ELSEVIER

Computational Statistics & Data Analysis 39 (2002) 165–186

COMPUTATIONAL STATISTICS & DATA ANALYSIS

www.elsevier.com/locate/csda

Multivariate locally adaptive density estimation

Stephan R. Sain

Department of Statistical Science, Southern Methodist University, Dallas, TX 75275-0332, USA

Received 1 January 2001

Abstract

Multivariate versions of variable bandwidth kernel density estimators can lead to improvement over kernel density estimators using global bandwidth choices. These estimators are more flexible and better able to model complex (multimodal) densities. In this work, two variable bandwidth estimators are discussed: the balloon estimator which varies the smoothing matrix with each estimation point and the sample point estimator which uses a different smoothing matrix for each data point. A binned version of the sample point estimator is developed that, for various situations in low to moderate dimensions, exhibits less error (MISE) than the fixed bandwidth estimator and the balloon estimator. A practical implementation of the sample point estimator is shown through simulation and example to do a better job at reconstructing features of the underlying density than fixed bandwidth estimators. Computational details, including parameterization of the smoothing matrix, are discussed throughout.

Keywords: Binning; Cross-validation; Mean integrated squared error; Smoothing matrix; Variable bandwidth kernel

1. Introduction

The kernel density estimator has become a staple in the data analysis tool box largely because of the flexibility of the method. Research has shown theoretical superiority over such naive estimators as histograms and the practical benefits are clear. With cross-validation (Rudemo, 1982; Bowman, 1984) and plug-in (Sheather and Jones, 1991) rules for choosing smoothing parameters, data-driven, automatic bandwidth selection has achieved a certain level of maturity.

E-mail address: ssain@mail.smu.edu (S.R. Sain).

0167-9473/02/$ - see front matter
PII: S 0167-9473(01)00053-6

Bandwidths are generally chosen based on the relatively simple but important idea of balancing bias and variance globally. Consider the multivariate kernel density estimator given by

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

where $\mathbf{x}_1, \ldots, \mathbf{x}_n$ is a d-variate random sample with density f. The kernel, K, is taken to be a d-variate density function with $\int K(\mathbf{w})\,\mathrm{d}\mathbf{w} = \mathbf{1}$, $\int \mathbf{w}K(\mathbf{w})\,\mathrm{d}\mathbf{w} = \mathbf{0}$, and $\int \mathbf{w}\mathbf{w}'K(\mathbf{w})\,\mathrm{d}\mathbf{w} = \mathbf{I}_d$. The contours of the kernel are restricted to be spherically symmetric and the smoothing parameter, h, controls the size of the kernel.

Straightforward asymptotic approximations yield integrated squared bias (ISB) and integrated variance (IV) equal to

$$\mathrm{IV} = \frac{R(K)}{nh^d} \quad \text{and} \quad \mathrm{ISB} = \frac{1}{4}h^4\sigma_K^4 \int \mathrm{tr}^2\{\bigtriangledown^2 f(\mathbf{x})\}\,\mathrm{d}\mathbf{x},$$

where $R(K) = \int K^2(\mathbf{w})\,\mathrm{d}\mathbf{w}$, tr indicates the trace of a matrix, and $\bigtriangledown^2 f(\mathbf{x})$ is the Hessian (matrix of second partial derivatives) of f. Combining these terms yields an estimate of the mean integrated squared error, $\mathrm{MISE} = E\int\{\hat{f}(\mathbf{x}) - f(\mathbf{x})\}^2\,\mathrm{d}\mathbf{x}$. The optimal bandwidth can then be easily derived and is equal to

$$h_{\mathrm{AMISE}} = \left[\frac{\mathrm{d}R(K)}{\sigma_K^4 \int \mathrm{tr}^2\{\bigtriangledown^2 f(\mathbf{x})\}\,\mathrm{d}\mathbf{x}}\right]^{1/(d+4)} n^{-1/(d+4)}.$$

This choice of h yields an $\mathrm{MISE} = O(n^{-4/(d+4)})$.

This bias-variance trade-off works well for many densities, especially those densities that are unimodal and are not extremely skewed. However, as the dimensionality increases, the so-called curse of dimensionality becomes influential. The effect on multivariate density estimates is well noted (Friedman and Stuetzle, 1981; Scott and Wand, 1991). Due to the sparseness of data in higher dimensions, multivariate neighborhoods are generally empty, particularly in the "tails" of the density (Exercise 4.1 in Wand and Jones, 1995). Furthermore, there is less probability mass near the mode. Not only does this have the effect of slowing the convergence rate of the MISE as dimensionality increases, but the relative contributions of variance and bias change. In the univariate setting, the ratio of bias to variance is $4:1$. For general dimensions, this ratio is $4:d$. Hence, as the dimension increases, larger and larger bandwidths are necessary to control the increased variability, particularly the contributions from the tails. However, this also has adverse effect of averaging away features near the modes.

Sain and Scott (1996) give another example of where this trade-off between variance and bias fails. Consider a bimodal normal mixture of the form $f(x) = 3/4\phi(x+3/2)+1/4\phi_{1/3}(x-3/2)$, where $\phi_\sigma(x-\mu)$ is a normal density with mean μ and variance σ^2. This mixture is characteristic of a density that is difficult for the kernel estimator in that the modes are of equal height but have differing scales. The globally optimal smoothing parameter of a sample of size $n = 200$ is $h = 0.248$. Considering the sample sizes associated with each mode, the optimal smoothing parameters are $h = 0.403$ and 0.175, respectively. Again, the global value reflects an attempt to find some sort

of middle ground between what is optimal for each mode. However, this value of h will undersmooth or oversmooth, depending on the mode, and the analyst will be faced with the difficult decision of choosing which features are real and which are noise.

Graphical ideas such those proposed by Minnotte and Scott (1993) and Chaudhuri and Marron (1999) have proven to be effective aids for exploring structure in univariate multimodal densities. Both approaches use a family of kernel estimates, where the bandwidth is allowed to range from small to large. These methods are attractive in that they allow the user to change "resolution" by focusing on large bandwidths to gain insight into general structure and then smaller bandwidths for finer detail. However, they serve to emphasize the lesson that a single smoothing parameter can be ineffective for more complex densities.

In low to moderate dimensions, varying the amount of smoothing is a possible solution. Intuitively, the task seems straightforward. More smoothing is necessary where data are scarce (i.e. tails and valleys) and less smoothing is necessary near modes. However, it is not as simple as modeling the bandwidth as a function of the height of the underlying density alone. Some attempt to account for the curvature must be made as well (Sain and Scott, 1996). Actually, it has proven difficult to gain understanding of exactly how to vary the bandwidth and, furthermore, how to apply this understanding to varying the bandwidth in practice, particularly in the multivariate case.

In this paper, variable bandwidth estimators will be studied in the multivariate setting. Following a brief introduction to the fixed bandwidth multivariate kernel estimator (Section 2), a multivariate version of the binned sample-point estimator is introduced and a comparison of two common approaches to variable bandwidth estimators is presented (Sections 3 and 4). Section 5 details practical algorithms for implementing a variable bandwidth estimator and discusses some of the issues and problems associated with variable bandwidth estimation in practice.

2. Fixed bandwidth estimators

The general multivariate kernel density estimator is given by

$$\hat{f}(\mathbf{x}) = \frac{1}{n|\mathbf{H}|^{1/2}} \sum_{i=1}^{n} K(\mathbf{H}^{-1/2}(\mathbf{x} - \mathbf{x}_i)) = \frac{1}{n} \sum_{i=1}^{n} K_{\mathbf{H}}(\mathbf{x} - \mathbf{x}_i). \tag{2}$$

Unlike (1) the smoothing parameter, $\mathbf{H}$, is actually a symmetric positive-definite matrix that is analogous to the covariance matrix of K. Hence, $K_{\mathbf{H}}(\mathbf{x}) = |\mathbf{H}|^{-1/2} K(\mathbf{H}^{-1/2}\mathbf{x})$.

