ELSEVIER Computational Statistics & Data Analysis 22 (1996) 1–16

COMPUTATIONAL STATISTICS & DATA ANALYSIS

Accuracy of binned kernel functional approximations[1]

W. González-Manteiga[a], C. Sánchez-Sellero[a], M.P. Wand[b,*]

[a]*Departamento de Estadística e Investigación Operativa, Facultad de Matemáticas, Universidad de Santiago de Compostela, 15771 Santiago de Compostela, Spain*

[b]*Australian Graduate School of Management, University of New South Wales, Sydney, NSW 2052, Australia*

Received 1 March 1995; revised 1 July 1995

Abstract

Virtually all common bandwidth selection algorithms are based on a certain type of kernel functional estimator. Such estimators can be computationally very expensive, so in practice they are often replaced by fast binned approximations. This is especially worthwhile when the bandwidth selection method involves iteration. Results for the accuracy of these approximations are derived and then used to provide an understanding of the number of binning grid points required to achieve a given level of accuracy. Our results apply to both univariate and multivariate settings. Multivariate contexts are of particular interest since the cost due to having a higher number of grid points can be quite significant.

Keywords: Bandwidth selection; Binned kernel estimator; Density estimation; Kernel estimator

1. Introduction

1.1. Motivation

In many circumstances the application of non-parametric smoothing techniques is greatly enhanced by the availability of a good automatic procedure for choosing the smoothing parameters. An illustration of smoothing parameter selection is depicted in Fig. 1. Fig. 1(a) is a bivariate kernel density estimate based on $n = 640$ data points representing earthquake activity in the Mount St Helens region of Washington state,

* Corresponding author.

[1] Partially supported by the grants PTR91-0058, PB91-0794 and XUGA20703B93.

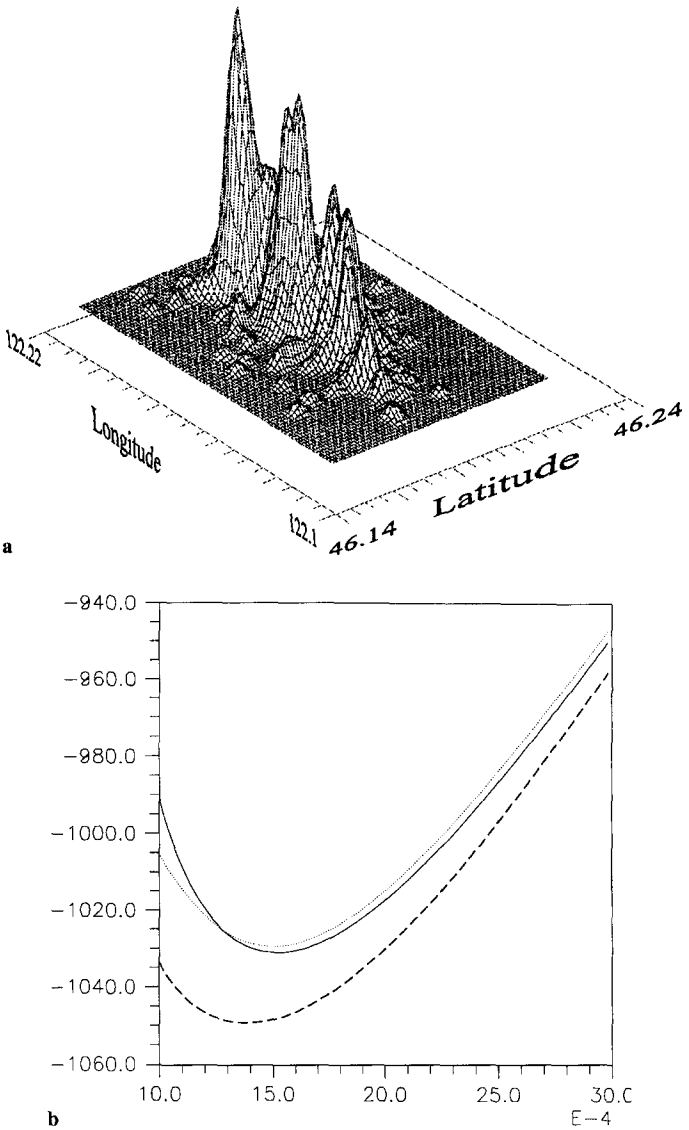


Fig. 1. (a) Bivariate kernel density estimate based on $n = 640$ data points representing earthquake activity in the Mount St Helens region of Washington state, U.S.A. (b) The least-squares cross-validation function and two binning approximations to it. The solid curve is the exact function, the dashed curve is its binned approximation with $M=75$ and the dotted curve is the binned approximation with $M=150$.

USA. These data are analysed by O'Sullivan and Pawitan (1993). The kernel is the bivariate normal density using a single bandwidth for both coordinate directions. Silverman (1986) provides a good introduction to kernel density estimation. Fig. 1(b) shows how the bandwidth was chosen. The solid curve is the least-squares cross-validation criterion function (Rudemo, 1982; Bowman, 1984) for these data. The chosen bandwidth is the value that minimises this criterion function, in this case $\hat{h} = 0.0015$.

The computational labour required to produce the solid curve is considerably heavy – involving $O(n^2)$ evaluations of a kernel function for each value of the bandwidth. Two fast-to-compute approximations of the criterion function are shown by the broken curves. Each is based on the simple binning strategy, described in Section 2, on an $M \times M$ mesh over the smallest rectangle containing the data. The dashed curve corresponds to $M = 75$ while the dotted curve is for $M = 150$. For $M = 75$ we see that there is quite a substantial difference between the exact curve and its approximation, particularly in terms of the bandwidth that minimises each. Even with $M = 150$ there is some noticeable approximation error, although not enough to affect the results markedly. The main point is that there is a definite loss in accuracy due to the use of binning.

The aim of this paper is to study the accuracy of binning approximations used in bandwidth selection algorithms. Since virtually all common rules depend on the computation of a particular type of kernel functional estimator, the problem reduces to the study of the accuracy of binned kernel functional approximations. For simplicity and brevity our study is confined to the density estimation context. However, the conclusions apply to other settings where bandwidth selection is used, such as kernel regression.

