c_2(s_2-s_1) \quad \forall\ n,\ 0 \leq s_1 \leq s_2 \leq 1,$$

for some constant c_2. Similarly, for some $c_3 > 0$ and $c_4 > 0$

$$|\sqrt{n} \sum_{|k|\leq m_n} \bar{t}_k \left(\frac{1}{n} \sum_{j=1}^{n} \alpha_0(x_j) e^{-2\pi ikx_j} - \int_0^1 \alpha_0(x) e^{-2\pi ikx} dx\right)|$$

$$\leq \sqrt{n} \sum_{|k|\leq m_n} c_3 \frac{|s_2-s_1|}{k} \frac{k}{n} \leq c_4 |s_2-s_1|$$

$\forall n$, $0 \leq s_1 \leq s_2 \leq 1$. Hence

$$E|p_n(s_2)-p_n(s_1)|^4$$

$$\leq c_5\left(|s_2-s_1|^4 + E\left|\frac{1}{\sqrt{n}} \sum_{j=1}^{n} (y_j-\alpha(x_j)) \left(\sum_{|k|\leq m_n} \bar{t}_k e^{-2\pi ikx_j}\right)\right|^4\right)$$

$\forall n$, $0 \leq s_1 \leq s_2 \leq 1$, some $c_5 > 0$ (use $(a+b)^4 \leq 8(a^4+b^4)$). So, for (A.24), it will be enough to show that for some $c_6 > 0$

$$E\left|\frac{1}{\sqrt{n}} \sum_{j=1}^{n} (y_j-\alpha(x_j)) \left(\sum_{|k|\leq m_n} \bar{t}_k e^{-2\pi ikx_j}\right)\right|^4 \leq c_6|s_2-s_1|^2$$

$\forall n$, $0 \leq s_1 \leq s_2 \leq 1$.

Let

$$D_m(x) = \sum_{|k|\leq m} e^{-2\pi ikx} = \frac{\sin \pi(2m+1)x}{\sin \pi x}$$

(the Dirichlet kernel), and let

$$I_{s_1,s_2,m}(x) = \int_{s_1}^{s_2} D_m(x-y)dy.$$

Then

$$\sum_{|k|\leq m_n} \bar{t}_k e^{-2\pi ikx_j} = \sum_{|k|\leq m_n} \int_{s_1}^{s_2} e^{-2\pi ik(x-x_j)} dx$$

$$= I_{s_1,s_2,m_n}(x_j)$$

and

$$E\left|\frac{1}{\sqrt{n}} \sum_{j=1}^{n} (y_j-\alpha(x_j)) \left(\sum_{|k|\leq m_n} \bar{t}_k e^{-2\pi ikx_j}\right)\right|^4$$

$$= \frac{1}{n^2} \sum_{j=1}^{n} E(y_j-\alpha(x_i))^4 I_{s_1,s_2,m_n}(x_j)^4$$

$$+ \frac{3}{n^2} \sum_{j=1}^{n} \sum_{\substack{k=1 \\ k\neq j}}^{n} V(x_j)V(x_k) I_{s_1,s_2,m_n}(x_j)^2 I_{s_1,s_2,m_n}(x_k)^2$$

$$\leq \frac{c_7}{n^2} \sum_{j=1}^{n} \sum_{k=1}^{n} I_{s_1,s_2,m_n}(x_j)^2 I_{s_1,s_2,m_n}(x_k)^2$$

$$= c_7 (\frac{1}{n} \sum_{j=1}^{n} I_{s_1,s_2,m_n}(x_j)^2)^2.$$

It is easily demonstrated that

$$\sup_{0\leq s_1\leq s_2\leq 1} \sup_{x\in[0,1]} |\int_{s_1}^{s_2} D_m(x-y)dy|$$