Straightforward asymptotic analysis involving multivariate Taylor's series expansions can yield an estimate of the MISE. Following Scott (1992) or Wand and Jones (1995), an asymptotic approximation of the MISE is given by

$$\mathrm{AMISE} = \frac{R(K)}{n|\mathbf{H}|^{1/2}} + \frac{1}{4} \int \mathrm{tr}^2(\mathbf{H} \bigtriangledown^2 f(\mathbf{x}))\, \mathrm{d}\mathbf{x}, \tag{3}$$

where $|\cdot|$ indicates the determinant of a matrix.

Choosing the form of $\mathbf{H}$ depends on the complexity of the underlying density and the number of parameters that must be estimated. Wand and Jones (1993) give an excellent discussion of the issues in the bivariate case. There are three primary classes for parameterizing the smoothing matrix. Assuming that K is a multivariate normal kernel, i.e. $K(\mathbf{x}) = (2\pi)^{-d/2}\exp(-\mathbf{x}'\mathbf{x}/2)$, the first class was introduced in (1). This kernel has contours that are spherically symmetric and it has only one independent smoothing parameter. In keeping with the notation of Wand and Jones (1993), this class will be denoted as $\mathscr{H}_1$ where $\mathscr{H}_1 = \{h^2\mathbf{I};\ h > 0\}$.

The second class, $\mathscr{H}_2$, allows kernels that are ellipsoidal. However, the axes are restricted to be parallel to the coordinate axis. Here, d independent smoothing parameters are allowed and $\mathscr{H}_2 = \{\mathrm{diag}(h_1^2,\ldots,h_d^2);\ h_1,\ldots,h_d > 0\}$, where diag indicates a diagonal matrix. This class is commonly referred to as the product kernel estimator and allows different amounts of smoothing in each dimension. The optimal bandwidths, $h_1,\ldots,h_d$ are still proportional to $n^{-1/(d+4)}$ and the AMISE is of the same order as $\mathscr{H}_1$. However, some gain can be obtained by using different smoothing parameters for each dimension, especially when the scales of the variables differ.

The final class, in which a full smoothing matrix is employed, involves $d(d+1)/2$ independent smoothing parameters and is denoted as $\mathscr{H}_3$. This class allows ellipsoidal kernels of arbitrary orientation and is given by

$$\mathscr{H}_3 = \left\{ \begin{bmatrix} h_1^2 = h_{11} & \cdots & h_{1d} \\ \vdots & \ddots & \vdots \\ h_{1d} & \cdots & h_d^2 = h_{dd} \end{bmatrix};\ h_1,\ldots,h_d > 0,\ |h_{ij}| < h_i h_j,\ i \neq j \right\}.$$

Unfortunately, there is no closed form expression for the optimal smoothing matrix in this case. Numerical methods are required to find $\mathbf{H}$ for a candidate density and a particular sample size.

In practice, a common approach is to scale the data so that the sample variances are the same in each dimension or to sphere the data in which a linear transformation is applied that yields an identity sample covariance matrix. These approaches are essentially dimension reduction ideas as they allow for the use of a single smoothing parameter ($\mathscr{H}_1$) with the transformed data. However, Wand and Jones (1993) urge caution when using these approaches as they are not guaranteed to give the correct transformation or rotation to achieve the gains possible by using the full smoothing matrix. Many authors have noted that for most densities, in particular unimodal ones, allowing different amounts of smoothing for each dimension (the product kernel estimator, $\mathscr{H}_2$) is adequate. With more complex densities, especially multimodal ones, the situation is less clear, although rotations can help if the structure of the distribution can be aligned with the coordinate axis (Wand and Jones, 1993).

3. Locally adaptive density estimation

One could consider a bandwidth function that adapts to not only the point of estimation, but also the observed data points and the shape of the underlying density. As

a matter of practice, however, two simplified versions are common. The first varies the bandwidth at each estimation point and is referred to as the *balloon estimator*. The term balloon estimator was first used by Terrell and Scott (1992) and is based on a suggestion of Tukey and Tukey (1981). The second varies the bandwidth for each data point and is referred as the *sample-point estimator*. See Jones (1990) for a detailed comparison of the two estimators in the univariate case.

3.1. Balloon estimators

The general form of the balloon estimator is given by

$$\hat{f}_{\mathrm{B}}(\mathbf{x}) = \frac{1}{n|\mathbf{H}(\mathbf{x})|^{1/2}} \sum_{i=1}^{n} K(\mathbf{H}(\mathbf{x})^{-1/2}(\mathbf{x}-\mathbf{x}_i)) = \frac{1}{n} \sum_{i=1}^{n} K_{\mathbf{H}(\mathbf{x})}(\mathbf{x}-\mathbf{x}_i), \tag{4}$$

where $\mathbf{H}(\mathbf{x})$ is the smoothing matrix for the estimation point $\mathbf{x}$. Considered pointwise, the construction of this estimator is the same as the fixed bandwidth estimator given in (2). For each point at which the density is to be estimated, kernels of the same size and orientation are centered at each data point and the density estimate is computed by taking the average of the heights of the kernels at the estimation point.

This type estimator was introduced by Loftsgaarden and Quesenberry (1965) as the kth nearest-neighbor estimator which can be written as in (4) by taking K to be a uniform density on the unit sphere and by restricting the smoothing matrix to $\mathscr{H}_1$. Thus, the bandwidth function can be written as $h_k(\mathbf{x})$ which measures the distance from $\mathbf{x}$ to the kth nearest data point.

Much has been written about the kth nearest neighbor estimator that suggests it is not an effective density estimator in the univariate case. The bandwidth function is discontinuous and these discontinuities manifest themselves directly in the density estimate (Silverman, 1986). Furthermore, the estimator suffers from severe bias problems, particularly in the tails (Mack and Rosenblatt, 1979; Hall, 1983; Terrell and Scott, 1992). However, Terrell and Scott (1992) show that the kth nearest-neighbor estimator improves as dimensionality increases and will perform well in dimensions greater than 4.

Terrell and Scott (1992) also study error properties of the general balloon estimator and found some remarkable results. Applied pointwise, the balloon estimator behaves just like the fixed bandwidth estimator. The authors show that by choosing the orientation of the smoothing matrix appropriately, the bias can be dramatically reduced. In fact, the bias will be of higher order for points in certain regions of the underlying density. These regions correspond to where the density is saddle shaped, e.g. outside the unit ball for a multivariate standard normal density.

3.2. Sample-point estimators

The multivariate sample-point estimator is given by

$$\hat{f}_{\mathrm{S}}(\mathbf{x}) = \frac{1}{n} \sum_{i=1}^{n} \frac{1}{|\mathbf{H}(\mathbf{x}_i)|^{1/2}} K(\mathbf{H}(\mathbf{x}_i)^{-1/2}(\mathbf{x}-\mathbf{x}_i)) = \frac{1}{n} \sum_{i=1}^{n} K_{\mathbf{H}(\mathbf{x}_i)}(\mathbf{x}-\mathbf{x}_i), \tag{5}$$

where $\mathbf{H}(\mathbf{x}_i)$ is the smoothing matrix associate with $\mathbf{x}_i$. The sample-point estimator still places a kernel at each data point, but these kernels each have their own size and orientation regardless of where the density is to be estimated.

This type of estimator was introduced by Breiman et al. (1977), who suggested using $\mathbf{H}(\mathbf{x}_i) = h_k(\mathbf{x}_i)\mathbf{I}$. Asymptotically, this is equivalent to choosing bandwidths proportional to $f(\mathbf{x}_i)^{-1/d}$ where d is the dimension of the data.

