RC 20751 (02/26/97)
Computer Science/Mathematics

IBM Research Report

On the Bias and Variance of Finite Element Discretizations for FFT-Based Kernel Density Estimation

Ramesh Natarajan

IBM Research Division
T.J. Watson Research Center
Yorktown Heights, New York

Research Division
Almaden · T.J. Watson · Tokyo · Zurich

On the Bias and Variance of Finite Element Discretizations for FFT-Based Kernel Density Estimation

by

Ramesh Natarajan
IBM T. J. Watson Research Center
P. O. Box 218
Yorktown Heights, N.Y. 10598.
e-mail: ramesh@watson.ibm.com

Abstract

A computationally-efficient procedure for kernel density estimation using the FFT algorithm has been given by Silverman ([10]), with extensions by Jones and Lotwick ([7]). This procedure requires the empirical characteristic function to be interpolated on a regular mesh which leads to high-frequency approximation errors in this step. In the case of density estimation, these high-frequency errors are subsequently damped by the smoothing effect of the kernel multiplier. Nevertheless, these small errors can lead to a significant loss of accuracy whenever the kernel density estimates or its derivatives are used as a part of some larger statistical procedure. In this paper, we describe systematic finite element discretization procedures for improving the accuracy of the FFT-based algorithms. We derive the bias and variance of the kernel density estimates for the FFT-based algorithms, and this analysis suggests modifications to the computational procedure to obtain estimates free from interpolation bias. Simulation studies that verify the results of the analysis are presented. Finally, an application that motivated the study described in this paper is discussed.

1 Introduction

An efficient computational procedure for kernel density estimation using the FFT algorithm has been given by Silverman ([10]), along with some extensions by Jones and Lotwick ([7]). This algorithm requires the empirical characteristic function to be interpolated on a uniform mesh, which introduces high-frequency approximation errors. In the case of density estimation, these high-frequency errors are subsequently damped by the smoothing effect of the kernel multiplier, but nevertheless, they can still be significant in applications where the density estimates and its derivatives are required as part of some larger statistical procedure. For example, Silverman ([11]) describes several applications of the use of density estimation in statistical procedures such as discriminant analysis, clustering, bump-hunting and projection pursuit.

Our specific interest in this error analysis arose in the context of using higher-order optimization methods in projection pursuit analysis. There, the usual low-order discretization methods for the FFT-based kernel density estimates lead to inaccurate and even inconsistent results in the evaluation of the gradient and Hessian for the optimization procedure (see Section (4)).

Let $g_N(x)$ be the N-point sample kernel density estimate of the density function $g(x)$, so that

$$g_N(x) = \frac{1}{Nh}\sum_{i=1}^{N} K\left(\frac{x - X_i}{h}\right), \tag{1}$$

where $\{X_i\}_{i=1}^N$ denote *i.i.d.* random variables sampled from $g(x)$, and h denotes the kernel bandwidth parameter. In this paper, we are specifically concerned with symmetric, non-negative kernels, so that

$$\begin{gathered} K(x) \geq 0, \quad \int_{-\infty}^{\infty} K(x)dx = 1, \\ \int_{-\infty}^{\infty} x^n K(x)dx = 0, \quad \text{for } n \text{ odd.} \end{gathered} \tag{2}$$

Since the regularity of the kernel $K(x)$ is inherited by $g_N(x)$, a suitable choice in the case when high-order differentiability is required is the Gaussian kernel

$$K(x) = (2\pi)^{-1/2} e^{-x^2/2}. \tag{3}$$

The Fourier transform of this particular kernel is explicitly known, which is advantageous for the FFT-based computational algorithm described below.

We note that the straightforward direct evaluation of $g_N(x)$ at M points from (1) requires $O(MN)$ operations. In contrast, for M points on a fixed regular grid, the FFT-based algorithm requires $O(M \log M + N)$ operations. Greengard and Strain [4] have described a "fast Gauss transform" algorithm, that for arbitrarily distributed M points requires only $O(N+M)$ operations (with the constant proportionality factor depending on the desired accuracy). For the multivariate case, the Greengard-Strain algorithm is likely to be method of choice, but for the typical values of M of interest in the univariate case, its performance advantage over the FFT-based algorithm is unclear, particularly given its greater programming complexity. In addition, the specific use of the properties of Hermite expansions to the Gaussian kernel in the Greengard-Strain algorithm renders it unusable for other choices of the kernel function.

Now, denoting the Fourier transform by

$$\widetilde{g_N}(s) = (2\pi)^{-1/2} \int_{-\infty}^{\infty} e^{isx} g_N(x)dx, \tag{4}$$

we have from (1),

$$\widetilde{g_N}(s) = \widetilde{K}(hs) \sum_{i=1}^{N} \frac{e^{isX_i}}{N}, \tag{5}$$

which in the specific case of the Gaussian kernel, this yields

$$\widetilde{g_N}(s) = (2\pi)^{-1/2} e^{-h^2 s^2/2} \sum_{i=1}^{N} \frac{e^{isX_i}}{N}. \tag{6}$$

Denoting the empirical characteristic function by

$$\gamma_N(s) = \sum_{i=1}^{N} \frac{e^{isX_i}}{N}, \tag{7}$$

it can be seen from (6) that the components of $\gamma_N(s)$ for $hs \gg 1$ are strongly damped in the evaluation of $\widetilde{g_N}(s)$ (with the magnitude of h controlling the range of the frequencies that are damped, and hence the resulting smoothness of the kernel density estimate). This feature is advantageously used in the FFT-based algorithms, where the empirical characteristic function is replaced by another estimate with possible high-frequency errors, since these errors are in any case damped by convolution with the Gaussian kernel. Nevertheless, as shown below, depending on the magnitude of h, these high-frequency errors can still be crucial in determining the convergence and accuracy of the resulting approximations to the density estimate and its derivatives from (1).

2 Discretization and Computational Details

Consider the interval (a, b), which is chosen large enough to completely contain the sample data $\{X_i\}_{i=1}^N$. This interval is partitioned into $M = 2^m$ equal subintervals, where $\delta = (b-a)/M$ is the discretization parameter, and the grid points are given by $x_k = a + k\delta$, $k = 0, 1, \cdots, M-1$, respectively.