$$\leq \int_{-\frac{1}{2m+1}}^{\frac{1}{2m+1}} D_m(x)dx \leq 2.$$

Hence

$$\frac{1}{n} \sum_{j=1}^{n} I_{s_1,s_2,m_n}(x_j)^2$$

$$= \frac{1}{n} \sum_{j=1}^{n} \int_{s_1}^{s_2} \int_{s_1}^{s_2} \sum_{|k|\leq m_n} \sum_{|\ell|\leq m_n} e^{2\pi ik(x-x_j)} e^{-2\pi i\ell(y-x_j)} dydx$$

$$= \int_{s_1}^{s_2} \int_{s_1}^{s_2} \sum_{|k|\leq m_n} e^{2\pi ik(x-y)} dydx$$

$$= \int_{s_1}^{s_2} \int_{s_1}^{s_2} D_{m_n}(x-y)dydx \leq 2(s_2-s_1)$$

$$\Rightarrow E|\frac{1}{\sqrt{n}} \sum_{j=1}^{n} (y_j-\alpha(x_j))(\sum_{|k|\leq m_n} \bar{\tau}_k e^{-2\pi ikx_j})|^4$$

$$\leq c_7(\frac{1}{n} \sum_{j=1}^{n} I_{s_1,s_2,m_n}(x_j)^2)^2$$

$$\leq 4c_7(s_2-s_1)^2,$$

and this establishes (A.24), and completes the proof. □

III. Nonparametric Density Estimation by Maximum Likelihood.

Notation:

1. X R^1-valued random variable with absolutely continuous distribution
2. $x_1, x_2, \ldots$ i.i.d. observations of X
3. $\alpha_0(x)$ density function for X
4. $S_n = \{\alpha(x): \alpha(x) = \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi\sigma_n^2}} e^{-(x-y)^2/2\sigma_n^2} F(dy)$

   F a probability distribution function$\}$
5. $A_n = \{\alpha\in S_n: \prod_{i=1}^{n} \alpha(x_i) = \sup_{\beta\in S_n} \prod_{i=1}^{n} \beta(x_i)\}$
6. $S_n^2 = \{\alpha(x): \alpha(x) = \frac{1}{n} \sum_{j=1}^{n} \frac{1}{\sqrt{2\pi\sigma_n^2}} e^{-(x-y_j)^2/2\sigma_n^2}\}$
7. $A_n^2 = \{\alpha\in S_n^2: \prod_{i=1}^{n} \alpha(x_i) = \sup_{\beta\in S_n^2} \prod_{i=1}^{n} \beta(x_i)\}$

<u>Proposition 1</u>. A_n is not empty, and $\alpha\in A_n$

$$\Rightarrow \qquad \alpha(x) = \sum_{j=1}^{n} p_j \frac{1}{\sqrt{2\pi\sigma_n^2}} e^{-(x-y_j)^2/2\sigma_n^2}$$

for some $y_1,\ldots,y_n$ and $p_1,\ldots,p_n$ satisfying $p_i \geq 0$ $1 \leq i \leq n$, $\sum_{i=1}^{n} p_i = 1$. Furthermore, if $\min(x_1,\ldots x_n) < \max(x_1,\ldots x_n)$, then $\min(x_1,\ldots x_n) < \min(y_1,\ldots y_n) \leq \max(y_1,\ldots y_n) < \max(x_1,\ldots x_n)$.

<u>Proof</u>: Define $k(x) = \frac{1}{\sqrt{2\pi}} e^{-x^2/2}$, $x_M = \max(x_1,\ldots x_n)$, and $x_m = \min(x_1,\ldots x_n)$. If $x_M = x_m$, then obviously A_n contains only

$$\alpha(x) = \frac{1}{\sigma_n} k(\frac{x-x_m}{\sigma_n}).$$

Suppose $x_m < x_M$. Take $\alpha_k(x)$ $k=1,2,\ldots,$ such that $\alpha_k\in S_n$ and

$$\lim_{k\to\infty} \prod_{i=1}^{n} \alpha_k(x_i) = \sup_{\beta\in S_n} \prod_{i=1}^{n} \beta(x_i).$$