Abramson (1982a, b) suggested the square-root law, i.e. using bandwidths proportional to $f(\mathbf{x}_i)^{-1/2}$ and, in practice, using a pilot estimate of the density to calibrate the bandwidth function. This formulation of the bandwidth function has been popular in no small part due to the early results that show that the pointwise bias associated with the square-root law was of higher order (Silverman, 1986; Hall and Marron, 1988; Jones, 1990). However, recent work has shown that this early result may not always hold globally due to bias contributions from the tails (Hall, 1992; McKay, 1993; Terrell and Scott, 1992; Hall et al., 1994; Sain and Scott, 1996).

The square-root law also suffers from a certain inflexibility by restricting the bandwidth to be only a function of the height of the density. In order to provide a more general study of the properties of the sample-point estimator, Sain and Scott (1996) introduce a binned version that uses a piecewise constant bandwidth function. In that work, the authors showed that the estimator did not exhibit a higher-order MISE but it did lead to substantial improvement over the fixed bandwidth estimator in the univariate case.

A multivariate version of the binned sample-point estimator is given by

$$\hat{f}_{\mathrm{S}}(\mathbf{x}) = \frac{1}{n}\sum_{j=1}^{m} \frac{n_j}{|\mathbf{H}(\mathbf{t}_j)|^{1/2}} K(\mathbf{H}(\mathbf{t}_j)^{-1/2}(\mathbf{x}-\mathbf{t}_j)) = \frac{1}{n}\sum_{j=1}^{m} n_j K_{\mathbf{H}(\mathbf{t}_j)}(\mathbf{x}-\mathbf{t}_j), \qquad (6)$$

where n_j is the number of data points in the jth bin, $\mathbf{t}_j$ is the center of the jth bin, and $\mathbf{H}(\mathbf{t}_j)$ is the smoothing matrix associated with the jth bin. In general, an equally spaced mesh of points is laid down over the support of the density to define the bins, although other binning rules such as the linear binning defined in Hall and Wand (1996) could be considered.

Binning has been used in density estimation for a variety of reasons. Hall (1982) studied rounded and truncated data. Silverman (1982), Härdle and Scott (1992), Wand (1994), and Hall and Wand (1996) use binning as a device that can radically reduce computing time. Scott and Sheather (1985) build on the result of Hall (1982) and show that binning results in an inflated bias but that the MISE is insensitive to reasonable amounts of binning.

For adaptive estimation, binning becomes much more than a computational tool. Computing the bias of an adaptive estimator is difficult, if not impossible, without specifying the form of the bandwidth function as was done with the square root law. Through binning, the expectation of the estimator in (6) is easily computed by noting that the only random quantities are the n_j, $j = 1, \ldots, m$. These counts can be thought of as a realization of a multinomial distribution with parameters $p_j = \int_{B_j} f(\mathbf{x})\,\mathrm{d}\mathbf{x}$, where B_j denotes the jth multivariate bin.

Assuming that K is a multivariate normal kernel, the MISE of the multivariate binned-sample point estimator is explicitly given by

$$\begin{aligned}
\mathrm{MISE} &= E\int [\hat{f}_{\mathrm{S}}(\mathbf{x}) - f(\mathbf{x})]^2\,\mathrm{d}\mathbf{x} \\
&= \frac{1}{n^2}\sum_j En_j^2 \int \phi^2_{\mathbf{H}_j}(\mathbf{x}-\mathbf{t}_j)\,\mathrm{d}\mathbf{x} \\
&\quad + \frac{1}{n^2}\sum_{i\neq j}\sum En_in_j \int \phi_{\mathbf{H}_i}(\mathbf{x}-\mathbf{t}_i)\phi_{\mathbf{H}_j}(\mathbf{x}-\mathbf{t}_j)\,\mathrm{d}\mathbf{x} \\
&\quad - \frac{2}{n}\sum_j En_j \int \phi_{\mathbf{H}_j}(\mathbf{x}-\mathbf{t}_j) f(\mathbf{x})\,\mathrm{d}\mathbf{x} + R(f) \\
&= \frac{1}{n(2\sqrt{\pi})^d}\sum_j \frac{p_j(1-p_j)+np_j^2}{|\mathbf{H}_j|^{1/2}} + \frac{n-1}{n}\sum_{i\neq j}\sum p_ip_j\phi_{\mathbf{H}_i+\mathbf{H}_j}(\mathbf{t}_i-\mathbf{t}_j) \\
&\quad - \frac{2}{n}\sum_j p_j \int \phi_{\mathbf{H}_j}(\mathbf{x}-\mathbf{t}_j) f(\mathbf{x})\,\mathrm{d}x + R(f), \qquad (7)
\end{aligned}$$

where $\mathbf{H}_j = \mathbf{H}(\mathbf{t}_j)$. Note that the normal integrals follow directly from formulae such as those presented in the appendices to Wand and Jones (1995). By specifying f, usually as some sort of normal mixture, the probabilities p_j, $j=1,\ldots,m$ can be computed and then the MISE function optimized over the collection of smoothing matrices $\mathbf{H}_j$, $j=1,\ldots,m$.

4. A comparison of variable bandwidth methods

Comparing optimal bandwidth functions for the balloon and the sample-point estimator is a difficult task, especially in the multivariate setting. However, developing an understanding of how optimal smoothing parameters behave is important and will yield insight into designing practical algorithms. It would seem reasonable that there is some sort of fundamental relationship between the two methods, at least asymptotically. Sain and Scott (1996) noticed some similarities between the sample-point estimator and the so-called zero-bias bandwidths discussed by Hazelton (1998), Sain and Scott (2001), and Devroye and Lugosi (2000). However, for finite multivariate samples, some differences appear.

To illustrate, let f to be a bivariate standard normal, K be standard normal, and $n=100$. Fig. 1 display optimal kernels for the fixed bandwidth estimator, the balloon estimator and the binned sample-point estimator using an equally spaced mesh with 9 bins per dimension. Two versions of the sample-point estimator are considered. One restricts the contours of the kernels to be circular ($\mathscr{H}_1$) while the second allows unrestricted size and orientation ($\mathscr{H}_3$). See Sain (1994) for details on computing the optimal kernels for the balloon estimator.

Four cases are considered, corresponding to bins centered at the origin and along the x-axis at 0.9, 1.8, and 2.7. These correspond to the mesh for the sample-point estimator being laid down on the square defined by $(-3,3)\times(-3,3)$. Since the

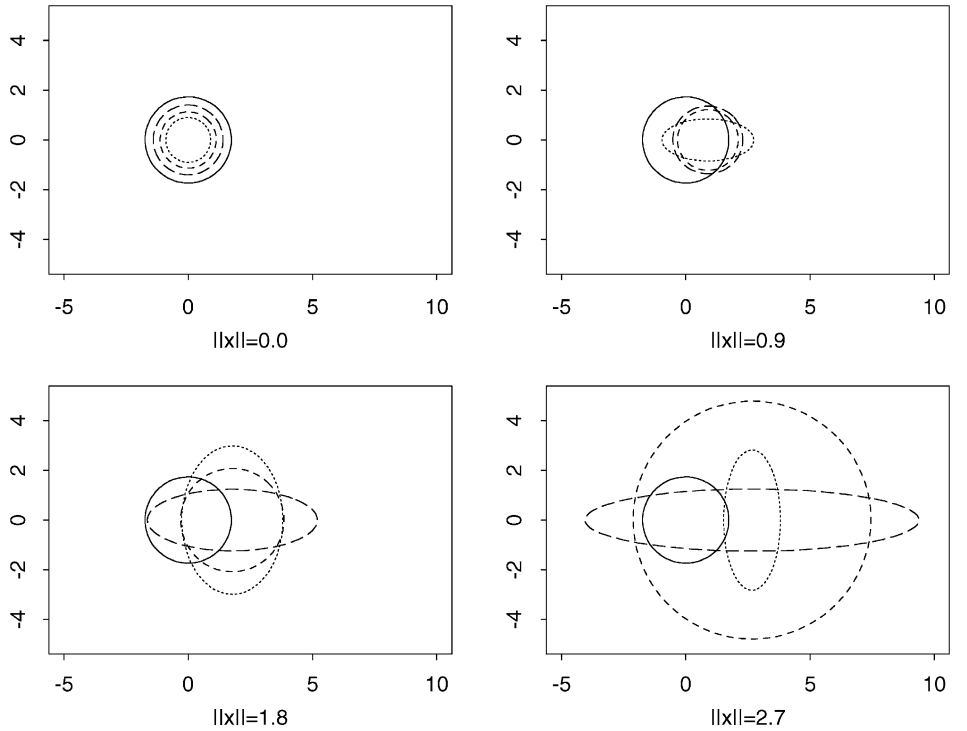

Fig. 1. Optimal kernels for a fixed bandwidth estimator (solid line), balloon estimator (dotted lines), sample-point estimator with the smoothing matrix restricted to $\mathscr{H}_1$ (short dashed lines), and sample-point estimator with an unrestricted smoothing matrix (long dashed lines). Ellipses are 95% contours of the kernels.

contours of the density are radially symmetric, it follows that the contours of the optimal kernels centered at different bins, but at the same distance from the origin, will behave similarly.