A finite element interpolation of the empirical characteristic function on this grid yields

$$\gamma_N(s) = \frac{1}{N}\sum_{i=1}^{N} e^{isX_i} \approx \sum_{k=0}^{M-1} \xi_k e^{isx_k}, \tag{8}$$

where

$$\xi_k = \frac{1}{N}\sum_{i=1}^{N} \phi_k(X_i), \tag{9}$$

with $\phi_k(x)$ denoting the finite-element basis function associated with the grid point x_k. Suitable basis functions include, for example, the piecewise constant $\phi_k^{(0)}$ (Silverman [11]), piecewise linear $\phi_k^{(1)}$ (Jones and Lotwick [6]), and piecewise quadratic $\phi_k^{(2)}$ cases (considered in this paper for the first time). In these specific cases, denoting

$$\mu_k = \frac{x - k\delta}{\delta} \tag{10}$$

we have in terms of the indicator function χ,

$$\begin{aligned}
\phi_k^{(0)}(x) &= \chi_{[x_{k-\frac{1}{2}}, x_{k+\frac{1}{2}}]}(x), && (11)\\
\phi_k^{(1)}(x) &= (1-|\mu_k|)\chi_{[x_{k-1}, x_{k+1}]}(x), && (12)\\
\phi_k^{(2)}(x) &= \begin{cases} \frac{1}{2}(1-|\mu_k|)(2-|\mu_k|)\chi_{[x_{k-2}, x_{k+2}]}(x), & \text{even k,} \\ (1-|\mu_k|)(1+|\mu_k|)\chi_{[x_{k-1}, x_{k+1}]}(x), & \text{odd k,} \end{cases} && (13)
\end{aligned}$$

and these are shown plotted in Figure (1).

The values of ξ_k in (9) are computed by histogramming and interpolation. For example, in the piecewise quadratic case, we have the following procedure

1. Set $\xi_k = 0$ for all k.
2. For each $i = 1, N$

   If $X_i \in [x_{2m}, x_{2m+2}]$, then with $k = 2m$ and $\alpha = 1/2N\delta^2$,

$$\begin{aligned}
\xi_k &\longleftarrow \xi_k + \alpha(x_{k+1} - X_i)(x_{k+2} - X_i) \\
\xi_{k+1} &\longleftarrow \xi_{k+1} + \alpha(x_{k+2} - X_i)(X_i - x_k) \\
\xi_{k+2} &\longleftarrow \xi_{k+2} + \alpha(X_i - x_k)(X_i - x_{k+1})
\end{aligned}$$

   Endif

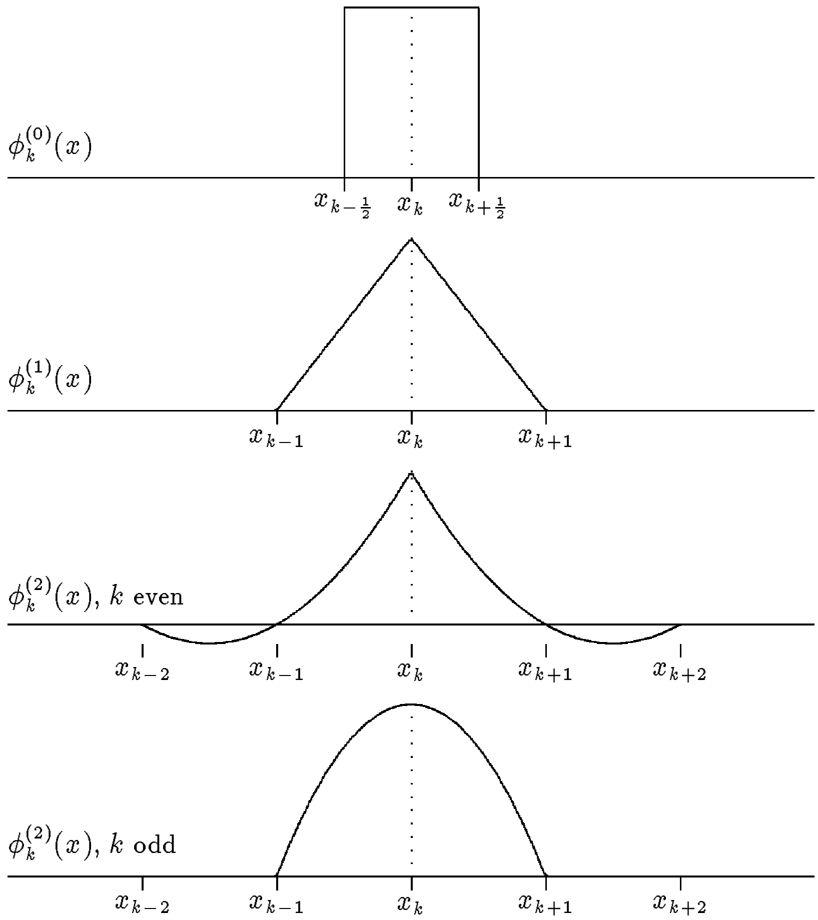

Figure 1: Finite element basis functions used in the discretization

Endfor

Now consider the following modification to the approximation in (8), defined by

$$\gamma_N^*(s) = e^{isa}\beta(s)\sum_{k=0}^{M-1}\xi_k e^{is\delta k}. \tag{14}$$

Here the frequency-dependent factor $\beta(s)$ has been introduced to adjust for any bias introduced by the interpolation (as discussed in detail below). We note that the unadjusted case in (8) can be recovered by merely setting $\beta(s) = 1$.

We consider this sum for the discrete values of s given by

$$s_j = \frac{2\pi j}{(b-a)} = \frac{2\pi j}{M\delta}, \text{ for } j = 0, \ldots, M-1, \tag{15}$$

since, if we assume that the kernel density estimate can be periodically extended outside the interval (a, b), then these discrete values are sufficient to reconstruct the kernel density estimate on (a, b). In practice, this is hardly a restrictive assumption since (a, b) can be chosen sufficiently larger than the range of the sample values. Then, denoting $\beta_j \equiv \beta(s_j)$, we have

$$\gamma_N^*(s_j) = e^{is_j a}\beta_j\sum_{k=0}^{M-1}\xi_k e^{i\frac{2\pi jk}{M}} = e^{is_j a}\beta_j\hat{\xi}_j, \tag{16}$$

in which the sum is evaluated in terms of the DFT components $\{\hat{\xi}_j\}_{j=0}^{M-1}$ of the sequence $\{\xi_j\}_{j=0}^{M-1}$.

Now, the Fourier integral in (14) can be discretized by the trapezoidal rule,

$$g_N(x_k) \approx \frac{(2\pi)^{\frac{1}{2}}}{M\delta}\sum_{j=-M/2+1}^{M/2}\hat{g}_j e^{-is_j x_k} = \frac{(2\pi)^{\frac{1}{2}}}{M\delta}\sum_{j=-M/2+1}^{M/2}\hat{g}_j e^{-is_j a}e^{-i\frac{2\pi jk}{M}} = g_N^*(x_k), \tag{17}$$

in which all the modes upto the Nyquist frequency on the grid have been retained (although in practice, the contribution of the higher-order modes on the grid to this integral will be negligibly small). Finally, combining (16) and (17) yields in the specific case of the Gaussian kernel,

$$g_N^*(x_k) = \frac{1}{M\delta}\sum_{j=-M/2+1}^{M/2}\left[\beta_j\xi_j e^{-\frac{1}{2}\left(\frac{2\pi h}{M\delta}\right)^2 j^2}\right]e^{-i\frac{2\pi jk}{M}}. \tag{18}$$

This sum can be evaluated by an inverse DFT to obtain the required approximation to the kernel density estimates at the grid points x_k.