Define F_k, $k=1,2,\ldots,$ by

$$\alpha_k(x) = \int_{-\infty}^{\infty} \frac{1}{\sigma_n} k(\frac{x-y}{\sigma_n}) F_k(dy),$$

and then define $\tilde{F}_k$ by

$$\tilde{F}_k(B) = F_k(B) \quad \text{for} \quad B\subset(x_m,x_M)$$

$$\tilde{F}_k(\{x_m\}) = F_k((-\infty,x_m]), \quad \text{and}$$

$$\tilde{F}_k(\{x_M\}) = F_k([x_M,\infty)).$$

Finally, let

$$\tilde{\alpha}_k(x) = \int_{-\infty}^{\infty} \frac{1}{\sigma_k} k(\frac{x-y}{\sigma_n}) \tilde{F}_k(dy).$$

Then, clearly, $\tilde{\alpha}_k\in S_n$, $k=1,2,\ldots,$ and $\prod_{i=1}^{n} \tilde{\alpha}_k(x_i) \geq \prod_{i=1}^{n} \alpha_k(x_i)$. Hence, also,

$$\lim_{k\to\infty} \prod_{i=1}^{n} \tilde{\alpha}_k(x_i) = \sup_{\beta\in S_n} \prod_{i=1}^{n} \beta(x_i).$$

Since $\{\tilde{F}_k\}$ is tight, there exists F_∞ such that $\tilde{F}_k \to F_\infty$ weakly. If

$$\alpha_\infty(x) = \int_{-\infty}^{\infty} \frac{1}{\sigma_n} k(\frac{x-y}{\sigma_n}) F_\infty(dy),$$

then $\alpha_\infty \in S_n$ and $\alpha_\infty(x) = \lim_{k\to\infty} \tilde{\alpha}_k(x)$ at each x. Hence

$$\prod_{i=1}^{n} \alpha_\infty(x_i) = \sup_{\beta\in S_n} \prod_{i=1}^{n} \beta(x_i),$$

and it follows that A_n is not empty.

Take $\alpha \in A_n$ and let F be such that

$$\alpha(x) = \int_{-\infty}^{\infty} \frac{1}{\sigma_n} k(\frac{x-y}{\sigma_n}) F(dy).$$

For any s in the support of F, any $\varepsilon > 0$, and any z, define a measure $G_{\varepsilon,s,z}$ by

$$G_{\varepsilon,s,z}(B) = F([s-\varepsilon,s+\varepsilon] \cap (B-z))$$

($G_{\varepsilon,s,z}$ is a rigid right shift, by distance z, of the measure F restricted to $[s-\varepsilon,s+\varepsilon]$). Let $F_{\varepsilon,s} = F-G_{\varepsilon,s,0}$. Observe that $F_{\varepsilon,s}+G_{\varepsilon,s,z}$ is a distribution function for any z. If

$$\alpha_{\varepsilon,s,z}(x) = \int_{-\infty}^{\infty} \frac{1}{\sigma_n} k(\frac{x-y}{\sigma_n}) F_{\varepsilon,s}(dy)$$
$$+ \int_{-\infty}^{\infty} \frac{1}{\sigma_n} k(\frac{x-y}{\sigma_n}) G_{\varepsilon,s,z}(dy),$$

then $\alpha = \alpha_{\varepsilon,s,0}$, and since $\alpha \in A_n$ we must have

$$\frac{d}{dz} \sum_{i=1}^{n} \log \alpha_{\varepsilon,s,z}(x_i)\Big|_{z=0} = 0$$

i.e.