Binning techniques for fast kernel estimation were first proposed by Silverman (1982), Scott (1985) and Härdle and Scott (1992). Wand (1994) describes the extension of binning ideas to multivariate functional estimation. Studies in the approxi mation accuracy of binned kernel estimators include those of Jones and Lotwick (1983), Scott and Sheather (1985) and Hall and Wand (1995). The class of kernel functional estimators studied here was introduced by Hall and Marron (1987) and Jones and Sheather (1991). For access to the large literature on automatic bandwidth selection methods and their relative merits, see, for example, Cao et al. (1994) and Jones et al. (1995).

Section 2 contains the theoretical results required for our investigation. In Section 3 we apply the results to a set of specific problems to develop an understanding of the effect of binning on kernel functional estimation and, therefore, the effect on bandwidth selection algorithms. Conclusions of this study are given in Section 4.

1.2. Notation

Elements of a d-vector a will be denoted by $(a_1, \ldots, a_d)$ and the sum of the entries of a will be denoted by $|a|$. For a scalar p we define $a^p = (a_1^p, \ldots, a_d^p)$. We will use e_i to denote the d-vector having 1 in the ith entry and 0 elsewhere. For a d-variate function u and d-vector r, partial derivatives of u of order r will be denoted by

$$u^{(r)}(x) = \frac{\partial^{|r|}}{\partial x_1^{r_1} \cdots \partial x_d^{r_d}} u(x).$$

Lastly, $\int u(x)\,dx$ will be taken to mean integration of u over the whole of d-dimensional Euclidean space.

2. Main results

2.1. General densities

Suppose that we observe a d-variate random sample $X_1,\ldots,X_n$ having common density f. An important class of kernel functional estimators is that having generic member

$$\hat{\theta} = n^{-2}\sum_{i=1}^{n}\sum_{i'=1}^{n} L(X_i - X_{i'}) \tag{2.1}$$

for some d-variate functional L. For simplicity we will assume that $L(-x) = L(x)$ for all d-vectors x. For example, plug-in bandwidth selection rules use $L = K_h^{(r)}$ where $|r|$ is some even number. Here $K_h(t) = K(t_1/h_1,\ldots,t_d/h_d)/(h_1\cdots h_d)$ for some d-variate kernel K and bandwidth vector $h = (h_1,\ldots,h_d)$. Estimators of type (2.1) are the main components of virtually all common bandwidth selection procedures including least-squares cross-validation, biased cross-validation (Scott and Terrell, 1987; Sain et al., 1994), plug-in rules (Park and Marron, 1990; Sheather and Jones, 1991; Wand and Jones, 1994) and bootstrap or smoothed cross-validation (Taylor, 1989; Hall et al., 1992; Cao et al., 1994).

In practice, for reasons of computational speed, it is common to replace $\hat{\theta}$ by a binned approximation. Let $\mathbb{Z}$ denote the set of integers and δ be a d-vector having all entries positive. Then

$$\{g_\ell = (\delta_1\ell_1,\ldots,\delta_d\ell_d) : \ell \in \mathbb{Z}^d\}$$

is a mesh of *grid points* in d-dimensional space, equally spaced in each direction. We will refer to δ_i as the *binwidth* for direction i. The *simple binning* rule assigns *grid counts* c_ℓ to each of the g_ℓ according to

$$c_\ell = \sum_{i=1}^{n}\prod_{j=1}^{d} I(-\tfrac{1}{2} < \delta_j^{-1}X_{ij} - \ell_j \le \tfrac{1}{2}),$$

where $I(A)$ is the indicator of the event A. An alternative binning procedure is *linear binning* (Jones and Lotwick, 1983) for which

$$c_\ell = \sum_{i=1}^{n}\prod_{j=1}^{d}(1 - |\delta_j^{-1}X_{ij} - \ell_j|)_+,$$

where $x_+ = \max(0,x)$. Thus, simple binning involves each observation being moved to its nearest grid point, while linear binning involves each observation being "linearly split up" among its neighbouring grid points. These notions are graphically illustrated by Figure 1 of Wand (1994).

In each case, the binned approximation of $\hat{\theta}$ is

$$\tilde{\theta}_\delta = n^{-2}\sum_{\ell\in\mathbb{Z}^d}\sum_{\ell'\in\mathbb{Z}^d} L(g_\ell - g_{\ell'})c_\ell c_{\ell'}.$$

The main advantage of $\tilde{\theta}_\delta$ is that the number of distinct arguments of the function L is relatively small and, combined with the fact that a finite number of the c_ℓ are non-zero, this can represent enormous savings compared to the $O(n^2)$ evaluations of L required for computation of $\hat{\theta}$; see, for example, Wand (1994).

Our interest here centres on the accuracy of the approximation of $\hat{\theta}$ by $\tilde{\theta}_\delta$. For this we appeal to the following theorem, the proof of which is given in Appendix A.

Theorem 1. *Suppose that the third-order partial derivatives of L are each continuous and bounded. Let $\delta_i \to 0$ for $i = 1,\ldots,d$, such that $\delta_i/\delta_j \to C_{ij}$ where $0 < C_{ij} < \infty$. If $\tilde{\theta}_\delta$ is based on simple binning then*

$$E\{(\hat{\theta}-\tilde{\theta}_\delta)^2\} = \sum_{i=1}^{d} \frac{\delta_i^2(n-1)}{3n^3}\left[E\{L^{(e_i)}(X_1-X_2)^2\}\right.$$
$$\left.+(n-2)E\{L^{(e_i)}(X_1-X_2)L^{(e_i)}(X_1-X_3)\}\right] + o\left(\sum_{i=1}^{d}\sum_{j=1}^{d}\delta_i\delta_j\right).$$

If $\tilde{\theta}_\delta$ is based on linear binning then

$$E\{(\hat{\theta}-\tilde{\theta}_\delta)^2\} = \sum_{i=1}^{d} \frac{\delta_i^4}{180n^3}\Big((1+5n)L^{(2e_i)}(0)^2$$
$$+(2+10n)(n-1)L^{(2e_i)}(0)E\{L^{(2e_i)}(X_1-X_2)\}$$
$$+5(n-1)(n-2)(n-3)[E\{L^{(2e_i)}(X_1-X_2)\}]^2 + 11(n-1)E\{L^{(2e_i)}(X_1-X_2)^2\}$$
$$+21(n-1)(n-2)E\{L^{(2e_i)}(X_1-X_2)L^{(2e_i)}(X_1-X_3)\}\Big)$$
$$+\sum\sum_{i\neq j}\frac{\delta_i^2\delta_j^2}{36n^3}\Big(nL^{(2e_i)}(0)L^{(2e_j)}(0) + 2n(n-1)L^{(2e_i)}(0)E\{L^{(2e_j)}(X_1-X_2)\}$$
$$+(n-1)(n-2)(n-3)E\{L^{(2e_i)}(X_1-X_2)\}E\{L^{(2e_j)}(X_1-X_2)\}$$
$$+2(n-1)E\{L^{(2e_i)}(X_1-X_2)L^{(2e_j)}(X_1-X_2)\}$$
$$+4(n-1)(n-2)E\{L^{(2e_i)}(X_1-X_2)L^{(2e_j)}(X_1-X_3)\}\Big) + o\left(\sum_{i=1}^{d}\sum_{j=1}^{d}\delta_i^2\delta_j^2\right).$$