The optimal kernel for the fixed bandwidth estimator has contours that are circular regardless of where the density is being estimated, which data point, or which parameterization of the smoothing matrix is used. The 95% contour of this kernel is plotted in Fig. 1 as a solid line. It is centered at the origin as a reference and an indication of scale for the variable bandwidth kernels.

When examining the optimal kernels for the balloon estimator and the sample-point estimator, it is important to keep in mind the construction of the estimator. For the balloon estimators, the same kernel is used for all data points and it is allowed to vary with the estimation point. For the binned-sample point estimator, the density is estimated by placing a weighted kernel at the center of the bin with each bin having a different size and orientation. In practice, the kernels are assumed to be the same for each data point in the bin. These kernels do not change with the estimation point.

First, consider the kernels at the origin ($||x|| = 0.0$) in Fig. 1. All of the kernels are circular which is no surprise considering the structure of this density at the origin. However, all of the variable bandwidth kernels are smaller than the fixed bandwidth

Table 1
MISE values for fixed and sample-point adaptive methods using a normal kernel with $\mathscr{H}_1$ and a $N(\mathbf{0}, \mathbf{I}_d)$ density. Sample size is $n = 100$

Estimator	Bins per dimension	$d = 1$	$d = 2$	$d = 3$
Fixed		0.00541	0.00431	0.00240
Sample-point	$m = 5$	0.00926	0.00655	0.00201
	$m = 9$	0.00333	0.00306	0.00173
	$m = 15$	0.00305	0.00273	0.00162

kernel. In the next plot, $||x|| = 0.9$, the center of the bin is still within the unit circle. The fixed bandwidth kernel is still much larger than the sample-point kernels which are still roughly circular. The balloon estimator kernel is already adjusting to the local curvature and the major axis of the ellipse is parallel to the line between the origin to the center of the bin.

The third plot, $||x|| = 1.8$ shows the kernels for the bin just outside the unit circle. Now the variable bandwidth kernels are much larger than the fixed bandwidth kernel. However, a distinct change in orientation is noted between the balloon kernel and the unrestricted sample-point kernel. Recall the balloon kernel has higher-order bias in this region and is based on the fourth-order derivatives of the density. Hence, the orientation changes, with the major axis of the ellipse being perpendicular to the line between the origin and the center of the bin. The major axis of the unrestricted kernel retains the lower-order bias behavior with the major axis of the kernel parallel to the origin.

Moving even further out into the tails of the density, $||x|| = 2.7$, the variable bandwidth kernels dwarf the fixed bandwidth kernel, with the restricted sample-point kernel being exceptionally large. This phenomenon is explained by the sample-point estimator with the restricted kernels attempt to minimize variability in the tails by trading a much larger size for the correct orientation. Note that the kernel for the sample-point estimator is much larger than the balloon estimator kernel.

The binned sample-point estimator allows head-to-head comparisons between a general sample-point estimator and the fixed bandwidth estimator or the balloon estimator. In actuality, the binned sample-point estimator is not entirely general in that the smoothing matrix is assumed to be constant for all data points in a particular bin. However, through the optimization, the smoothing matrices are allowed to adapt to both level and curvature, something that the square-root law cannot accomplish.

Scott and Sheather (1985) and Hall and Wand (1996) showed that binning can inflate the bias. Sain and Scott (1996) showed that the binned sample-point estimator needed enough bins to counteract this bias and give sufficient flexibility to improve on the fixed bandwidth estimator. The same is true in the multivariate case. In fact, Table 1 shows that more bins per dimension may be needed to achieve that same gain. Table 1 is based on a multivariate standard normal density and $n = 100$. For the fixed bandwidth method, the MISE was computed using the expressions derived by Worton (1989) and Marron and Wand (1992). The kernels for the binned sample-point estimator were restricted to be spherically symmetric ($\mathscr{H}_1$) for

Table 2
MISE values for fixed, sample-point, and balloon estimators using a normal kernel and a $N(\mathbf{0}, \mathbf{I}_2)$. For the sample-point estimator, $m = 9$ when $n = 100$ and $m = 15$ when $n = 1000$

Estimator		$n = 100$	$n = 1000$
Fixed		0.00431	0.00106
Sample-point	$\mathbf{H} \in \mathscr{H}_1$	0.00306	0.000644
	$\mathbf{H} \in \mathscr{H}_3$	0.00211	0.000450
Balloon		0.00252	0.000542

simplicity. An equally spaced mesh was laid down over the approximate support of the density (assumed to be $(-3, 3)$ in each dimension for consistency).

What is clear from Table 1 is that considerable gain in terms of the MISE can be achieved with a relatively small number of bins per dimension and restricting the shape of the kernels. For $d = 1$, using only five bins was actually worse than the fixed bandwidth estimator. However, for $d = 3$, using five bins per dimension (125 total bins) led to some gain over the fixed bandwidth estimator.

Using $m = 15$ bins achieved a 44% gain in MISE for $d = 1$ and a 33% gain for $d = 3$. This suggests that it may be necessary to use more bins per dimension or use the more general smoothing matrices for each bin. In Table 2, MISE values are also computed for the bivariate standard normal. However, a sample-point estimator ($m = 9$) is included that allows unrestricted smoothing matrices. A considerable gain in terms of the MISE is realized as the restriction on the kernels is removed. The restricted kernel sample-point estimator with $m = 9$ and $n = 100$ bins leads to a 29% improvement in the MISE while the unrestricted smoothing matrices leads to a 51% reduction in the MISE.

The binned sample-point estimator also allows a direct comparison with the balloon estimator. In this case, the unrestricted sample-point estimator actually gives a slight improvement over the balloon estimator. The error of the balloon estimator is dominated by contributions from the mode as it is of higher order in the tails. This implies that the sample-point estimator is doing a better job than the balloon estimator even near the mode. The improvement offered by the sample-point estimator is not a small sample curiosity as the improvement is also achieved for the more moderate sample size of $n = 1000$. However, for larger d, the results will change as the impact of the tails (where the balloon estimator is of higher order) becomes more pronounced.