$$\sum_{i=1}^{n} \frac{1}{\alpha(x_i)} \frac{d}{dz} \int_{-\infty}^{\infty} \frac{1}{\sigma_n} k(\frac{x_i-y}{\sigma_n}) G_{\varepsilon,s,z}(dy)\Big|_{z=0} = 0$$

$$\Rightarrow \sum_{i=1}^{n} \frac{1}{\alpha(x_i)} \int_{[s-\varepsilon,s+\varepsilon]} \frac{(x_i-y)}{\sigma_n^3} k(\frac{x_i-y}{\sigma_n}) F(dy) = 0$$

$$\Rightarrow \sum_{i=1}^{n} \frac{1}{\alpha(x_i)} \frac{1}{F([s-\varepsilon,s+\varepsilon])} \int_{[s-\varepsilon,s+\varepsilon]} (x_i-y) k(\frac{x_i-y}{\sigma_n}) F(dy) = 0.$$

Now let $\varepsilon \downarrow 0$:

$$\sum_{i=1}^{n} \frac{(x_i-s)}{\alpha(x_i)} k(\frac{x_i-s}{\sigma_n}) = 0 \qquad \text{(A.25)}$$

for any s in the support of F.

Introduce the function

$$T(y) = \sum_{i=1}^{n} \frac{(x_i-y)}{\alpha(x_i)} k(\frac{x_i-y)}{\sigma_n}$$
$$= e^{-y^2/2\sigma_n^2} \{ \sum_{i=1}^{n} (\frac{x_i e^{-x_i^2/2\sigma_n^2}}{\alpha(x_i)}) e^{x_i y/\sigma_n^2}$$
$$- \sum_{i=1}^{n} (\frac{e^{-x_i^2/2\sigma_n^2}}{\alpha(x_i)}) y e^{x_i y/\sigma_n^2} \}.$$

The collection of functions

$$\{e^{x_i y/\sigma_n^2}\}_{i=1}^{n} \cup \{y e^{x_i y/\sigma_n^2}\}_{i=1}^{n}$$

forms an extended Tchebycheff system with at most 2n distinct elements. Two consequences for T are (c.f. Karlin and Studden [12]):

1. $z_1 \equiv \{y: \ T(y) = 0\}$ has at most 2n-1 elements,

and 2. $z_2 \equiv \{y: \ T(y) = 0, \ \frac{d}{dy} T(y) \leq 0\}$ has at most n elements.

Since the support of F lies in z_1(see (A.25)), F is discrete with at most 2n-1 jumps. We wish to show that, in fact, F has at most n jumps. Because of 2, it will be enough to show that

$$\frac{d}{dy} T(y)|_{y=s} \le 0$$

for any s in the support of F.

For α, we may now write

$$\alpha(x) = \sum_{j=1}^{q} p_j \frac{1}{\sigma_n} k(\frac{x-s_j}{\sigma_n}),$$

where $(s_1,\dots s_q)$ is the support of F, $q \le 2n-1$, $p_j > 0$, $j=1,2,\dots q$, and $p_1+p_2+\dots+p_q=1$. Fix $\ell\in(1,2,\dots q)$, and define for every $\varepsilon > 0$, $\alpha_\varepsilon(x)$ by

$$\alpha_\varepsilon(x) = \sum_{j\neq\ell} p_j \frac{1}{\sigma_n} k(\frac{x-s_j}{\sigma_n})$$

$$+ \frac{p_\ell}{2} \frac{1}{\sigma_n} k(\frac{x-s_\ell-\varepsilon}{\sigma_n}) + \frac{p_\ell}{2} \frac{1}{\sigma_n} k(\frac{x-s_\ell+\varepsilon}{\sigma_n}).$$

Then $\alpha_\varepsilon \in S_n$ and $\alpha=\alpha_0$. Hence

$$\frac{d^2}{d\varepsilon^2} \sum_{i=1}^{n} \log \alpha_\varepsilon(x_i)|_{\varepsilon=0} \le 0.$$

Straightforward calculation reveals that

$$\frac{d^2}{d\varepsilon^2} \sum_{i=1}^{n} \log \alpha_\varepsilon(x_i)|_{\varepsilon=0} = \frac{p_\ell}{\sigma_n^2} \frac{d}{dy} T(y)|_{y=s_\ell},$$

and consequently, as required,

$$\frac{d}{dy} T(y)|_{y=s} \le 0$$

for any s in the support of F.