3 Analysis of Discretization Errors

The estimates $g_N(x)$ and $g_N^*(x)$ along with their derivatives are all functions of the *i.i.d* random variables $\{X_i\}_{i=1}^N$. Therefore, letting $g_N^{(r)}(x)$ and $g_N^{*(r)}(x)$ denote the corresponding r'th derivatives, the integrated mean square discretization error $(I.M.S.D.E.)$ is given by

$$\begin{aligned} I.M.S.D.E. \equiv \mathcal{M}^{(r)} &= E\int_{-\infty}^{\infty}(g_N^{(r)}(x) - g_N^{*(r)}(x))^2 dx \\ &= \int_{-\infty}^{\infty} s^{2r}\widetilde{K}(hs)^2 E(|\gamma_N(s) - \gamma_N^*(s)|^2)ds, \end{aligned} \tag{19}$$

with the second step following from Parseval's theorem. Now, we have the following standard results for the expectation and variance of the empirical characteristic function,

$$E(\gamma_N(s)) = \gamma(s), \quad E(|\gamma_N(s)|^2) = \frac{1}{N} + \frac{N-1}{N}|\gamma(s)|^2, \tag{20}$$

where $\gamma(s) = E(e^{isx})$ denotes the characteristic function of $g(x)$. Using these, yields after some manipulations (Lotwick and Jones [6])

$$\begin{aligned} E(|\gamma_N(s) - \gamma_N^*(s)|^2) &= E(|\gamma_N(s)|^2) + E(|\gamma_N^*(s)|^2) - E(\gamma_N(s)\overline{\gamma_N^*(s)}) - E(\overline{\gamma_N(s)}\gamma_N^*(s)) \\ &= \frac{1}{N} + \frac{N-1}{N}|\gamma(s) - \beta(s)\sum_{k=0}^{M-1} E(\phi_k(x))e^{isx_k}(s)|^2 \\ &\quad -\frac{2\beta(s)}{N}\sum_{k=0}^{M-1} E(\phi_k(x)\cos s(x-x_k)) + \frac{\beta(s)^2}{N}\sum_{k=0}^{M-1} E(\phi_k(x)^2) \\ &\quad +\frac{2\beta(s)^2}{N}\sum_{k=0}^{M-1}\sum_{k<l}^{M-1} E(\phi_k(x)\phi_l(x))\cos s(x_k - x_l)). \end{aligned} \tag{21}$$

The arrangement of the terms in (21) exposes a square bias term, since the evaluation of the bias correction factor $\beta(s)$ is also of interest below. The various terms in (21) can now be evaluated using the following lemma, which is a direct consequence of the Euler-Maclaurin summation formula for periodic functions

Lemma 3.1 *Let the density function $g(x) \in C^\infty(a,b)$ be extended periodically outside this interval, and let $g^{(n)}(x), n = 0, 1, \ldots$ denote its derivatives. Then, using the notation $g_k^{(n)} \equiv g^{(n)}(x_k), k = 0, 1, \ldots, M-1$, we have*

$$(-is_j)^n\gamma_j = \sum_{k=0}^{M-1}\delta g_k^{(n)} e^{is_jx_k}, \quad \textit{for } s_j = \frac{2\pi j}{b-a}, \quad j = 0, 1, \ldots, M-1, \tag{22}$$

where $\gamma_j \equiv \gamma(s_j) = \int_a^b g(x) e^{is_jx} dx$. *Similarly, for even* M,

$$\frac{1}{2}(-is_j)^n \gamma_j = \sum_{\substack{k=0 \\ k\,even}}^{M-2} \delta g_k^{(n)} e^{is_jx_k} = \sum_{\substack{k=1 \\ k\,odd}}^{M-1} \delta g_k^{(n)} e^{is_jx_k}. \tag{23}$$

Finally, $\gamma_0 \equiv \gamma(s_0) = 1$.

First, consider the evaluation of the bias terms in (21), which are given by the following lemma

Lemma 3.2 *For* $s_j = 2\pi j/(b-a)$, $j = 0, 1, \ldots, M-1$, *let* $\gamma_{N,j}^{[i]} \equiv \gamma_N^{[i]}(s_j), i = 0, 1, 2$ *respectively denote the approximation to* $\gamma_N(s_j)$ *obtained by using piecewise constant, linear and quadratic basis functions as in (11)–(13). Then,*

$$E(\gamma_{N,j}^{[i]}) = \alpha_j^{[i]} \beta_j \gamma_j, \text{ for } i = 0, 1, 2, \tag{24}$$

where

$$\alpha_j^{[0]} = \frac{2 \sin \frac{1}{2} \delta s_j}{\delta s_j}, \tag{25}$$

$$\alpha_j^{[1]} = \frac{2(1 - \cos \delta s_j)}{(\delta s_j)^2}, \tag{26}$$

$$\alpha_j^{[2]} = \frac{(2 \sin \delta s_j - \sin 2\delta s_j)}{(\delta s_j)^3} + \frac{(3 - 4 \cos \delta s_j + \cos 2\delta s_j)}{2(\delta s_j)^2}. \tag{27}$$

We sketch the details of the proof for the piecewise quadratic case. First, note that

$$E(\gamma_{N,j}^{[2]}) = \beta_j \sum_{k=0}^{M-1} E(\phi_k^{[2]}(x)) e^{is_jx_k}, \tag{28}$$

and in the case of even k,

$$\begin{aligned} E(\phi_k^{[2]}(x)) &= \int_{(k-2)\delta}^{(k+2)\delta} \frac{1}{2}(1 - |\frac{x-k\delta}{\delta}|)(2 - |\frac{x-k\delta}{\delta}|) g(x) dx \\ &= \delta \int_0^2 \frac{1}{2}(1-z)(2-z) \left[\cosh \delta z D\right] g_k dz \\ &= \left[\frac{2 \sinh 2\delta D}{(\delta D)^3} - \frac{(3 + \cosh 2\delta D)}{(\delta D)^2}\right] \delta g_k. \end{aligned} \tag{29}$$

Here, the integral has been evaluated using the variable transformation $z = x - k\delta$, followed by the Taylor's series expansion of $g(x)$ within the integrand, and finally a term-by-term integration. The last two steps have been simplified by using the symbolic notation

$$[e^{zD}]g_k = g(k\delta + z) = \sum_{n=1}^{\infty} \frac{z^n}{n!} g_k^{(n)}. \tag{30}$$

Similarly, in the case of odd k,

$$\begin{aligned} E(\phi_k^{[2]}(x)) &= \int_{(k-1)\delta}^{(k+1)\delta} (1 - |\frac{x - k\delta}{\delta}|)(1 + |\frac{x - k\delta}{\delta}|) g(x) dx \\ &= 2\delta \int_0^1 \frac{1}{2}(1 - z)(1 + z) \left[\cosh \delta z D\right] g_k dz \\ &= 4 \left[\frac{\cosh \delta D}{(\delta D)^2} - \frac{(\sinh \delta D)}{(\delta D)^3} \right] \delta g_k. \end{aligned} \tag{31}$$

Substituting (29) and (31) in (28), we then note that from (23) in Lemma (3.1) that the symbol D can be replaced by $(-is_j)$ to yield the desired result in (27).