While varying the bandwidth can lead to improvements over the traditional bias-variance trade-off, the real power of the sample-point estimator may lie in the ability to model more complex, multimodal distributions. Consider the normal mixture density, $f(\mathbf{x}) = \sum w_j \phi_{\Sigma_j}(\mathbf{x} - \mu_j)$, labeled "K" in Wand and Jones (1993). This is a bivariate, trimodal density comprised of three normal components with parameters

$$w_1 = 3/7, \quad w_2 = 3/7, \quad w_3 = 1/7,$$
$$\mu_1 = (-1, 0)^t, \quad \mu_2 = \left(1, \frac{2\sqrt{3}}{3}\right)^t, \quad \mu_3 = \left(1, -\frac{2\sqrt{3}}{3}\right)^t,$$

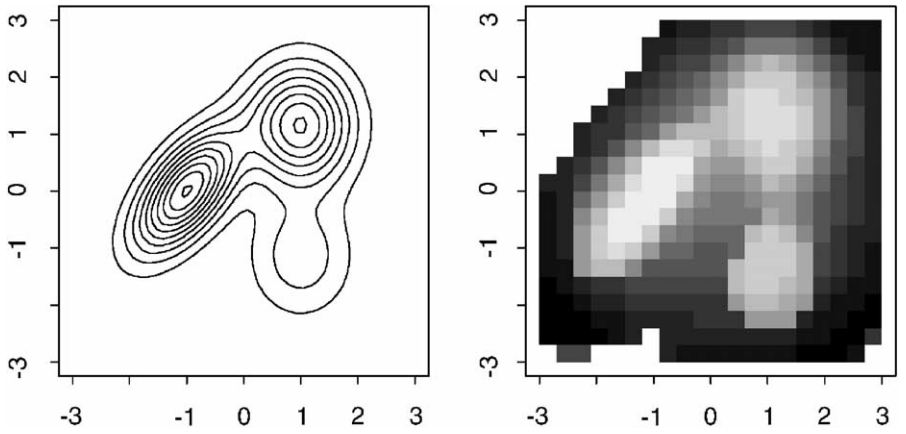

Fig. 2. Contours of normal mixture distribution (left frame) and image plot of the optimal bandwidth function (right frame) for the binned sample-point estimator ($m = 20$). Lighter gray scale indicate smaller bandwidths.

Table 3
MISE values for fixed and sample-point ($m = 20$) estimators using a normal kernel and a trimodal normal mixture. Sample size is $n = 100$

Estimator		MISE
Fixed	$\mathbf{H} \in \mathscr{H}_1$	0.00940
	$\mathbf{H} \in \mathscr{H}_3$	0.00864
Sample-point	$\mathbf{H} \in \mathscr{H}_1$	0.00561
	$\mathbf{H} \in \mathscr{H}_3$	0.00330

$$\Sigma_1 = \begin{bmatrix} (\frac{3}{5})^2 & \frac{3}{5}\frac{3}{5}\frac{7}{10} \\ \frac{3}{5}\frac{3}{5}\frac{7}{10} & (\frac{7}{10})^2 \end{bmatrix}, \quad \Sigma_2 = \begin{bmatrix} (\frac{3}{5})^2 & 0 \\ 0 & (\frac{7}{10})^2 \end{bmatrix}, \quad \Sigma_3 = \begin{bmatrix} (\frac{3}{5})^2 & 0 \\ 0 & (\frac{7}{10})^2 \end{bmatrix}.$$

Contours of this mixture are shown in the left frame of Fig. 2. This density poses problems for the fixed bandwidth kernel, even with unrestricted orientation because the covariance structure of the entire density is not far from the identity. The modes, however, have varying heights and different covariance structures.

Table 3 shows a comparison of the MISE for a fixed kernel estimator with a smoothing matrix in $\mathscr{H}_1$ and a unrestrictive smoothing matrix in $\mathscr{H}_3$. The sample-point estimator with $m = 20$ bins per dimension is also considered for the two smoothing matrix structures. The results for the two parameterizations of the fixed bandwidth estimator are not terribly different with a slight improvement going to the unrestricted smoothing matrix. However, allowing the kernels to vary can lead to significant reduction in the MISE. Even just allowing the size of the kernels to vary can give significant gains with the restricted sample-point estimator giving a 35% reduction in the MISE. An image plot of the optimal bandwidth function is shown in the right frame of Fig. 2 where each square in the image plot represents a bin. The log of the bandwidth is indicated through gray-scale with lighter shades indicating smaller

bandwidths. It is clear from that the size of the kernels mirrors the shape of the density. Also note that the kernels are not adapting simply based on heights alone. Curvature is also taken into account as the kernels near the left mode (first component of the mixture) are much smaller than those near the top right mode (second component of the mixture) despite the similar heights of these modes. Finally, the unrestricted sample-point estimator yields even more promising results and has a MISE that is 62% smaller than the unrestricted fixed estimator.

5. Practical algorithms

The previous section shows that varying the bandwidth in some fashion can lead to substantial theoretical gains. Designing a practical, databased algorithm to realize these gains in practice is a difficult task. There are more issues to consider than in the univariate case, not least of these is the parameterization of the kernel as well estimating the bandwidth function.

For example, the balloon estimator is attractive for a variety of reasons. It has tremendous potential when good pointwise estimates of the density in the tails are required. However, the orientation of the balloon kernel leading to a higher-order bias requires knowledge of the fourth-order derivatives. A pilot estimate could be used to calibrate the kernels, as suggested by Terrell and Scott (1992), but estimating fourth-order derivatives is generally perceived as harder than estimating the density itself and the gains from adaptivity would likely be overwhelmed from the error in estimating the derivatives.

For the sample-point estimator, there are currently no asymptotic approximations to the MISE that would allow a plug-in style bandwidth selector. This is due in part to the failure of the traditional asymptotic approximations to model the behavior of the optimal bandwidth function (Sain and Scott, 1996). However, least-squares cross-validation offers a method that is an unbiased estimate of the MISE and does not rely on an asymptotic approximation. Sain and Scott (1996) shows cross-validation to perform well in the univariate setting at estimating the optimal bandwidth function. Hence, it will be used as a test case to illustrate some of the difficulties that may be encountered.

Least-squares cross-validation, also referred to as unbiased cross-validation (UCV), was introduced by Rudemo (1982) and Bowman (1984) and is an approximation of the integrated squared error. On average, this criterion is equal to the MISE less a constant; hence, the term unbiased cross-validation. In this setting, the UCV function is given as

$$\mathrm{UCV} = R(\hat{f}) - \frac{2}{n}\sum_{i=1}^{n} \hat{f}_{-i}(\mathbf{x}_i),$$

where

$$\hat{f}_{-i}(\mathbf{x}_i) = \frac{1}{n-1}\sum_{j=1}^{m} n_{ij}^{*} K_{\mathbf{H}_j}(\mathbf{x}_i - \mathbf{t}_j)$$

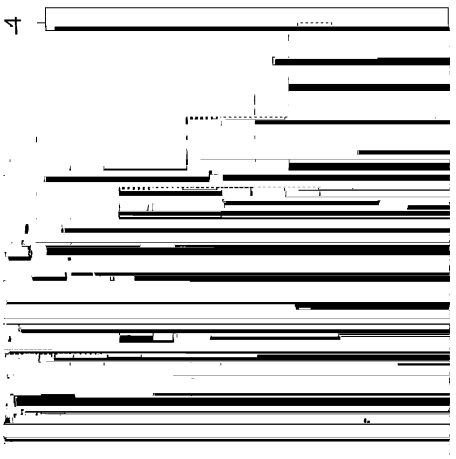

Fig. 3. Scatterplot of log cholesterol and log triglyceride levels for $n = 320$ males with heart disease. The grid indicates the mesh used for the sample-point estimator. The data have been scaled to have zero mean and unit variance in each dimension.

and

$$n_{ij}^{*} = \begin{cases} n_j - 1 & \text{if } \mathbf{x}_i \in B_j, \\ n_j & \text{otherwise.} \end{cases}$$

Note that this is not a fully binned version of UCV because the actual data are used in the estimation of the cross-product. As with the expression for the MISE, the bin counts are computed and then the criterion minimized over the space of the smoothing matrices, $\mathbf{H}_j,\ j = 1,\ldots,m$.