Finally, for the last statement in the proposition, observe that if $s \le x_m$ for some s in the support of F, then for every i=1,2,...n, $\alpha(x_i)$ is strictly increased by a sufficiently small increase in s. Hence, $\prod_{i=1}^{n} \alpha(x_i)$ is also increased, contrading $\alpha\in A_n$. A similar argument precludes $s_\ell \ge x_M$ as well. □

<u>Proposition 2</u>. Let $\alpha_0(x)$ be a density function satisfying $\int_{-\infty}^{\infty}\alpha_0(x) \log \alpha_0(x)dx < \infty$. If, for each n, A_n is a collection of density functions, and if

$$\lim_{n\to\infty} \sup_{\alpha\in A_n} \int_{-\infty}^{\infty} \alpha_0(x) \log \frac{\alpha_0(x)}{\alpha(x)} dx = 0,$$

then also

$$\lim_{n\to\infty} \sup_{\alpha\in A_n} \int_{-\infty}^{\infty} |\alpha(x)-\alpha_0(x)| dx = 0.$$

<u>Proof</u>: For each c > 1 define x_c by $x_c > 1$ and $x_c - 1 = c(x_c-1-\log x_c)$. Observe that

$$x-1 < c(x-1-\log x) \text{ for all } x > x_c, \tag{A.26}$$

$$\lim_{c\to\infty} x_c = 1, \tag{A.27}$$

and

$$x-1-\log x \ge 0 \quad \text{for all } x \ge 0. \tag{A.28}$$

Choose $c_n \uparrow \infty$ such that

$$\lim_{n\to\infty} c_n \sup_{\alpha\in A_n} \int_{-\infty}^{\infty} \alpha_0(x) \log \frac{\alpha_0(x)}{\alpha(x)} dx = 0.$$

$$\overline{\lim_{n\to\infty}} \sup_{\alpha\in A_n} \int_{-\infty}^{\infty} |\alpha(x)-\alpha_0(x)|dx$$

$$= 2\, \overline{\lim_{n\to\infty}} \sup_{\alpha\in A_n} \int_{\alpha(x)>\alpha_0(x)} \alpha_0(x)\left(\frac{\alpha(x)}{\alpha_0(x)} - 1\right)dx$$

$$= 2\, \overline{\lim_{n\to\infty}} \sup_{\alpha\in A_n} \int_{1<\frac{\alpha(x)}{\alpha_0(x)} \le x_{c_n}} \alpha_0(x)\left(\frac{\alpha(x)}{\alpha_0(x)} -1\right)dx$$

$$+ 2\, \overline{\lim_{n\to\infty}} \sup_{\alpha\in A_n} \int_{\frac{\alpha(x)}{\alpha_0(x)} > x_{c_n}} \alpha_0(x)\left(\frac{\alpha(x)}{\alpha_0(x)} - 1\right)dx$$

$\le$ (use (A.26))

$$2\, \overline{\lim_{n\to\infty}}\, c_n \sup_{\alpha\in A_n} \int_{\frac{\alpha(x)}{\alpha_0(x)} > x_{c_n}} \alpha_0(x)\left(\frac{\alpha(x)}{\alpha_0(x)} - 1 - \log \frac{\alpha(x)}{\alpha_0(x)}\right)dx$$

$$+ 2\, \overline{\lim_{n\to\infty}} (x_{c_n} -1)$$

$\le$ (use (A.27) and (A.28))

$$2\, \overline{\lim_{n\to\infty}}\, c_n \sup_{\alpha\in A_n} \int_{-\infty}^{\infty} \alpha_0(x)\left(\frac{\alpha(x)}{\alpha_0(x)} - 1 - \log \frac{\alpha(x)}{\alpha_0(x)}\right)dx$$

$$= 2\, \overline{\lim_{n\to\infty}}\, c_n \sup_{\alpha\in A_n} \int_{-\infty}^{\infty} \alpha_0(x)\log \frac{\alpha_0(x)}{\alpha(x)}\, dx = 0. \qquad \square$$

<ins>Lemma 5</ins>. Let $\alpha_0(x)$ be a density function with support on $[0,1]$.