For the case $\beta_j = 1$ (i.e., without introducing a bias adjustment factor), the $E(\gamma_{N,j}^{[i]})$ are unbiased estimates of γ_j only for $\delta s_j \longrightarrow 0$. An unbiased estimator for all δs_j can be obtained by taking $\beta_j = (\alpha_j^{[i]})^{-1}$, but in practice it is often sufficient to use the following approximations for $\delta s_j \ll 1$,

$$\beta_j^{[0]} \approx 1 + \frac{1}{24}(\delta s_j)^2 + \frac{7}{5760}(\delta s_j)^4 + \frac{31}{96780}(\delta s_j)^6 + O((\delta s_j)^8), \tag{32}$$

$$\beta_j^{[1]} \approx 1 + \frac{1}{12}(\delta s_j)^2 + \frac{1}{240}(\delta s_j)^4 + \frac{1}{6048}(\delta s_j)^6 + O((\delta s_j)^8), \tag{33}$$

$$\beta_j^{[2]} \approx 1 + \frac{1}{60}(\delta s_j)^4 - \frac{13}{7560}(\delta s_j)^6 + O((\delta s_j)^8). \tag{34}$$

These approximations are more computationally economical, and as shown in Figure (2) adequately provide the necessary bias adjustment in the low-to-moderate frequency regime of interest in kernel density estimation.

With these results in hand, the square bias term in (21) is then given by

$$|\gamma(s_j) - E(\gamma_N(s_j)|^2 = (1 - \beta_j^{[i]}\alpha_j^{[i]})^2 |\gamma_j|^2, \text{ for } i = 0, 1, 2. \tag{35}$$

The evaluation of the remaining terms in (21) uses similar manipulations as in (29)–(31) above. For brevity, we only illustrate as an example the evaluation of the fourth term in (21) involving the expectation of the self-overlap of the basis functions for the case of $\phi_k^{[2]}(x)$. First, consider the summation for the even terms,

$$\sum_{\substack{k=0 \\ k\,even}}^{M-2} E(\phi_k^{[2]}(x)^2) = \sum_{\substack{k=0 \\ k\,even}}^{M-2} \int_{(k-2)\delta}^{(k+2)\delta} \frac{1}{4}(1 - |\frac{x - k\delta}{\delta}|)^2 (2 - |\frac{x - k\delta}{\delta}|)^2 g(x) dx$$

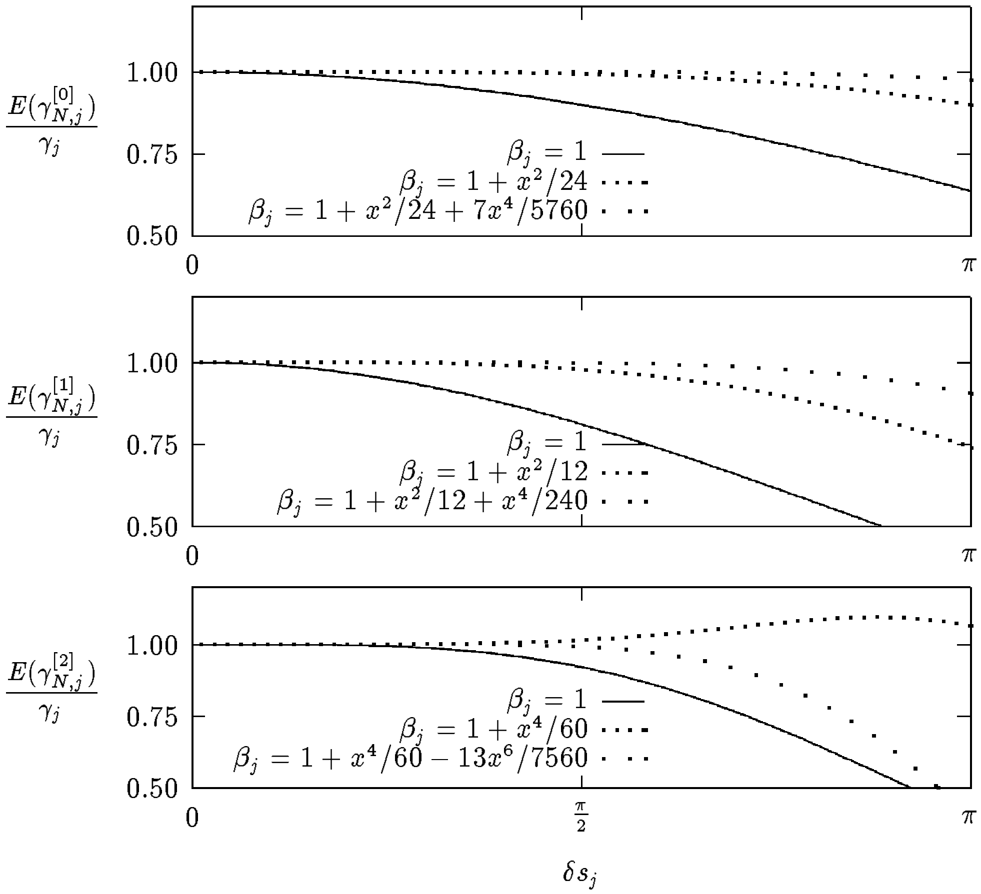

Figure 2: Effect of low-frequency approximation to the bias adjustment factor on the Expectation of the interpolated estimates for the empirical characteristic function

$$
\begin{aligned}
&= \sum_{\substack{k=0\\ k\,even}}^{M-2} \delta \int_0^2 \frac{1}{2}\delta(1-z)^2(2-z)^2 \left[\cosh \delta z D\right] g_k dz \\
&= \frac{4}{15}. \qquad (36)
\end{aligned}
$$