The most noticeable feature of Theorem 1 is that, for linear binning, the mean squared distance between $\tilde{\theta}_\delta$ and $\hat{\theta}$ has terms proportional to $\delta_i^2\delta_j^2$, while the rate is only $\delta_i\delta_j$ for simple binning. Therefore, linear binning has a distinct advantage over simple binning in terms of asymptotic approximation error. For finite grids the dominance of linear binning is not as clear since the constant multiples are not comparable for general L and f. To develop a better understanding of the relative and absolute performances of each binning strategy in practice it is necessary to

study individual examples, and this will be done in Section 3. For this it is desirable to have a set of examples for which the constant multiples are relatively simple to calculate. This is provided by the following subsection.

2.2. Normal mixture densities

The difficult-to-calculate quantities in the expressions of Theorem 1 are of the form

$$A_r \equiv E\{L^{(r)}(X_1 - X_2)\},$$

$$B_{rr'} \equiv E\{L^{(r)}(X_1 - X_2)L^{(r')}(X_1 - X_2)\},$$

$$C_{rr'} \equiv E\{L^{(r)}(X_1 - X_2)L^{(r')}(X_1 - X_3)\}.$$

However, it is possible to obtain exact closed form expressions for A_r, $B_{rr'}$ and $C_{rr'}$ if L is based on the normal kernel and f is a *normal mixture* density.

For d-vectors μ and σ, where σ has all entries positive, let

$$\phi_\sigma(x - \mu) = (2\pi)^{-d/2} \prod_{i=1}^{d} \exp\{-(x_i - \mu_i)^2/(2\sigma_i^2)\}/\sigma_i.$$

Note that $\phi_\sigma(\cdot - \mu)$ is simply the density of the d-variate normal distribution with mean vector μ and covariance matrix $\mathrm{diag}(\sigma^2)$. Let h and h' be vectors of bandwidths. The data-dependent component of most of the common bandwidth selection criteria can be written in terms of expressions of the form given by (2.1) where L is a linear combination of versions of ϕ_σ. Examples are

$$L = \begin{cases} \phi_{h2^{1/2}} - 2\left(\dfrac{n}{n-1}\right)\phi_h & \text{for least squares cross-validation,} \\ \phi_h^{(r)},\ |r| \text{ even} & \text{for biased cross-validation and plug-in} \\ \phi_{(2h^2+2h'^2)^{1/2}} - 2\phi_{(h^2+2h'^2)^{1/2}} + \phi_{h'2^{1/2}} & \text{for smoothed cross-validation.} \end{cases}$$

Therefore, in each of these cases, the constants A_r, $B_{rr'}$ and $C_{rr'}$ depend on linear combinations of

$$\mathscr{A}(r,\lambda) = E\{\phi_\lambda^{(r)}(X_1 - X_2)\},$$

$$\mathscr{B}(r,r',\lambda,\lambda') = E\{\phi_\lambda^{(r)}(X_1 - X_2)\phi_{\lambda'}^{(r')}(X_1 - X_2)\},$$

$$\mathscr{C}(r,r',\lambda,\lambda') = E\{\phi_\lambda^{(r)}(X_1 - X_2)\phi_{\lambda'}^{(r')}(X_1 - X_3)\}.$$

The class of normal mixture densities that we consider is those having the form

$$f(x) = \sum_{\ell=1}^{k} w_\ell \phi_{\sigma_\ell}(x - \mu_\ell), \tag{2.2}$$

where, for each $\ell = 1,\ldots,k$, μ_ℓ is an arbitrary d-vector, σ_ℓ is a d-vector with all entries positive, $w_\ell > 0$ and $\sum_{\ell=1}^{k} w_\ell = 1$. Note that this class is not as general as

that considered by, for example, Wand and Jones (1993), where the components had full covariance matrices. However, the current class still provides an interesting set of examples, without complicating the calculations.

If f is of the form (2.2) then it can be shown that

$$\mathscr{A}(r,\lambda)=\sum_{\ell_1=1}^{k}\sum_{\ell_2=1}^{k}w_{\ell_1}w_{\ell_2}\prod_{i=1}^{d}\alpha(r_i,\lambda_i,\mu_{\ell_1 i},\mu_{\ell_2 i},\sigma_{\ell_1 i},\sigma_{\ell_2 i}),$$

$$\mathscr{B}(r,r',\lambda,\lambda')=\sum_{\ell_1=1}^{k}\sum_{\ell_2=1}^{k}w_{\ell_1}w_{\ell_2}\prod_{i=1}^{d}\beta(r_i,r_i',\lambda_i,\lambda_i',\mu_{\ell_1 i},\mu_{\ell_2 i},\sigma_{\ell_1 i},\sigma_{\ell_2 i})$$

and

$$\mathscr{C}(r,r',\lambda,\lambda')=\sum_{\ell_1=1}^{k}\sum_{\ell_2=1}^{k}\sum_{\ell_3=1}^{k}w_{\ell_1}w_{\ell_2}w_{\ell_3}\prod_{i=1}^{d}\gamma(r_i,r_i',\lambda_i,\lambda_i',\mu_{\ell_1 i},\mu_{\ell_2 i},\mu_{\ell_3 i},\sigma_{\ell_1 i},\sigma_{\ell_2 i},\sigma_{\ell_3 i}),$$

where, for scalars r, λ, λ', μ_i, σ_i, $i=1,2,3$,

$$\alpha(r,\lambda,\mu_1,\mu_2,\sigma_1,\sigma_2)$$
$$=(-1)^r\phi_{(\lambda^2+\sigma_1^2+\sigma_2^2)^{1/2}}(\mu_1-\mu_2)H_r\left\{\frac{\mu_1-\mu_2}{(\lambda^2+\sigma_1^2+\sigma_2^2)^{1/2}}\right\}(\lambda^2+\sigma_1^2+\sigma_2^2)^{-r/2},$$