(a) For any sequence $\sigma_n \downarrow 0$ such that $n\sigma_n \to \infty$, there exists a sequence of density functions, β_n, such that $\beta_n \in S_n^2$ for all n, and

$$\int_0^1 \alpha_0(x)\log \beta_n(x)dx \to \int_0^1 \alpha_0(x)\log \alpha_0(x)dx$$

(finite or infinite),

and (b) for any sequence $\sigma_n \downarrow 0$ such that for some $\varepsilon > 0$ $n^{1/4-\varepsilon}\sigma_n \to \infty$,

$$\lim_{n\to\infty} \sup_{\alpha\in A_n^2} \left|\frac{1}{n}\sum_{i=1}^{n} \log \alpha(x_i) - \int_0^1 \alpha_0(x)\log \alpha(x)dx\right| = 0$$

a.s.

<ins>Proof</ins>: There is nothing new in the proof of (b), just another application of the "small ball" technique as it was used in Theorems 1 and 2. We will forego the details.

For (a), first define a new density, $\alpha_\varepsilon(x)$, as follows. For any $0 < \varepsilon < 1$, let

$$\hat{\alpha}_\varepsilon(x) = \begin{cases} 0 & x\notin[0,1] \\ \alpha_0(x) & x\in[0,1] \text{ and } \varepsilon \le \alpha_0(x) \le 1/\varepsilon \\ \varepsilon & x\in[0,1] \text{ and } \alpha_0(x) < \varepsilon \\ \frac{1}{\varepsilon} & x\in[0,1] \text{ and } 1/\varepsilon < \alpha_0(x). \end{cases}$$

Take $\varepsilon_0 > 0$ sufficiently small that

$$c_\varepsilon \equiv \int_0^1 \hat{\alpha}_\varepsilon(x)dx > 0 \quad \text{whenever} \quad 0 < \varepsilon < \varepsilon_0,$$

and define $\alpha_\varepsilon(x) = \hat{\alpha}_\varepsilon(x)/c_\varepsilon$ for all $0 < \varepsilon < \varepsilon_0$. Observe that $c_\varepsilon \uparrow 1$ as $\varepsilon \downarrow 0$.

We claim that

$$\int_0^1 \alpha_0(x)\log \alpha_\varepsilon(x)dx \to \int_0^1 \alpha_0(x)\log \alpha_0(x)dx \tag{A.29}$$

(finite or infinite) as $\varepsilon \to 0$. If $\int_0^1 \alpha_0(x)\log \alpha_0(x)dx < \infty$, then by Jensen's inequality,

$$\int_0^1 \alpha_0(x)\log \alpha_\varepsilon(x)dx - \int_0^1 \alpha_0(x)\log \alpha_0(x)dx$$

$$= \int_0^1 \alpha_0(x)\log(\frac{\alpha_\varepsilon(x)}{\alpha_0(x)})dx \le 0.$$

And, whether or not $\int_0^1 \alpha_0(x)\log \alpha_0(x)dx < \infty$:

$$\int_0^1 \alpha_0(x)\log \alpha_\varepsilon(x)dx \ge \int_0^1 \alpha_0(x)\log \hat{\alpha}_\varepsilon(x)dx$$

$$\ge \int_{\{\alpha_0(x)\le \frac{1}{\varepsilon}\}} \alpha_0(x)\log \alpha_0(x)dx + \int_{\{\alpha_0(x)>\frac{1}{\varepsilon}\}} \alpha_0(x)\log \frac{1}{\varepsilon}\, dx$$

$$\ge \int_{\{\alpha_0(x)\le \frac{1}{\varepsilon}\}} \alpha_0(x)\log \alpha_0(x)dx \to \int_0^1 \alpha_0(x)\log \alpha_0(x)dx$$

as $\varepsilon \to 0$, which proves (A.29).