Here, in the second step above, a simplification of the algebra is obtained by noting that by virtue of Lemma (3.1), a non-zero contribution to the integral arises only from the constant term in the Taylor's series expansion for $g(x)$ within the integrand. Therefore, it is appropriate to set the value of the symbol D to zero in evaluating the integral. Similarly, for the summation of the odd terms, we have

$$
\begin{aligned}
\sum_{\substack{k=1\\ k\,odd}}^{M-1} E(\phi_k^{[2]}(x)^2) &= \sum_{\substack{k=1\\ k\,odd}}^{M-1} \int_{(k-1)\delta}^{(k+1)\delta} (1-|\frac{x-k\delta}{\delta}|)^2(1+|\frac{x-k\delta}{\delta}|)^2 g(x) dx \\
&= \sum_{\substack{k=1\\ k\,odd}}^{M-1} \delta \int_0^1 2\delta(1-z)^2(1+z)^2 \left[\cosh \delta z D\right] g_k dz \\
&= \frac{8}{15}. \qquad (37)
\end{aligned}
$$

Finally, adding (36) and (37) yields the desired result for the required self-overlap term in (21). The final results can be summarized in the following two lemmas

Lemma 3.3 *For $s_j = 2\pi j/(b-a), j = 0, 1, \ldots, M-1$, the mean square discretization error of the characteristic function is given by*

$$
E(|\gamma_{N,j} - \gamma_{N,j}^{[i]}|^2) = \frac{1}{N} + \frac{N-1}{N}(1-\beta_j^{[i]}\alpha_j^{[i]})^2|\gamma_j|^2 - \frac{2}{N}\beta_j^{[i]}\alpha_j^{[i]} + \frac{1}{N}(\beta_j^{[i]})^2\psi_j^{[i]}, \quad (38)
$$

for $i = 0, 1, 2$, corresponding to the case of piecewise constant, piecewise linear and piecewise quadratic basis functions respectively. In (38), we have denoted

$$
\psi_j^{[0]} = 1, \quad \psi_j^{[1]} = \frac{1}{3}(2+\cos\delta s_j), \quad \psi_j^{[2]} = \frac{1}{15}(12 + 4\cos\delta s_j - \cos 2\delta s_j). \quad (39)
$$

Lemma 3.4 *For $\delta s_j \longrightarrow 0$, from (38) and (39) we have the following asymptotic estimates*

$$
\begin{aligned}
E(|\gamma_{N,j} - \gamma_{N,j}^{[0]}|^2) &= \frac{1}{12N}(\delta s_j)^2 + O((\delta s_j)^4), with\ \beta_j^{[0]} = 1, \qquad (40) \\
E(|\gamma_{N,j} - \gamma_{N,j}^{[1]}|^2) &= [\frac{1}{120N} + \frac{N-1}{144N}|\gamma_j|^2](\delta s_j)^4 + O((\delta s_j)^6), with\ \beta_j^{[1]} = 1, \qquad (41) \\
E(|\gamma_{N,j} - \gamma_{N,j}^{[2]}|^2) &= \frac{2}{945N}(\delta s_j)^6 + O((\delta s_j)^8), with\ \beta_j^{[0]} = 1. \qquad (42)
\end{aligned}
$$

The leading order term in (41) depends on the underlying density being estimated via the term involving $|\gamma_j|^2$. Since this dependence arises from the square bias term in (35), we can obtain the alternative estimate

$$E(|\gamma_{N,j} - \gamma_{N,j}^{[1]}|^2) = \frac{1}{720N}(\delta s_j)^4 + O((\delta s_j)^6), with\ \beta_j^{[1]} = 1 + \frac{1}{12}(\delta s_j)^2. \tag{43}$$

With these results in hand, we now note from (19) that

$$\mathcal{M}^{(r)} \approx \frac{(2\pi)^{\frac{1}{2}}}{M\delta} \sum_{j=-M/2+1}^{M/2} s_j^{2r} \widetilde{K}(hs_j)^2 E(|\gamma_N(s_j) - \gamma_N^*(s_j)|^2), \tag{44}$$

where again all the modes upto the Nyquist frequency on the grid have been retained although $\widetilde{K}(hs_j)$ will be heavily damped for the higher-order modes. The evaluation of the summation above can be carried in the case $\delta s_j \ll 1$, if we write

$$E(|\gamma_N(s_j) - \gamma_N^*(s_j)|^2) = \lambda(\delta s_j)^{2n} + O((\delta s_j)^{2n+1}). \tag{45}$$

Here the leading-order coefficient λ which depends on the discretization method, is assumed to be independent of underlying characteristic function $\gamma(s)$. From (35) and (45), we note that the dependence of λ on $\gamma(s)$ can only arise through the square bias terms, and it is therefore important to choose the value of β_j via (33) to at least cancel the bias in the leading order terms for small δs_j. The required values of λ and n for the various discretization procedures are given above in (40), (42) and (43). Thus finally, we have

$$\begin{aligned} \mathcal{M}^{(r)} &\approx \lambda\delta^{2n} \left\{ \frac{(2\pi)^{\frac{1}{2}}}{M\delta} \sum_{j=-M/2+1}^{M/2} s_j^{2(n+r)} \widetilde{K}(hs_j)^2 \right\} \\ &\approx \lambda\delta^{2n} \int_{-\infty}^{\infty} s^{2(n+r)} \widetilde{K}(hs)^2 ds = \lambda\delta^{2n} \mathcal{A}_{n+r}. \end{aligned} \tag{46}$$

Here, the integrals $\mathcal{A}_j$ depend only on the kernel function, and in the particular case of the Gaussian kernel are given by the recurrence

$$\mathcal{A}_0 = \frac{1}{2\sqrt{\pi}h}, \quad \mathcal{A}_j = -\frac{1}{2h}\frac{d\mathcal{A}_{j-1}}{dh}, \text{ for } j = 1, 2, \ldots, \tag{47}$$