5.1. An example

As an illustration, consider the lipid data of Scott et al. (1978). These data consist of measurements of cholesterol and triglycerides for 320 men diagnosed with coronary artery disease. The original paper showed that the data were bimodal, indicating an increased risk for heart disease associated with increased cholesterol level. Further study of the data reveals the potential of a third mode. Fig. 3 shows a scatterplot of the data with the mesh used to bin the data for the sample-point estimator shown by the dotted lines. An 11×11 mesh was used; only the bins containing data are indicated on the plot. Note that the data were scaled to have zero mean and unit variance in each dimension.

Fig. 4 shows perspective and contour plots for a fixed bandwidth estimator using a smoothing matrix in $\mathscr{H}_1$. The smoothing parameter was selected by cross-validation. The estimate shows evidence of two distinct modes, while the third mode is marginal. There is also a considerable amount of noise exhibited by spurious modes in the tails.

The bandwidths for the sample-point estimate with smoothing matrices restricted to $\mathscr{H}_1$ are shown in Fig. 5. The observed pattern is consistent with intuition, having larger bandwidths in the tails and smaller near the modes. The estimate is shown

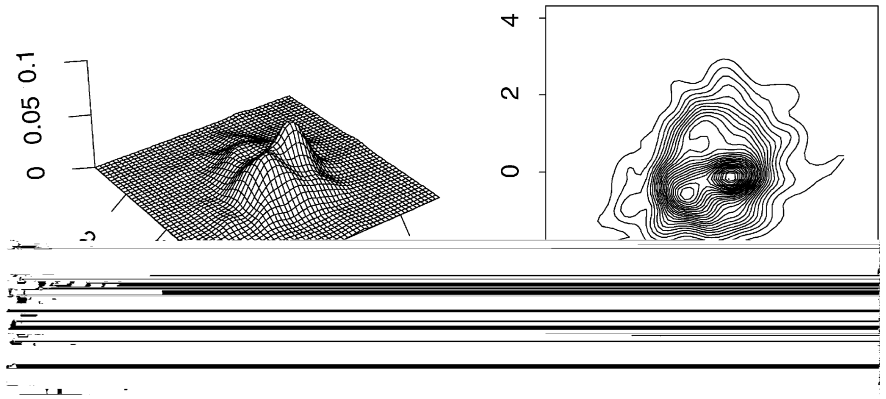

Fig. 4. Perspective and contour plots of the fixed bandwidth ($\mathscr{H}_1$) estimate for the lipid data.

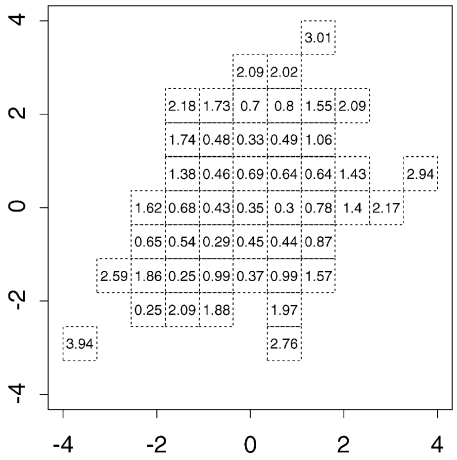

Fig. 5. Estimated bandwidths for the restricted sample-point estimate.

in Fig. 6 and initially appears oversmoothed. However, the modes are actually more pronounced (higher) and there are three clear modes in the adaptive estimate. There is also little evidence of the spurious modes in the tails of the distribution. The resulting estimator achieves the goals of the adaptive procedure, namely to correctly emphasize features (both heights and positions of modes) while minimizing noise and spurious modes.

Fig. 7 shows the sample-point estimate using unrestricted kernels ($\mathscr{H}_3$) chosen by UCV. Unfortunately, this estimate does not reflect the promise of the MISE results in the previous section. The estimate is quite rough with many spurious modes. To determine the cause, consider an examination of the estimated kernels. Those kernels near the modes are shown in Fig. 8. Here, the source of the difficulties can be found. Many of the kernels appear close to degenerate and most seem to have wildly

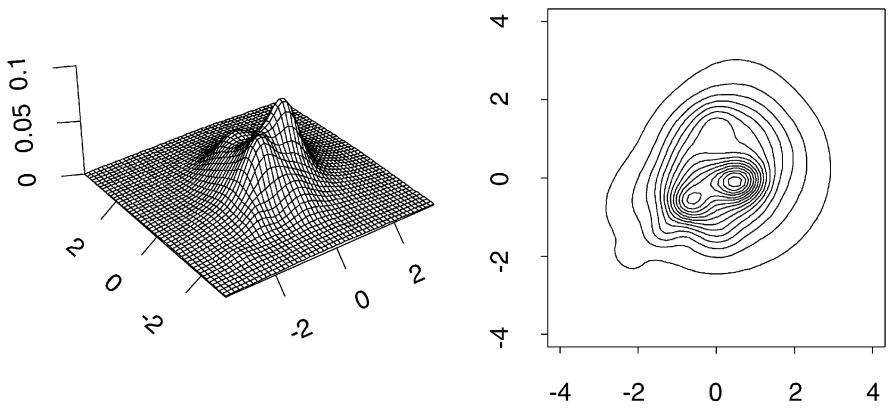

Fig. 6. Perspective and contour plots of the sample-point estimate ($\mathscr{H}_1$) for the lipid data.

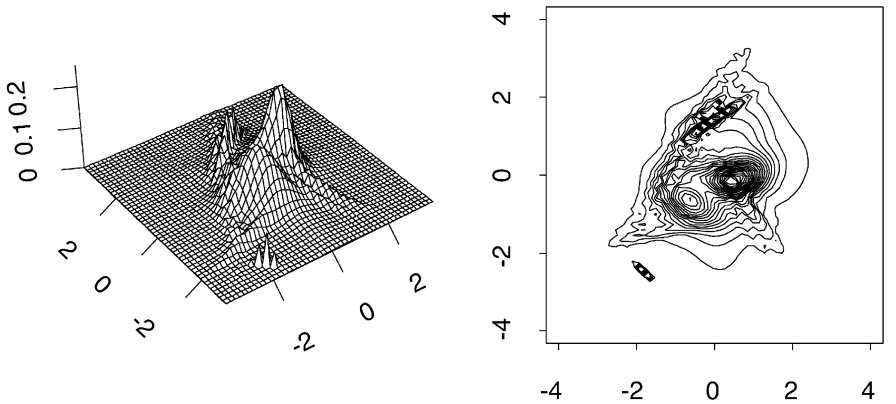

Fig. 7. Perspective and contour plots of the estimated density using a fully parameterized bandwidth matrix ($\mathscr{H}_3$).

varying orientations. Fig. 9 shows that the optimal kernels in the tails are highly elliptical and, with the scarcity of data in the tails, makes estimation difficult. This may be an artifact of the well-established aggressive behavior of cross-validation to yield highly variable bandwidths that are often much smaller than optimal. Unfortunately, there are just too many parameters to estimate, in this case $48 \times 3 = 144$, for such a moderate sample size. Clearly, attempting to estimate the fully parameterized smoothing matrices will require much more data or a more stable procedure.