Now define, for each $0 < \varepsilon < \varepsilon_0$ and each $n=1,2,\ldots,$

$$\gamma_{\varepsilon,n}(x) = \int_0^1 \frac{1}{\sqrt{2\pi\sigma_n^2}} \exp\{\frac{1}{2\sigma_n^2}(x-y)^2\}\alpha_\varepsilon(y)dy.$$

Then (Stein [15], Theorem 2, page 62)

$$\gamma_{\varepsilon,n}(x) \to \alpha_\varepsilon(x) \quad \text{a.e. } dx$$

as $n \to \infty$. Because $\frac{\varepsilon}{c_\varepsilon} \le \alpha_\varepsilon(x) \le \frac{1}{\varepsilon c_\varepsilon}$,

$$\gamma_{\varepsilon,n}(x) \le \frac{1}{\varepsilon c_\varepsilon},$$

and, whenever $\sigma_n < 1$, $x \in [0,1] \Rightarrow$

$$\gamma_{\varepsilon,n}(x) \ge \frac{\varepsilon}{c_\varepsilon}\int_0^1 \frac{1}{\sqrt{2\pi\sigma_n^2}} \exp\{\frac{1}{2\sigma_n^2}(x-y)^2\}dy \qquad (A.30)$$

$$\ge \frac{\varepsilon}{4c_\varepsilon} \ge \frac{\varepsilon}{4}.$$

By dominated convergence

$$\int_0^1 \alpha_0(x)\log \gamma_{\varepsilon,n}(x)dx \to \int_0^1 \alpha_0(x)\log \alpha_\varepsilon(x)dx \qquad (A.31)$$

as $n \to \infty$.

Define $y_0=0, y_n=1$, and, for each $0 < \varepsilon < \varepsilon_0$, choose $y_1,\ldots y_{n-1}$ such that

$$\int_0^{y_k} \alpha_\varepsilon(x)dx = \frac{k}{n}.$$

Let

$$\xi_{\varepsilon,n}(x) = \frac{1}{n}\sum_{k=1}^{n} \frac{1}{\sqrt{2\pi\sigma_n^2}}\, e^{-(x-y_k)^2/2\sigma_n^2}.$$

Then, if $x \in [y_{\ell-1}, y_\ell]$ for some $\ell=1,2,\ldots n$,

$$|\gamma_{\varepsilon,n}(x)-\xi_{\varepsilon,n}(x)|$$

$$= |\sum_{k=1}^{n}\int_{y_{k-1}}^{y_k} \{\frac{1}{\sqrt{2\pi\sigma_n^2}}\, e^{-(x-y)^2/2\sigma_n^2} - \frac{1}{\sqrt{2\pi\sigma_n^2}}\, e^{-(x-y_k)^2/2\sigma_n^2}\}\alpha_\varepsilon(y)dy|$$

$$\le \frac{1}{\sigma_n\sqrt{2\pi}} \sum_{k=1}^{n}\int_{y_{k-1}}^{y_k} |e^{-(x-y)^2/2\sigma_n^2} - e^{-(x-y_k)^2/2\sigma_n^2}|\, \alpha_\varepsilon(y)dy$$

$$\leq \frac{1}{\sigma_n\sqrt{2\pi}} \{\int_{y_{\ell-1}}^{y_\ell} |e^{-(x-y)^2/2\sigma_n^2} - e^{-(x-y_\ell)^2/2\sigma_n^2}|\alpha_\varepsilon(y)dy$$

$$+ \sum_{k=\ell+1}^{n} \int_{y_{k-1}}^{y_k} (e^{-(x-y_{k-1})^2/2\sigma_n^2} - e^{-(x-y_k)^2/2\sigma_n^2})\alpha_\varepsilon(y)dy$$