whose solution is

$$\mathcal{A}_j = \frac{1.3\ldots(2j-1)}{2^{j+1}\sqrt{\pi}h^{2j+1}}. \tag{48}$$

We therefore obtain the main theoretical result of this section,

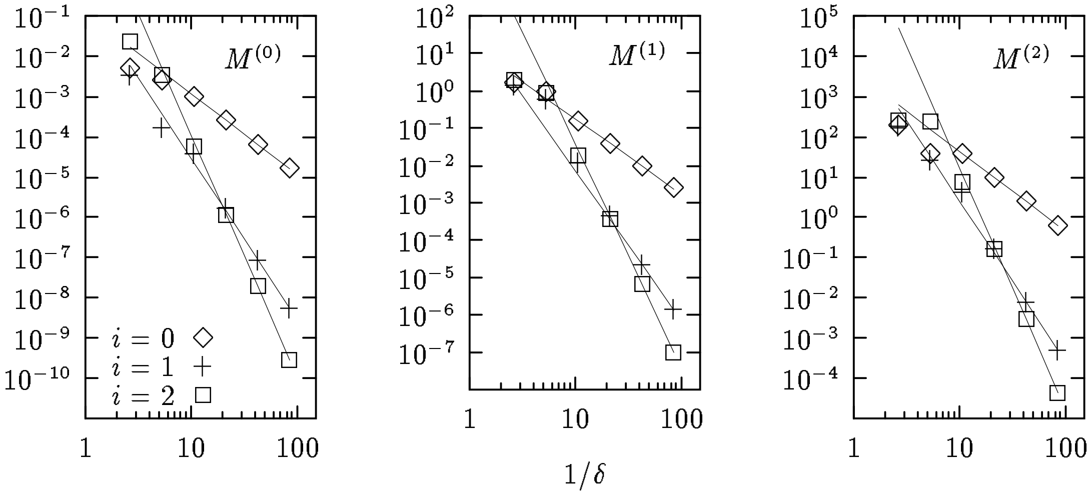

Figure 3: I.M.S.D.E. convergence for input data sampled from a standard gaussian, for piecewise constant ($i = 0$), piecewise linear ($i = 1$) and piecewise quadratic ($i = 2$) interpolation, with points (simulation results) and solid lines (theory).

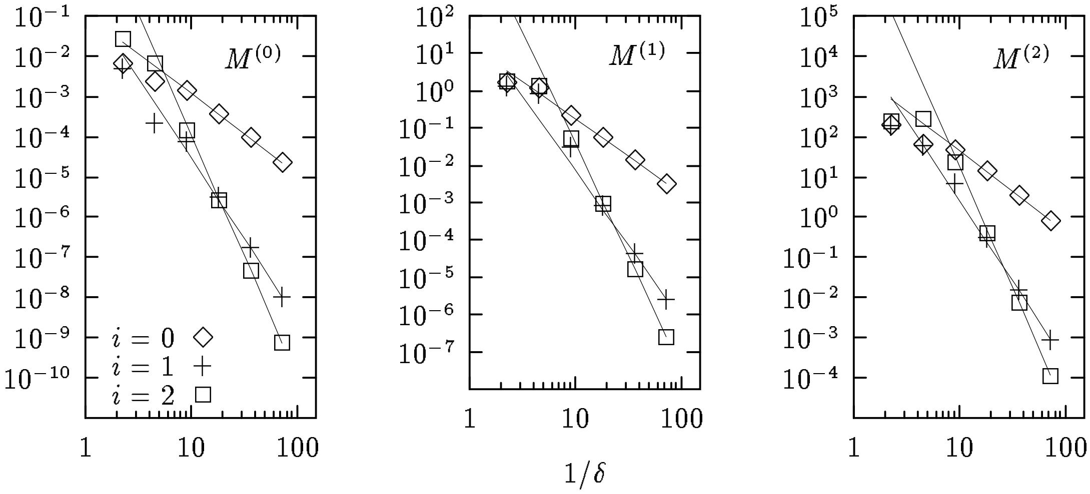

Figure 4: As in Figure (3), but for input data sampled from a 50-50 standard gaussian mixture centered at -1 and $+1$ respectively.

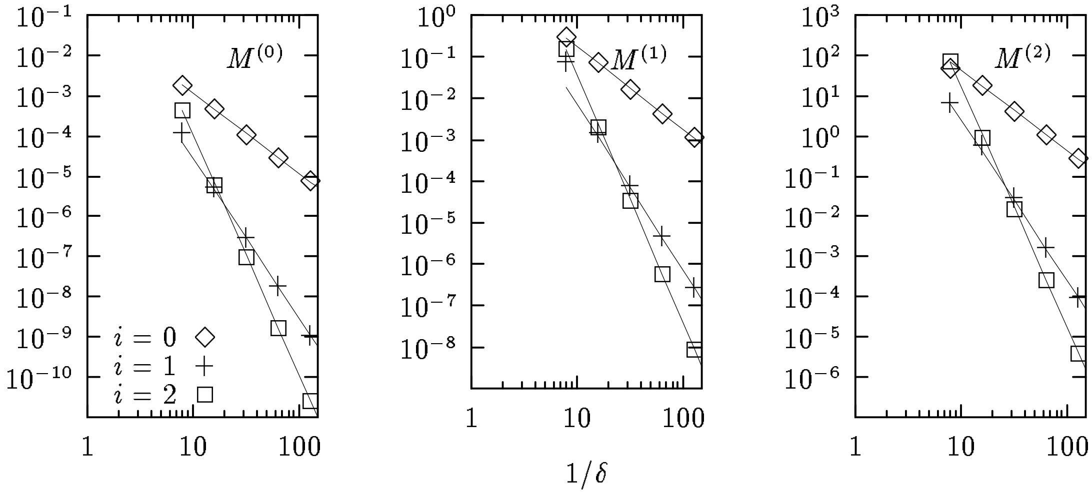

Figure 5: As in Figure (3), but on the interval $(-2, +2)$ for data sampled from a uniform distribution on the interval $(-1, +1)$.

Lemma 3.5 *The Integrated Mean Square Discretization Error (I.M.S.D.E.) of the approximations to the sample density estimates and its derivatives (as defined in (19)) is given asymptotically for $\delta \longrightarrow 0$ by*

1. *Piecewise constant:* $\mathcal{M}^{(r)} \approx \frac{1}{12N}\delta^2 \mathcal{A}_{1+r}$.
2. *Piecewise linear:* $\mathcal{M}^{(r)} \approx \frac{1}{720N}\delta^4 \mathcal{A}_{2+r}$, *when the bias correction factor is used as in Lemma 3.4.*
3. *Piecewise quadratic:* $\mathcal{M}^{(r)} \approx \frac{2}{945N}\delta^6 \mathcal{A}_{3+r}$.

The theoretical results are in excellent agreement with simulations that have been performed as described below. Figures (3)–(5) shows the convergence with decreasing δ when the density estimation is performed using random samples drawn from three different distributions, viz., a standard gaussian , a 50-50 gaussian mixture, and a uniform distribution. In all these simulations the kernel bandwidth was fixed at $h = 0.1$, and each input data realization consisted of $N = 100$ randomly-sampled points (with the results averaged over 125 such realizations). These simulation results are also plotted in a different way in Figure (6) to show the excellent agreement with the theoretical predictions for the constant factor in the asymptotic convergence rate.