5.2. Two simulated examples

To explore the behavior of the sample-point estimator further, 100 samples of size 500 were generated from two distributions, a skewed distribution labeled "C" in Wand and Jones (1993) and constructed from a mixture of three normals with

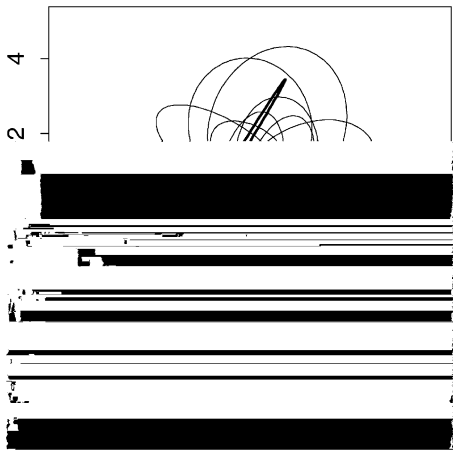

Fig. 8. Estimated kernels for fully parameterized bandwidth matrix. Kernels only shown for bins near the primary modes.

parameters

$$w_1 = 1/5, \quad w_2 = 1/5, \quad w_3 = 3/5,$$
$$\mu_1 = (0,0)^t, \quad \mu_2 = (1/2,1/2)^t, \quad \mu_3 = (13/12,13/12)^t,$$
$$\Sigma_1 = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}, \quad \Sigma_2 = \begin{bmatrix} (\frac{2}{3})^2 & 0 \\ 0 & (\frac{2}{3})^2 \end{bmatrix}, \quad \Sigma_3 = \begin{bmatrix} (\frac{5}{9})^2 & 0 \\ 0 & (\frac{5}{9})^2 \end{bmatrix}$$

and the trimodal mixture discussed previously. Both are good examples of densities that are difficult to estimate. Roughly 13 bins were placed over the region $(-4,4) \times (-4,4)$ for the sample-point estimator. Bandwidths for the product kernel estimator ($\mathscr{H}_2$) were computed using cross-validation (Sain et al., 1994) and plug-in (Wand and Jones, 1994).

The performance of the estimators was measured by a number of different criteria. First, the integrated squared error (ISE) was computed for each replication. Averages and standard deviations are shown in Table 4. For both densities, there are not dramatic differences in terms of the integrated squared error. The sample-point estimator holds a slight advantage for the skewed density while the fixed bandwidth estimators are slightly better for the trimodal density. Interestingly, the sample-point estimator has better consistency (less variability) than either of the fixed bandwidth approaches.

At first it seems disconcerting that the sample-point estimator is not clearly superior in practice, at least with respect to the squared error. Much of this is attributable to cross-validation. The increased variability associated with procedure forces a smaller number of bins than what is ideal. This keeps the variability of the estimated bandwidths down, but does not give quite enough flexibility to achieve the full benefit of adaptivity. Other approaches to estimating optimal bandwidths are currently being investigated.

While the ISE results are important, another view of performance of a density estimator is how well the estimate recovers the true structure of the underlying

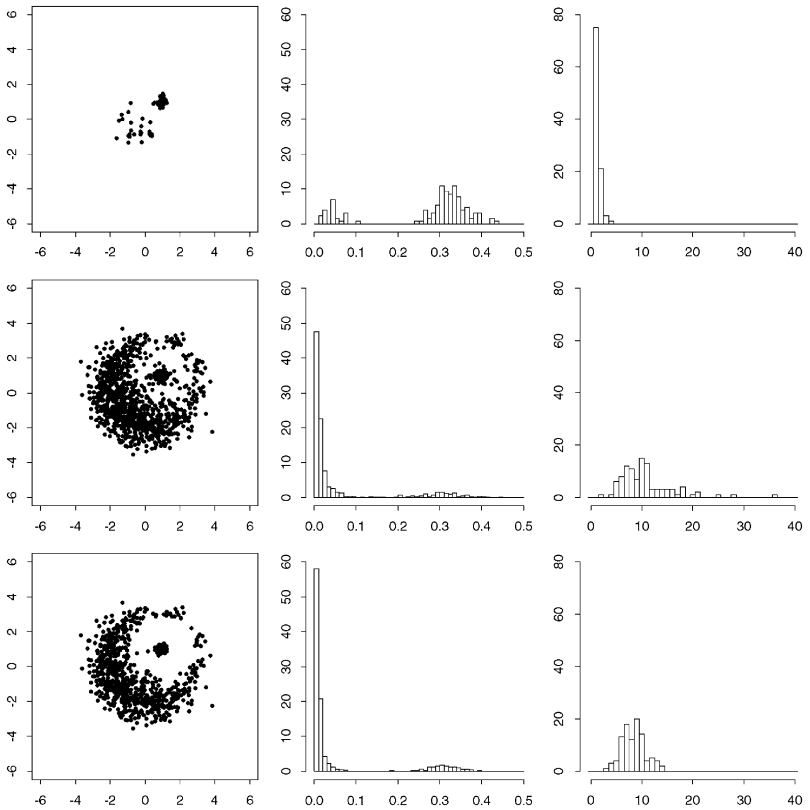

Fig. 9. Simulation results for the skewed distribution. The first row is for the sample-point estimator and the second and third rows are for the product kernel estimator with bandwidths estimated by UCV and plug-in, respectively. The left column shows locations of sample modes, the middle column shows the heights of the sample modes, and the right column shows the number of sample modes.

Table 4
Average/standard deviation of ISEs for the simulated examples. Values are written $\times 10^{-3}$

Distribution	Sample-point	UCV	Plug-in
Skewed	4.11/1.24	4.73/1.61	4.13/1.49
Trimodal	4.28/0.796	3.94/1.08	3.75/0.879

density. Figs. 9 and 10 address this issue by focusing on how well the estimates recover the primary modes of the density.

Scatterplots of the locations of the sample modes found in each replication are shown in the first column of both figures. All of the estimates seem to find the correct primary modes as evidence by the clear clusters near the true mode locations, but the fixed bandwidth procedures have a substantial number of sample modes in the tails. In fact, the last column in each plot (histograms of the number of modes in

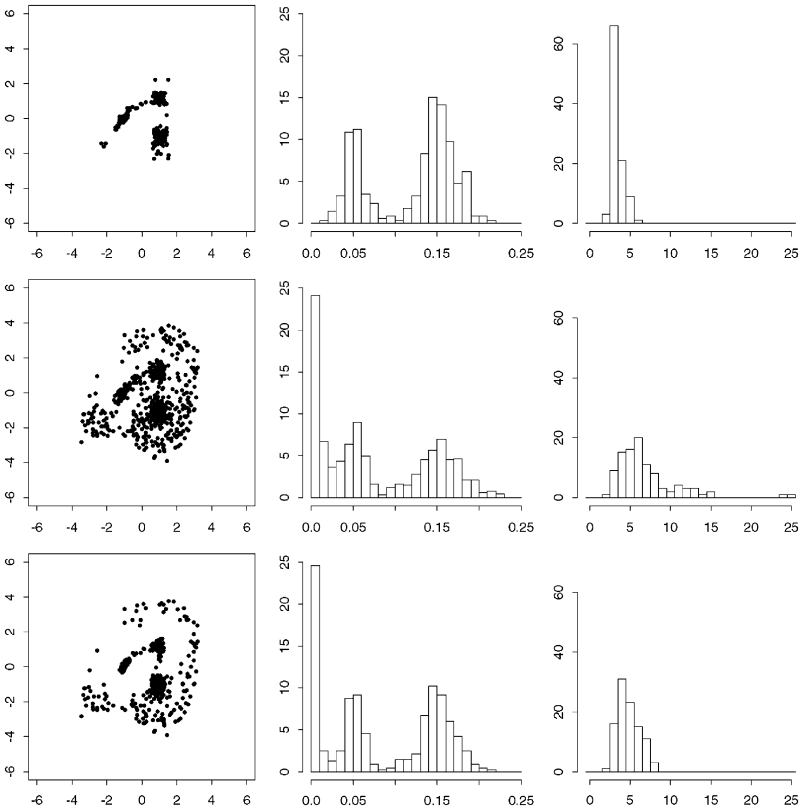

Fig. 10. Simulation results for the trimodal distribution.

each replication) shows that the fixed bandwidth procedures routinely give too many modes leaving the analyst to wonder which sample modes are real and which are spurious.

This phenomenon might lead one to conclude that the variable bandwidth procedure is simply oversmoothing. However, the middle column of plots shows histograms of the heights of the modes found in the simulation study. For the skewed distribution, the true height of the mode is 0.355. All of the estimators are biased downward (as expected), although the sample-point estimator has far less bias. For the trimodal distribution, the heights are 0.054, 0.162 and 0.203. The sample-point estimator does a much better job of distinguishing the difference in heights between the two primary peaks.