$$\beta(r,r',\lambda,\lambda',\mu_1,\mu_2,\sigma_1,\sigma_2)=(2\pi)^{-1/2}\lambda^{-r-1}\lambda'^{-r'-1}\tilde{\lambda}\phi_{(\sigma_1^2+\sigma_2^2+\tilde{\lambda}^2)^{1/2}}(\mu_1-\mu_2)$$
$$\times\sum_{j=0}^{r}\sum_{j'=0}^{r'}\sum_{k=0}^{r-j}\sum_{k'=0}^{r'-j'}\binom{r}{j}\binom{r'}{j'}\binom{r-j}{k}\binom{r'-j'}{k'}$$
$$\times v(j+j')v(k+k')\lambda^{-j-k}\lambda'^{-j'-k'}\sigma_1^{j+j'}\sigma_2^{k+k'}\tilde{\lambda}^{j+j'+2k+2k'}$$
$$\times(\sigma_1^2+\tilde{\lambda}^2)^{-(j+j'+k+k')/2}H_{r-j-k}\left\{\frac{(\mu_1-\mu_2)\tilde{\lambda}^2}{\lambda(\sigma_1^2+\sigma_2^2+\tilde{\lambda}^2)}\right\}$$
$$\times(\sigma_1^2+\sigma_2^2+\tilde{\lambda}^2)^{-(k+k')/2}H_{r'-j'-k'}\left\{\frac{(\mu_1-\mu_2)\tilde{\lambda}^2}{\lambda'(\sigma_1^2+\sigma_2^2+\tilde{\lambda}^2)}\right\}$$

and

$$\gamma(r,r',\lambda,\lambda',\mu_1,\mu_2,\mu_3,\sigma_1,\sigma_2,\sigma_3)$$
$$=(2\pi)^{-1/2}\phi_{\tilde{\sigma}}(\tilde{\mu})\sigma_1^{-1}\sum_{j=0}^{r}\sum_{j'=0}^{r'}\binom{r}{j}\binom{r'}{j'}v(j+j')\times H_{r-j}\left\{\frac{\mu_2-\tilde{\tilde{\mu}}}{(\lambda^2+\sigma_2^2)^{1/2}}\right\}$$
$$\times H_{r'-j'}\left\{\frac{\mu_3-\tilde{\tilde{\mu}}}{(\lambda'^2+\sigma_3^2)^{1/2}}\right\}\tilde{\sigma}^{-j-j'}\{(\lambda^2+\sigma_2^2)^{r+j+1}(\lambda'^2+\sigma_3^2)^{r'+j'+1}\}^{-1/2}.$$

Here H_j denotes the jth normalised Hermite polynomial, given by $H_j(t) = (-1)^j \phi_1^{(j)}(t)/\phi_1(t)$, $\nu(j) = (j-1)(j-3)\cdots 1$ for j even and positive, $\nu(0) = 1$, $\nu(j) = 0$ for j odd, $\tilde{\lambda} = \lambda\lambda'/(\lambda^2 + \lambda'^2)^{1/2}$,

$$\tilde{\mu} = \left\{ \frac{(\mu_2 - \mu_3)^2}{(\lambda^2 + \sigma_2^2)(\lambda'^2 + \sigma_3^2)} + \frac{(\mu_1 - \mu_2)^2}{\sigma_1^2(\lambda^2 + \sigma_2^2)} + \frac{(\mu_1 - \mu_3)^2}{\sigma_1^2(\lambda'^2 + \sigma_3^2)} \right\}^{1/2},$$

$$\tilde{\sigma} = \left\{ \frac{1}{\lambda^2 + \sigma_2^2} + \frac{1}{\lambda'^2 + \sigma_3^2} + \frac{1}{\sigma_1^2} \right\}^{1/2},$$

$$\tilde{\tilde{\mu}} = \frac{\mu_1(\lambda^2 + \sigma_2^2)(\lambda'^2 + \sigma_3^2) + \sigma_1^2\{\mu_2(\lambda'^2 + \sigma_3^2) + \mu_3(\lambda^2 + \sigma_2^2)\}}{(\lambda^2 + \sigma_2^2)(\lambda'^2 + \sigma_3^2) + \sigma_1^2(\lambda^2 + \sigma_2^2 + \lambda'^2 + \sigma_3^2)}.$$

3. Numerical Results

Suppose that the data $X_1, \ldots, X_n$ are contained in the 'box' $[a_1, b_1] \times \cdots \times [a_d, b_d]$ where, without loss of generality, a_i and b_i are integer multiples of δ_i, $i = 1, \ldots, d$. In practice binned kernel estimators are obtained by binning the data on an $M_1 \times \cdots \times M_d$ grid where

$$M_i = (b_i - a_i)/\delta_i + 1, \quad i = 1, \ldots, d.$$

We will call M_i the *grid size* for direction i. While Theorem 1 is in terms of the binwidths δ_i, it is the grid sizes M_i that have direct relevance to computational costs. To see this, let $\kappa_\ell = L(\delta_1 \ell_1, \ldots, \delta_d \ell_d)$, $\ell_i = -M_i, \ldots, M_i$, $i = 1, \ldots, d$. Then it is easy to show that

$$\tilde{\theta}_\delta = n^{-2} \sum_{\ell_1'=1}^{M_1} \cdots \sum_{\ell_d'=1}^{M_d} c_{\ell'} \left(\sum_{\ell_1=1}^{M_1} \cdots \sum_{\ell_d=1}^{M_d} \kappa_{\ell - \ell'} c_\ell \right).$$

The number of distinct arguments of L is $O(M_1 \cdots M_d)$. Moreover, $\tilde{\theta}_\delta$ requires $O(M_1^2 \cdots M_d^2)$ operations if computed directly, although considerable savings are possible by recognising that many of the c_ℓ and κ_ℓ are zero; see, for example, Scott (1992, pp.118, 121). Noting that the inner summation is a discrete convolution, the Fast Fourier Transform can be used to compute $\tilde{\theta}_\delta$ in $O(M_1 \log M_1 \cdots M_d \log M_d)$ operations (Wand, 1994). Either way, there is a genuine trade-off: increasing the M_i's leads to better accuracy, but also increases the computational labour and storage requirements. Therefore, it is of particular interest to develop an understanding of the minimum grid sizes required to achieve a certain level of accuracy.