The results in Figure (5) show the good agreement between simulations and the asymptotic results even in the case where the underlying density is not a C^∞

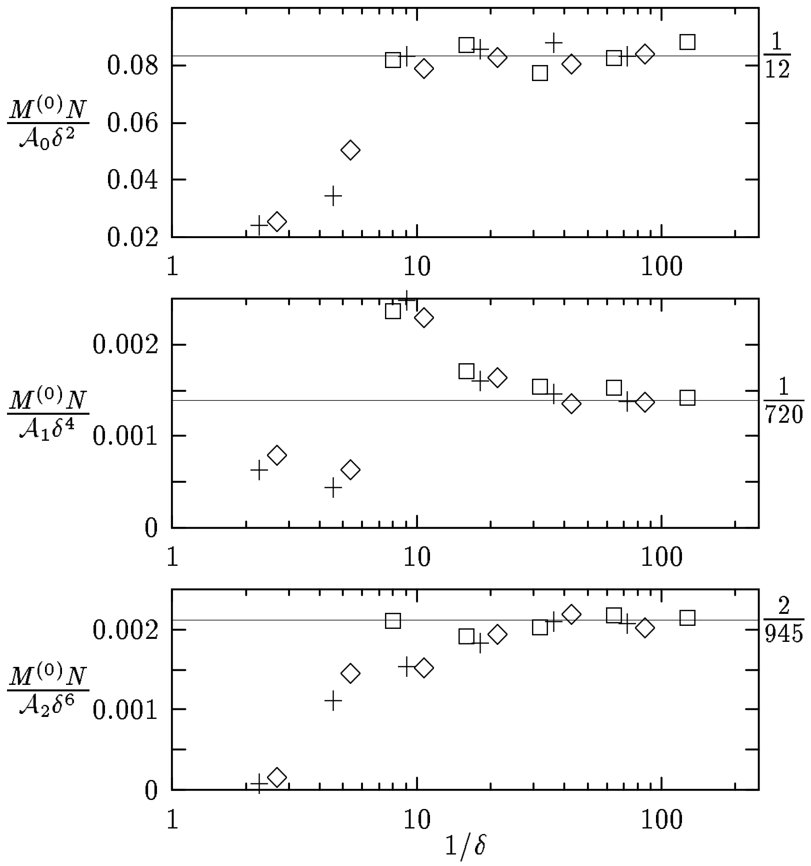

Figure 6: Comparison of theory and simulation results (same as Figures (3)-(5)) for the constant factor in the $\delta \longrightarrow 0$ asymptotic convergence rate. The legends are standard gaussian $\diamond$, gaussian mixture $+$, and uniform distribution $\square$.

function, as assumed in the theoretical analysis.

4 Application to Projection Pursuit.

In this section, we describe an application where the use of a higher-order finite element interpolation in an FFT-based algorithm for kernel density estimation is quite essential.

Projection pursuit is a multivariate data analysis technique that seeks to identify interesting structures such as clustering and multimodality from the examination of suitable low-dimensional linear projections of the data (see Friedman [2], Huber [5], Jones and Sibson [8]). Specifically, let $\{\boldsymbol{U}_i\}_{i=1}^N$ denote *i.i.d* random variables sampled from a multivariate distribution in $\mathbb{R}^P$ (we assume that a preliminary whitening transformation has been used to standardize the sample mean to zero and the sample covariance matrix to the identity, see Fukunaga [3]). In one class of projection pursuit algorithms, the objective is to find projections $\boldsymbol{\alpha} \in \mathbb{R}^P$ in which the data $\{\boldsymbol{U}_i\}_{i=1}^N$ reveals maximum departure from a gaussian behavior. Jones and Sibson [8], suggest that a suitable projection vector $\boldsymbol{\alpha}$ must minimize an entropy index computed from the corresponding marginal kernel density estimate for the data, which is

$$g_N(x, \boldsymbol{\alpha}) = \frac{1}{Nh} \sum_{i=1}^{N} K\left(\frac{x - X_i}{h}\right), \quad \text{with } X_i = \frac{\boldsymbol{\alpha}^T \boldsymbol{U}_i}{(\boldsymbol{\alpha}^T \boldsymbol{\alpha})^{1/2}}, \tag{49}$$

where, X_i is the projection of $\boldsymbol{U}_i$ in the direction $\boldsymbol{\alpha}$. The corresponding marginal entropy is then defined by

$$S(\boldsymbol{\alpha}) = -\int_{-\infty}^{\infty} g_N(x, \boldsymbol{\alpha}) \log g_N(x, \boldsymbol{\alpha}) dx. \tag{50}$$

This entropy measure takes its maximum value

$$S_{max} = \frac{1}{2} \log(2\pi e) \tag{51}$$

when the corresponding marginal density $g_N(x, \boldsymbol{\alpha})$ is the standard univariate gaussian. Thus $S(\boldsymbol{\alpha}) - S_{max} \leq 0$ is a measure of the departure of the marginal density from the standard gaussian, and therefore the directions $\boldsymbol{\alpha}$ where this quantity has a local minima are of particular interest in projection pursuit. The required local minina solutions can be computed by a gradient-optimization procedure (we have used nonlinear conjugate gradient and Quasi-Newton methods). If α_l denotes the

$l'th$ component of $\boldsymbol{\alpha}$, and $f_{N,l}(x,\boldsymbol{\alpha}) = \partial g_N(x,\boldsymbol{\alpha})/\partial\alpha_l$, then the required analytical gradients of the objective function $S(\boldsymbol{\alpha})$ are given by

$$\frac{\partial S}{\partial \alpha_l} = -\int_{-\infty}^{\infty} (1 + \log g_N(x,\boldsymbol{\alpha})) f_{N,l}(x,\boldsymbol{\alpha}) dx, \tag{52}$$

The integrals in (50) and (52) can be evaluated using the trapezoidal rule from the values of the integrand on a fixed grid. These values can be computed by the FFT-based algorithm, after noting that the Fourier transforms of $g_N(x,\boldsymbol{\alpha})$ and $f_{N,l}(x,\boldsymbol{\alpha})$ are given by

$$\widetilde{g_N}(s,\boldsymbol{\alpha}) = \widetilde{K}(hs)\sum_{i=1}^{N}\frac{e^{isX_i}}{N}, \quad \widetilde{f_{N,l}}(s,\boldsymbol{\alpha}) = is\widetilde{K}(hs)\sum_{i=1}^{N} Z_{i,l}\frac{e^{isX_i}}{N}, \tag{53}$$

where

$$Z_{i,l} = \frac{U_{i,l}}{\sqrt{\boldsymbol{\alpha}^T\boldsymbol{\alpha}}} - \frac{\alpha_l X_i}{\boldsymbol{\alpha}^T\boldsymbol{\alpha}}. \tag{54}$$

The details of the FFT-based algorithm are similar to the description in Section (2), requiring the interpolation of the marginal empirical characteristic function and its partial derivatives in (53)-(53) above. The use of lower-order finite element basis functions for this interpolation, leads to results that are inaccurate (compared to the direct summation of the individual terms in (53)-(53)). More seriously, it can also lead to inconsistencies between the function and gradient calculations, which in turn leads to convergence difficulties in the optimization procedure.