$$+ \sum_{k=1}^{\ell-1} \int_{y_{k-1}}^{y_k} (e^{-(x-y_k)^2/2\sigma_n^2} - e^{-(x-y_{k-1})^2/2\sigma_n^2})\alpha_\varepsilon(y)dy\}$$

$$\leq \frac{1}{n\sigma_n\sqrt{2\pi}} \{1 + \sum_{k=\ell+1}^{n} (e^{-(x-y_{k-1})^2/2\sigma_n^2} - e^{-(x-y_k)^2/2\sigma_n^2})$$

$$+ \sum_{k=1}^{\ell-1} (e^{-(x-y_k)^2/2\sigma_n^2} - e^{-(x-y_{k-1})^2/2\sigma_n^2})\}$$

$$= \frac{1}{n\sigma_n\sqrt{2\pi}} \{1 + e^{-(x-y_\ell)^2/2\sigma_n^2} - e^{-(x-y_n)^2/2\sigma_n^2}$$

$$+ e^{-(x-y_{\ell-1})^2/2\sigma_n^2} - e^{-x^2/2\sigma_n^2}\}$$

$$\leq \frac{3}{n\sigma_n\sqrt{2\pi}} .$$

i.e. $$\sup_{x\in[0,1]} |\gamma_{\varepsilon,n}(x)-\xi_{\varepsilon,n}(x)| \leq \frac{3}{n\sigma_n\sqrt{2\pi}} . \qquad (A.32)$$

Consequently:

$$|\int_0^1 \alpha_0(x)\log\xi_{\varepsilon,n}(x)dx - \int_0^1 \alpha_0(x)\log\gamma_{\varepsilon,n}(x)dx|$$

$$= |\int_0^1 \alpha_0(x)\log\{1 + \frac{\xi_{\varepsilon,n}(x)-\gamma_{\varepsilon,n}(x)}{\gamma_{\varepsilon,n}(x)}\}dx|$$

$$\leq \sup_{x\in[0,1]} |\log\{1 + \frac{\xi_{\varepsilon,n}(x)-\gamma_{\varepsilon,n}(x)}{\gamma_{\varepsilon,n}(x)}\}|$$

$$\leq -\log\{1 - \sup_{x\in[0,1]} |\frac{\xi_{\varepsilon,n}(x)-\gamma_{\varepsilon,n}(x)}{\gamma_{\varepsilon,n}(x)}|\}$$

$\leq$ (use (A.30) and (A.32))

$$-\log\{1 - \frac{12}{\varepsilon n\sigma_n\sqrt{2\pi}}\} \qquad (A.33)$$

whenever $\sigma_n < 1$.

Finally, choose a sequence $\varepsilon_n \downarrow 0$ sufficiently slowly that

1. $\varepsilon_n n\sigma_n \to \infty$, and
2. $|\int_0^1 \alpha_0(x)\log\gamma_{\varepsilon_n,n}(x)dx - \int_0^1 \alpha_0(x)\log\alpha_{\varepsilon_n}(x)dx| \to 0.$

Take $\beta_n = \xi_{\varepsilon_n,n} \in S_n^2$, and apply (A.29) and (A.33):

$$|\int_0^1 \alpha_0(x)\log\beta_n(x)dx - \int_0^1 \alpha_0(x)\log\alpha_0(x)dx|$$

$$\leq |\int_0^1 \alpha_0(x)\log\alpha_{\varepsilon_n}(x)dx - \int_0^1 \alpha_0(x)\log\alpha_0(x)dx|$$

$$+ |\int_0^1 \alpha_0(x)\log\gamma_{\varepsilon_n,n}(x)dx - \int_0^1 \alpha_0(x)\log\alpha_{\varepsilon_n}(x)dx|$$

$$+ |\int_0^1 \alpha_0(x)\log\xi_{\varepsilon_n,n}(x)dx - \int_0^1 \alpha_0(x)\log\gamma_{\varepsilon_n,n}(x)dx|$$

$$\to 0 \text{ as } n \to \infty .$$

□