While Theorem 1 provides a concise quantification of the accuracy of $\tilde{\theta}_\delta$, it suffers from the problem that the mean squared error $E\{(\tilde{\theta}_\delta - \hat{\theta})^2\}$ is scale dependent and therefore does not lend itself to meaningful interpretation. A convenient scale invariant adjustment is the *relative mean squared error* (RMSE) given by

$$\text{RMSE} = E\{(\tilde{\theta}_\delta - \hat{\theta})^2\}/E\{(\hat{\theta} - \theta)^2\},$$

where θ represents the 'target' of the estimator $\hat{\theta}$. Therefore, RMSE is the ratio of the error due to binning to the overall estimation error.

The form of θ depends on the particular setting. For plug-in bandwidth selection we have θ's of the form

$$\theta = \int f^{(r)}(x) f(x)\,dx, \quad |r| \text{ even},$$

while for least-squares cross-validation

$$\theta = E\left\{\int \hat{f}(x;h)^2\,dx - 2\int \hat{f}(x;h) f(x)\,dx\right\} - 2(n-1)^{-1}(h_1 \cdots h_d)^{-1}K(0),$$

where $\hat{f}(\cdot;h)$ is a kernel density estimator with bandwidth vector h and d-variate kernel K. For smoothed cross-validation

$$\theta = \int [E\{\hat{f}(x;h)\} - f(x)]^2\,dx,$$

the integrated squared bias of $\hat{f}(\cdot;h)$.

For brevity's sake we will confine our numerical results to plug-in bandwidth selection with $|r| = 4$. Functionals with $|r| = 4$ are of particular importance in plug-in bandwidth selection since they correspond to the unknown functionals in expressions in the asymptotically optimal bandwidth formulae; see, for example, Wand and Jones (1994) and Sain et al. (1994). In this case the mean squared error of $\hat{\theta}$ can be shown to be

$$\begin{aligned} E\{(\hat{\theta} - \theta)^2\} = {} & n^{-2}\{\phi_h^{(r)}(0) + (n-1)\mathscr{A}(r,h) - n\mathscr{A}(r,0)\}^2 \\ & + 2n^{-3}(n-1)\{\mathscr{B}(r,r,h,h) - \mathscr{A}(r,h)^2\} \\ & + 4n^{-3}(n-1)(n-2)\{\mathscr{C}(r,r,h,h) - \mathscr{A}(r,h)^2\}. \end{aligned}$$

For simplicity we will take $a_i = -3$, $b_i = 3$, $i = 1, \cdots, d$ and $M_1 = \ldots = M_d = M$ in all of our examples. In each setting we will be interested in the *minimum grid size* $M^*(\alpha)$ required to achieve $100\alpha\%$ accuracy, $0 < \alpha < 1$, defined by

$$M^*(\alpha) = \text{smallest } M \text{ such that the approximate RMSE} \leq \alpha.$$

The approximate RMSE is simply the RMSE with $E\{(\tilde{\theta}_\delta - \hat{\theta})^2\}$ replaced by its leading term as $\delta_i \to 0$, given by Theorem 1. Note that RMSE also depends on the bandwidths used by $\hat{\theta}$ and $\tilde{\theta}_\delta$. We will take $h_1 = \cdots = h_d$ to equal the single bandwidth that minimises $E\{(\hat{\theta} - \theta)^2\}$.

Table 1 contains the minimum grid sizes for estimation of $\int f^{(4)}(x) f(x)\,dx$ for the 15 normal mixture densities of Marron and Wand (1992). These are pictured in Fig. 2(a). The results show that a grid size of about 500 guarantees a very accurate approximation for the majority of situations. It is only those densities with a high degree of 'fine structure' that require grid-sizes in the thousands to achieve this 1% accuracy level.

In most situations we see that linear binning is more accurate than simple binning. However, it is interesting to see that there are cases where the asymptotic (as $\delta \to 0$) dominance of linear binning has not taken effect, and simple binning is more accurate.

The minimum grid sizes required in each direction for some bivariate settings are given in Table 2. Density S is the standard bivariate normal density, while densities C, E and F are as defined in the study by Wand and Jones (1993). Fig. 2(b) gives contour plots of these densities. Here we see that we require about the same number

Table 1
Minimum grid sizes to achieve 1% approximate relative MSE for plug-in bandwidth selection and for 15 example normal mixture densities ($r = 4$)

	Sample size					
	$n = 100$		$n = 1000$		$n = 10\,000$	
Density	Simple	Linear	Simple	Linear	Simple	Linear
1	51	30	75	48	114	76
2	77	46	111	73	168	117
3	505	341	673	528	896	828
4	368	252	490	392	692	632
5	486	288	708	462	1078	747
6	73	44	102	71	153	115
7	100	62	140	98	208	156
8	108	65	148	103	209	165
9	117	69	162	110	223	176
10	363	262	498	416	703	662
11	1285	1104	2000	1936	2904	3259
12	690	507	1053	879	1515	1472
13	1153	958	1810	1687	2687	2881
14	1086	885	1628	1516	2306	2502
15	621	478	858	767	1155	1209

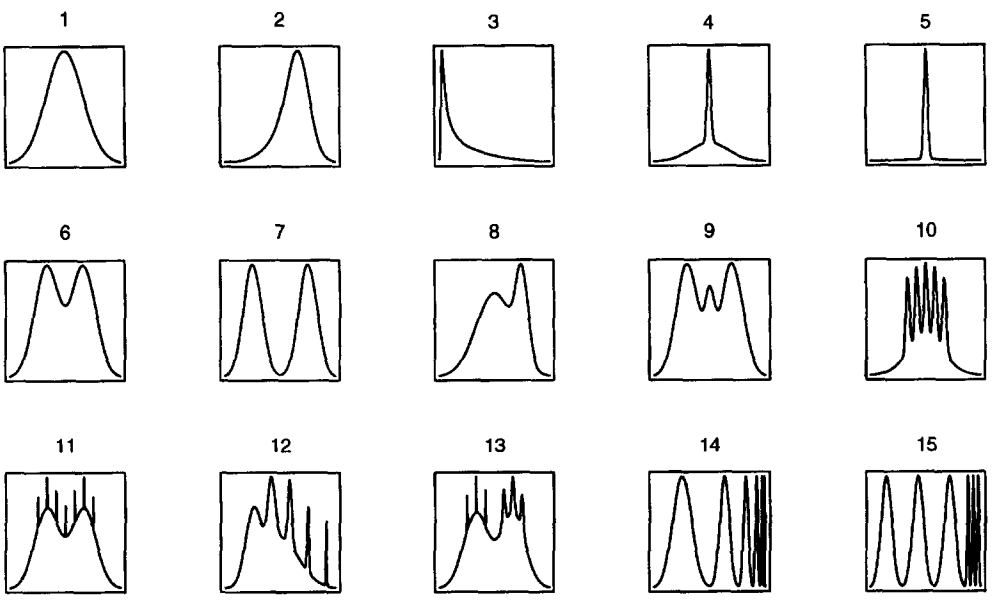
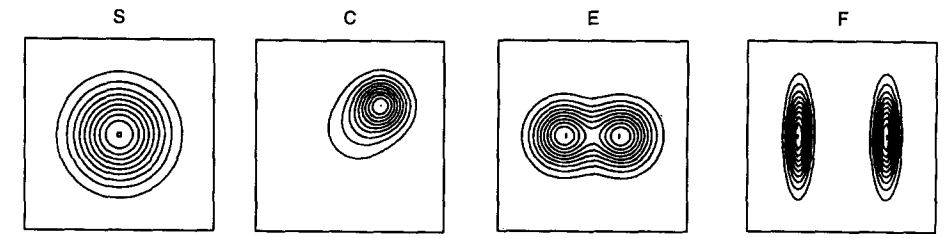