To illustrate this point, we consider the Taylor expansion of $S_(\boldsymbol{\alpha})$ about some $\boldsymbol{\alpha}_c$. Let $\delta\boldsymbol{\alpha} \in \mathbb{R}^P$ denote a random perturbation vector and let $\boldsymbol{g}(\boldsymbol{\alpha}) = \nabla_{\boldsymbol{\alpha}} S(\boldsymbol{\alpha}) \in \mathbb{R}^P$ denote the gradient vector. Then, define the following truncated Taylor expansion

$$t_\epsilon = S(\boldsymbol{\alpha}_c + \epsilon\delta\boldsymbol{\alpha}) - S(\boldsymbol{\alpha}_c) - \epsilon\boldsymbol{g}(\boldsymbol{\alpha}_c)^T\delta\boldsymbol{\alpha} = O(\epsilon^2), \tag{55}$$

where the function values $S(\boldsymbol{\alpha}_c + \epsilon\delta\boldsymbol{\alpha}), S(\boldsymbol{\alpha}_c)$ and the gradients $\boldsymbol{g}(\boldsymbol{\alpha}_c)$ are all evaluated from the marginal density estimates computed using either exact summation or an FFT-based algorithm. In order to check the accuracy and consistency of gradient evaluations, Averbukh, Figueroa and Schlick [1] have suggested looking at the ratio $r_\epsilon = t_{2\epsilon}/t_\epsilon$ as ϵ is decreased, since the theoretical value of $\lim_{\epsilon\longrightarrow 0} r_\epsilon = 4$ if the function and gradient evaluations are consistent. In practice, the evaluation of t_ϵ for very small ϵ is affected by arithmetic round-off error, so that reliable estimates for $\lim_{\epsilon\longrightarrow 0} r_\epsilon$ are obtained from a series of values for r_ϵ with decreasing ϵ in the the range when ϵ is neither too large or small.

ϵ	FFT-based Algorithm			Direct Sum
	$i = 0$	$i = 1$	$i = 2$	
$.5000E+0$	2.032	2.411	2.409	2.410
$.2500E+0$	2.111	2.995	3.015	3.016
$.1250E+0$	4.274	3.697	3.549	3.553
$.6250E-1$	6.879	3.059	3.795	3.819
$.3125E-1$	2.650	3.310	3.830	3.926
$.1562E-1$	–1.029	2.428	3.883	3.968
$.7812E-2$	1.314	2.548	3.889	3.986
$.3906E-2$	8.689	2.381	3.839	3.993
$.1953E-2$	1.840	2.743	3.715	3.997
$.9766E-3$	9.900	2.559	3.515	3.998
$.4883E-3$	–200.5	2.615	3.201	3.999
$.2441E-3$	2.000	2.364	2.859	4.000
$.1221E-3$	2.000	2.200	2.547	4.000

Table 1: Table of values for r_ϵ used to determine the consistency of function and gradient evaluations. The FFT-based kernel density estimation results include interpolation by piecewise constant ($i = 0$), linear ($i = 1$) and quadratic ($i = 2$) basis functions.

We have carried out this test using the FLEA data set (Lubischew [9]) for which $P = 6, N = 74$. The values of $\boldsymbol{\alpha}_c$ and $\delta\boldsymbol{\alpha}$ were chosen randomly, and h was taken to be 0.45 (a value that was intentionally chosen to oversmooth the data, which is bimodal or trimodal in most projections). The quadratures were computed with $M = 256$ points, which is more than is necessary to obtain a converged result. Table (1) shows the plot of r_ϵ versus ϵ, and the exact summation for the kernel density estimation yields fully consistent results, while the FFT-based algorithm requires at least the piecewise-quadratic interpolation to obtain a reasonably consistent function and gradient evaluation. Practical experience shows that this lack of consistency results in poor convergence and stagnation in the projection pursuit optimization procedure.

5 Summary Remarks

The accuracy and convergence of FFT-based algorithms for kernel density estimation can be improved by using higher-order finite-element interpolation procedures. This paper has considered the case of piecewise constant, piecewise linear and piecewise quadratic finite element approximations to the empirical characteristic function. Convergence results have been derived for the Integrated Mean Square Discretization Error ($I.M.S.D.E.$) resulting from these approximations in the FFT-based algorithm. The methods used in the proof, which are generalizations of the approach of Jones and Lotwick [6], can be easily extended to other finite element approximations (such as piecewise cubic or piecewise Hermite cubic), although a symbolic manipulation package might be required for the detailed algebraic derivation. The extension of the analysis to multivariate density estimation in high dimensions is also of interest, as is the performance of the FFT-based algorithm relative to the Greengard-Strain algorithm ([4]) in the multivariate case.

References

[1] V. Averbukh, S. Figueroa and T. Schlick, *ALGORITHM 566: HESFCN - A FORTRAN Package of Hessian Subroutines for Testing Nonlinear Optimization Software*, ACM Transactions on Mathematical Software, 20 (1994), pp. 282-285.

[2] J. H. Friedman, *Exploratory projection pursuit*, J. Amer. Stat. Assoc., 820 (1987), pp. 249-266.

[3] K. Fukunaga, *Introduction to Statistical Pattern Recognition*, Second Edition, Academic Press (1990).

[4] L. Greengard and J. Strain, *The Fast Gauss Transform*, SIAM J. of Stat. Comput., 12 (1991), pp. 79-94.

[5] P. J. Huber, *Projection Pursuit (With Discussion)*, Annals of Statistics, 9 (1985), pp. 435-475.

[6] M. C. Jones and H. W. Lotwick, *On the errors involved in computing the empirical characteristic function*, J. Statist. Comput. Simul., 17 (1983), pp. 133-149.

[7] M. C. Jones and H. W. Lotwick, *Remark AS R50: A Remark on Algorithm AS 176: Kernel density estimation using the Fast Fourier Transform*, Applied Statistics, 33 (1984), pp. 120-122.

[8] M. C. Jones and R. Sibson, *What is projection pursuit?*, J. Royal Stat. Soc., 150 (1987), pp. 1-37.

[9] A. A. Lubischew, *On the Use of Discriminant Functions in Taxonomy*, Biometrics, 18 (1962) pp. 455-477.

[10] B. W. Silverman, *Algorithm AS 176: Kernel density estimation using the Fast Fourier Transform*, Applied Statistics, 31 (1982), pp. 93-99.

[11] B. W. Silverman, *Density Estimation for Statistics and Data Analysis*, Chapman and Hall, London (1986).