Fig. 2. (a) Graphs of the 15 example univariate densities. (b) Contour plots of the 4 example bivariate densities.



b

Fig. 2. (Continued)

Table 2
Minimum grid sizes to achieve 1% approximate relative MSE for plug-in bandwidth selection and for 4 example bivariate normal mixture densities

	Sample size					
	$n = 100$		$n = 1000$		$n = 10\,000$	
Density	Simple	Linear	Simple	Linear	Simple	Linear
$r_1 = 2, r_2 = 2$						
S	56	42	75	65	103	104
C	88	67	113	104	153	167
E	82	60	103	91	135	144
F	125	101	175	163	242	263
$r_1 = 0, r_2 = 4$						
S	55	38	72	60	100	96
C	85	62	110	96	150	153
E	77	58	100	90	137	144
F	73	49	105	78	151	125
$r_1 = 4, r_2 = 0$						
S	55	38	72	60	100	96
C	85	62	110	96	150	153
E	77	56	97	87	131	140
F	176	160	218	245	283	388

of bins in each direction as for the univariate samples, but these values have to be squared to obtain the total number of grid points in the bivariate mesh.

4. Conclusions

Problems requiring automatic smoothing of large data sets are becoming increasingly more abundant. It is vital that fast and efficient algorithms for performing the required calculations are developed and explored. Binned kernel functional approximations are an important component of one solution to this problem and this paper, for the first time, provides an analysis of their accuracy. As expected, the level of accuracy depends on the problem, but for a wide variety of situations a high level

of accuracy is achieved using a relatively low number of bins in each direction. For univariate settings the minimum grid sizes are all reasonable for practice using contemporary computing environments. The bivariate results indicate that some caution does need to be taken when using binned approximations. For problems where the level of detail is about the same as the example bivariate densities considered in Section 3 grid sizes of about 100×100 will be adequate. More detailed surfaces will require much bigger meshes which will sometimes be beyond storage capabilities in many of the current computing environments.

In closing we remark that the binned approximation is one of a few recent ideas for speeding up computation of kernel estimators. For other ideas see, for example, Loader (1994) and Seifert et al. (1994). It would be interesting to see how these other ideas adapt to the multivariate functional estimation problem and if their accuracy can be quantified so that practical recommendations for their use can be made, as the current paper does for the binned approach.

Appendix A. Proof of Theorem 1.

We first treat simple binning. Taylor expansion leads to

$$\tilde{\theta}_\delta - \hat{\theta} = n^{-2}\sum_{i=1}^{n}\sum_{i'=1}^{n}\sum_{j=1}^{d} L^{(e_j)}(X_i - X_{i'})\delta_j \left\{ Q\left(\delta_j^{-1}X_{i'j}\right) - Q\left(\delta_j^{-1}X_{ij}\right)\right\} + o_p\left\{\left(\sum_{j=1}^{d}\delta_j^2\right)^{1/2}\right\} \tag{A.1}$$

Denote the main part on the right-hand side of (A.1) by $\mathscr{M}$. Then it is clear that

$$E\{(\hat{\theta} - \tilde{\theta}_\delta)^2\} = E(\mathscr{M}^2) + o\left(\sum_{j=1}^{d}\delta_j^2\right).$$

We will decompose $E(\mathscr{M}^2)$ in the mean and the variance terms. Consider V_{ij}, $i = 1,\ldots,n$, $j = 1,\ldots,d$, mutually independent random variables uniformly distributed on the interval $(-\frac{1}{2},\frac{1}{2})$. By applying Lemma 1 (part (a)) below we obtain for the mean,

$$E(\mathscr{M}) = n^{-2}\sum_{i=1}^{n}\sum_{i'=1}^{n}\sum_{j=1}^{d}\delta_j E\left\{L^{(e_j)}(X_i - X_{i'})\right\} E(V_{i'j} - V_{ij}) + o\left(\sum_{j=1}^{d}\delta_j\right)$$

$$= o\left(\sum_{j=1}^{d}\delta_j\right).$$

We obtain the expression in Theorem 1 for simple binning after some straightforward calculations for $\mathrm{Var}(\mathscr{M})$ and the application of Lemma 1 (part (a)) in a similar way as above.

For linear binning the arguments are similar to those for simple binning, although somewhat more long-winded. Taylor expansion leads to

$$\tilde{\theta}_\delta - \hat{\theta} = \frac{1}{2} n^{-2} \sum_{i=1}^{n} \sum_{i'=1}^{n} \sum_{j=1}^{d} L^{(2e_j)}(X_i - X_{i'}) \delta_j^2 \left(S_{ij} + S_{i'j}\right) + o_p\left\{\left(\sum_{j=1}^{d} \delta_j^4\right)^{1/2}\right\}, \tag{A.2}$$

where $S_{ij} = R_{ij}\left(1 - R_{ij}\right)$ and $R_{ij} = R\left(\delta_j^{-1} X_{ij}\right)$ for $i = 1,\ldots,n$, $j = 1,\ldots,d$.

As before, the squared error is dominated by the second moment of the main term on the right-hand side of (A.2).

In this case we must consider U_{ij}, $i = 1,\ldots,n$, $j = 1,\ldots,d$, mutually independent random variables uniformly distributed on the unit interval, and so, after some calculations and applications of Lemma 1 (part (b)) we obtain the desired result for linear binning.

Lemma 1. *Let Z be a d-dimensional random vector with continuous density f. Define $V_{\delta i} = Q\left(\delta_i^{-1} Z_i\right)$, $U_{\delta i} = R\left(\delta_i^{-1} Z_i\right)$ for all $i = 1,\ldots,d$ with the functions Q and R defined as $Q(z) = z -$ (closest integer to z), $R(z) = z - \lfloor z \rfloor$, where $\lfloor z \rfloor$ denotes the integer part of z. Then*

(a) $P(V_\delta \le v, Z \le z) \to \left(\prod_{i=1}^{d} (\tfrac{1}{2} + v_i)\right) \quad P(Z \le z)$ *as* $\delta \to 0$

for all $v \in (-\tfrac{1}{2}, \tfrac{1}{2})^d$ *and* $z \in \mathbb{R}^d$,

(b) $P(U_\delta \le u, Z \le z) \to \left(\prod_{i=1}^{d} u_i\right) \quad P(Z \le z)$ *as* $\delta \to 0$

for all $u \in (0,1)^d$ *and* $z \in \mathbb{R}^d$,

where $Z \le z$ means that $Z_i \le z_i$ for all $i = 1,\ldots,d$.

Proof. This lemma is an extension of Lemma 3 by Hall (1983) to the multidimensional case. We will only prove part (b) of the lemma.

To this aim, it suffices to show that for all $u \in (0,1)^d$ and $z, z' \in \mathbb{R}^d$ $(z < z')$,

$$P\left(U_\delta \le u, z < Z \le z'\right) \to \left(\prod_{i=1}^{d} u_i\right) P\left(z < Z \le z'\right) \tag{A.3}$$

as $\delta \to 0$. Let $Y = \delta^{-1} Z = \left(\delta_1^{-1} Z_1, \ldots, \delta_d^{-1} Z_d\right)$, $y_\delta = \delta^{-1} z$ and $y'_\delta = \delta^{-1} z'$. Then

$$P\left(U_\delta \le u, z < Z \le z'\right) = \left(Y_i - [Y_i] \le u_i, y_{\delta i} < Y_i \le y'_{\delta i}\right)$$

$$= \sum_{y_{\delta 1} < m_1 \le y'_{\delta 1}} \cdots \sum_{y_{\delta d} < m_d \le y'_{\delta d}} P(m_i < Y_i \le m_i + u_i \quad \forall i = 1,\ldots,d) + r_\delta,$$

where

$$|r_\delta| \le 2\sum_{i=1}^{d} \sup_{y_i \in \mathbb{R}} R(y_i < Y_i \le y_i + 1) = 2\sum_{i=1}^{d} \sup_{z_i \in \mathbb{R}} R(z_i < Z_i \le z_i + \delta_i) \to 0$$

as $\delta \to 0$. Thus,

$$P(U_\delta \le u, z < Z \le z') = \sum_{y_{\delta 1} < m_1 \le y'_{\delta 1}} \cdots \sum_{y_{\delta d} < m_d \le y'_{\delta d}} \times \int_{m_1\delta_1}^{(m_1+u_1)\delta_1} \cdots \int_{m_d\delta_d}^{(m_d+u_d)\delta_d} q(z)\, \mathrm{d}z + \mathrm{o}(1) \tag{A.4}$$

as $\delta \to 0$. Since q is uniformly continuous on (z, z'), the series on the right in (A.4) may be written as

$$\left(\prod_{i=1}^{d} u_i\delta_i\right) \sum_{y_{\delta 1} < m_1 \le y'_{\delta 1}} \cdots \sum_{y_{\delta d} < m_d \le y'_{\delta d}} q(m_1\delta_1, \ldots, m_d\delta_d) + \mathrm{o}(1).$$

For the same reason, this quantity equals

$$\left(\prod_{i=1}^{d} u_i\right) \sum_{y_{\delta 1} < m_1 \le y'_{\delta 1}} \cdots \sum_{y_{\delta d} < m_d \le y'_{\delta d}} \int_{m_1\delta_1}^{(m_1+1)\delta_1} \cdots \int_{m_d\delta_d}^{(m_d+1)\delta_d} q(s)\, ds + \mathrm{o}(1)$$

$$= \left(\prod_{i=1}^{d} u_i\right) \int_z^{z'} q(s)\, \mathrm{d}s + \mathrm{o}(1).$$

The result (A.3) follows immediately.

Appendix B. Derivation of results in Section 2.2

The non-trivial part is the derivation of closed-form expressions for

$$\alpha(r, \lambda, \mu_1, \mu_2, \sigma_1, \sigma_2) = \int\int \phi_\lambda^{(r)}(x - y)\phi_{\sigma_1}(x - \mu_1)\phi_{\sigma_2}(y - \mu_2)\,\mathrm{d}x\,\mathrm{d}y,$$

$$\beta(r, r', \lambda, \lambda', \mu_1, \mu_2, \sigma_1, \sigma_2)$$

$$= \int\int \phi_\lambda^{(r)}(x - y)\phi_{\lambda'}^{(r')}(x - y)\phi_{\sigma_1}(x - \mu_1)\phi_{\sigma_2}(y - \mu_2)\,\mathrm{d}x\,\mathrm{d}y$$

and

$$\gamma(r, r', \lambda, \lambda', \mu_1, \mu_2, \mu_3, \sigma_1, \sigma_2, \sigma_3)$$

$$= \int\int\int \phi_\lambda^{(r)}(x - y)\phi_{\lambda'}^{(r')}(x - z)\phi_{\sigma_1}(x - \mu_1)\phi_{\sigma_2}(y - \mu_2)\phi_{\sigma_3}(z - \mu_2)\,\mathrm{d}x\,\mathrm{d}y\,\mathrm{d}z.$$

The expression for α follows from the general result

$$\int \phi_\sigma^{(r)}(x-\mu)\phi_{\sigma'}^{(r')}(x-\mu')\,dx = (-1)^r \phi_{(\sigma^2+\sigma'^2)^{1/2}}^{(r+r')}(\mu-\mu'), \tag{A.5}$$

which is proved, for example, by Wand and Jones (1993). The expression for γ can be obtained by a change of integration order, application of (A.5) and the result

$$\begin{aligned}
&\int \phi_{\sigma_1}^{(r_1)}(x-\mu_1)\phi_{\sigma_2}^{(r_2)}(x-\mu_2)\phi_{\sigma_3}(x-\mu_3)\,dx \\
&= (2\pi)^{-1/2}\phi_{\tilde{\sigma}}(\tilde{\mu})\sum_{j_1=0}^{r_1}\sum_{j_2=0}^{r_2}\binom{r_1}{j_1}H_{r_1-j_1}\left(\frac{\mu_1-\tilde{\tilde{\mu}}}{\sigma_1}\right)\binom{r_2}{j_2}H_{r_2-j_2}\left(\frac{\mu_2-\tilde{\tilde{\mu}}}{\sigma_2}\right) \\
&\times \sigma_1^{-r_1-j_1}\sigma_2^{-r_2-j_2}(\sigma_1\sigma_2\sigma_3)^{-1}\tilde{\tilde{\sigma}}^{j_1+j_2}v(j_1+j_2)
\end{aligned} \tag{A.6}$$

where $\tilde{\mu} = \left[\sum\sum_{i<j}\{(\mu_i-\mu_j)/(\sigma_i\sigma_j)\}^2\right]^{1/2}$, $\tilde{\tilde{\mu}} = \sum_{i=1}^3 \sigma_i^{-2}\mu_i \Big/ \sum_{i=1}^3 \sigma_i^{-2}\mu_i$ and $\tilde{\sigma}^2 = \sum_{i=1}^3 \sigma_i^{-2}$. The proof of (A.6) is given in an unpublished manuscript by B. Aldershof, J.S. Marron, B.U. Park and M.P. Wand. It can be accomplished by straightforward, albeit long-winded, algebra using the explicit formula for the coefficients of a normalised Hermite polynomial. The result for β can be derived using the same type of argument that is required to derive (A.6) together with the result

$$\phi_\sigma(x-\mu)\phi_{\sigma'}(x-\mu') = \phi_{(\sigma^2+\sigma'^2)^{1/2}}(\mu-\mu')\phi_{\sigma\sigma'/(\sigma^2+\sigma'^2)^{1/2}}(x-\mu^*),$$

where $\mu^* = (\sigma'^2\mu+\sigma^2\mu')/(\sigma^2+\sigma'^2)$ (Wand and Jones, 1993).

References

Bowman, A.W., An alternative method of cross-validation for the smoothing of density estimates, *Biometrika*, **71** (1984) 353–360.

Cao, R., A. Cuevas and W. González-Manteiga, A comparative study of several smoothing methods in density estimation, *Comput. Statist. Data Anal.*, **17** (1994) 153–176.

Hall, P., Edgeworth expansion of the distribution of Stein's statistic, *Math. Proc. Cambridge Philos. Soc.*, **93** (1983) 163–175.

Hall, P. and J.S. Marron, Estimation of integrated squared density derivatives, *Statist. Probab. Lett.*, **6** (1987) 109–115.

Hall, P., J.S. Marron and B.U. Park, Smoothed cross-validation, *Probab. Theory Rel. Fields*, **92** (1992) 1–20.

Hall, P. and M.P. Wand, On the accuracy of binned kernel density estimators, *J. Multiv. Anal.*, (1996) to appear.

Härdle, W. and D.W. Scott, Smoothing by weighted averaging using rounded points, *Comput. Statist.*, **7** (1992) 97–128.

Jones, M.C. and H. Lotwick, On the errors involved in computing the empirical characteristic function, *J. Statist. Comput. Simulation*, **17** (1983) 133–149.

Jones, M.C., J.S. Marron, and S.J. Sheather, Progress in data-based bandwidth selection for kernel density estimation, *J. Amer. Statist. Assoc.*, to appear.

Jones, M.C. and S.J. Sheather, Using non-stochastic terms to advantage in kernel-based estimation of integrated squared density derivatives, *Statist. Probab. Lett.*, **11** (1991) 511-514.

Loader, C. Computing nonparametric function estimates, in: *Computing Science and Statistics: Proc. 26th Symp. on the Interface*, Interface Foundation of North America, 1994, to appear.

Marron, J.S. and M.P. Wand, Exact mean integrated squared error, *Ann. Statist.*, **13** (1992) 712–736.

O'Sullivan, F. and Y. Pawitan, Multivariate density estimation by tomography, *J. Roy. Statist. Soc. Ser. B*, **55** (1993) 509–521.

Park, B.U. and J.S. Marron, Comparison of data-driven bandwidth selectors, *J. Amer. Statist. Assoc.*, **85** (1990) 66–72.

Rudemo, M. Empirical choice of histograms and kernel density estimators, *Scand. J. Statist.*, **9** (1982) 65–78.

Sain, S., K. Baggerly, and D.W. Scott, Cross-validation of multivariate densities, *J. Amer. Statist. Assoc.*, **89** (1994) 807–817.

Scott, D.W. Average shifted histograms: effective nonparametric density estimators in several dimensions, *Ann. Statist.*, **13** (1985) 1024–1040.

Scott, D.W., *Multivariate Density Estimation: Theory, Practice and Visualization* (Wiley, New York; 1992).

Scott, D.W. and S.J. Sheather, Kernel density estimation with binned data, *Comm. Statist. Theory Metheods*, **14** (1985) 1353–1359.

Scott, D.W. and G.R. Terrell, Biased and unbiased cross-validation in density estimation, *J. Amer. Statist. Assoc.*, **82** (1987) 1131–1156.

Seifert, B., M. Brockmann, J. Engel, and Th. Gasser, Fast algorithms for nonparametric curve estimation. *J. Comput. Graph. Statist.*, **3** (1994) 192–213.

Sheather, S.J. and M.C. Jones, A reliable data-based bandwidth selection method for kernel density estimation, *J. Roy. Statist. Soc. Ser. B*, **53** (1991) 683–690.

Silverman, B.W. Kernel density estimation using the fast Fourier transform, *Appl. Statist.*, **31** (1982) 93–99.

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

Taylor, C.C., Bootstrap choice of the smoothing parameter in kernel density estimation, *Biometrika*, **76** (1989) 705–712.

Wand, M.P., Fast computation of multivariate kernel estimators, *J. Comput. Graph. Statist.*, **3** (1994) 433–445.

Wand, M.P. and M.C. Jones, Comparison of smoothing parametrizations for bivariate kernel density estimation, *J. Amer. Statist. Assoc.*, **88** (1993) 520–528.

Wand, M.P. and M.C. Jones, Multivariate plug-in bandwidth selection, *Comput. Statist.*, **9** (1994) 97–116.