5.3. Another approach

Optimization over a large number of bins can be difficult. One way to reduce the dimensionality of the problem is based on the notion that the fixed bandwidth procedure is adequate for unimodal distributions. Hence, a pilot estimate could be used to calibrate the bandwidth function, using the same bandwidth for each data

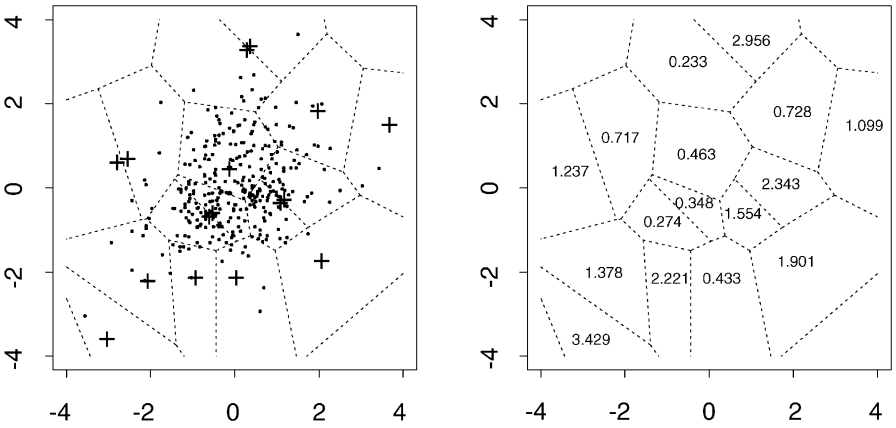

Fig. 11. Left plot: modes and partitions for calibrating the smoothing matrix. Right plot: estimated bandwidths based on the partitions.

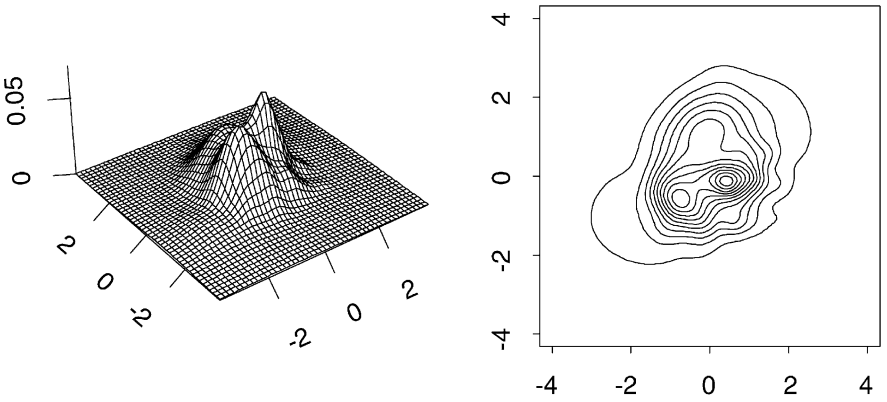

Fig. 12. Perspective and contour plots of the estimated density using the bandwidth function calibrated by the pilot estimate.

point near a sample mode. An example is shown in Fig. 11. Here the fixed bandwidth estimate from Fig. 4 is used to partition the data. There are 16 sample modes in the estimate. Each data point is assigned to the nearest mode and each mode is assigned a separate smoothing parameter. This has the effect of defining a piecewise constant bandwidth function. Only now the mesh is not equally spaced.

Fig. 11 shows the estimated bandwidths using the restricted smoothing matrices ($\mathscr{H}_1$). The behavior is as expected with smaller bandwidths near dominant modes and larger bandwidths out in the tails. The estimate using these bandwidths is plotted in Fig. 12 and it shows the two modes clearly and the third mode is apparent. Most of the variability in the tails is also diminished. While not as good as the sample-point estimate shown in Fig. 6, this estimate is still an improvement over the fixed bandwidth estimate, particularly in the tails where the number of spurious modes is reduced. Estimates using more general parameterizations of the smoothing

matrices (not shown) suffer from the same problems as the estimates in Figs. 6 and 8, although not to the same degree, as data are sparse in some of the bins. One interpretation of these result is that improvement can be gained even with a more restrictive bandwidth function that puts the adaptivity where it may be needed the most, that is out in the tails.

6. Conclusions

There are two primary conclusions to be made from this work. First, locally adaptive density estimators that allow smoothing to vary in some fashion can lead to substantial gains over fixed bandwidths for low to moderate dimensional density estimation. The results are even more promising when the underlying density is complex, exhibiting multiple modes with differences in scale and orientation.

Second, it is possible to achieve gains in practice, but it is not necessarily easy, and overparameterizing the smoothing matrices is a concern. In the fixed bandwidth case, theory suggests that kernel density estimators using smoothing matrices in $\mathscr{H}_3$ are generally superior to other parameterizations. However, in practice, difficulties in estimating the full smoothing matrix lead practitioners to use kernels in $\mathscr{H}_2$. In the variable bandwidth case, a similar finding arises from this work. Fully parameterized smoothing matrices are theoretically superior, but they are difficult to estimate. In practice, a variable bandwidth estimator using smoothing matrices in $\mathscr{H}_1$ gives sufficient flexibility to model complex densities while still improving upon fixed bandwidth estimators.

More work is required in understanding the complex theory behind locally adaptive methods and their connections to mixture models. Inspired by the desire to parametrically model data that appears to be generated from several subpopulations, mixture models have the surprising ability to model distributions whose components are not necessarily in the parametric family of the components of the mixture (see Sain et al. (1999) for an example in practice, among others). There is other work that is also exploring the connection between kernels and mixtures, for example Priebe (1994) and Marchette et al. (1996).

Additionally, more work is needed in building practical algorithms. Least-squares cross-validation holds a lot of promise, but it certainly has problems as well. Recognizing that the binned kernel estimator is similar to a mixture model with prespecified means and mixing parameters, the expectation-maximization (EM) algorithm is an interesting alternative to cross-validation and should be studied further (see McLachlan and Peel (2000) for an overview of the EM for mixtures). However, while constrained versions of the EM algorithm exist, it is generally not immune to some of the problems with degenerate and non-singular covariance matrices mentioned here.

This research is blurring the lines between parametric and nonparametric density estimation. What has been proposed here is certainly inspired by kernel estimators and motivated by a desire to learn more about variable bandwidth kernel estimators and applications in higher dimensions. Kernel estimators already possess a great flexibility, and locally adaptive kernel estimators add significantly to an already rich class.

Finally, much of the work done here required some sort of numerical optimization. Most of this optimization was done in S-Plus using the function nlminb (Statistical Sciences, 1995) which is based on subroutines from the PORT Mathematical Subroutine Library (AT& T Bell Laboratories, 1984). In some cases the gradient was also computed and used in the optimization which dramatically improved performance.

References

AT& T Bell Laboratories, 1984. PORT Mathematical Subroutine Library Manual.

Abramson, I., 1982a. On bandwidth variation in kernel estimates—a square root law. Ann. Statist. 10, 1217–1223.

Abramson, I., 1982b. Arbitrariness of the pilot estimator in adaptive kernel methods. J. Multivariate Anal. 12, 562–567.

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

Breiman, L., Meisel, W., Purcell, E., 1977. Variable kernel estimates of multivariate densities. Technometrics 19, 135–144.

Chaudhuri, P., Marron, J.S., 1999. SiZer for exploration of structures in curves. J. Am. Statist. Assoc. 94, 269–276.

Devroye, L., Lugosi, T., 2000. Variable kernel estimates: on the impossibility of tuning the parameters. In: Giné, E., Mason, D. (Eds.), High-Dimensional Probability. Springer, New York.

Friedman, J.H., Stuetzle, W., 1981. Projection pursuit regression. J. Am. Statist. Assoc. 76, 817–823.

Hall, P., 1982. The influence of rounding errors on some nonparametric estimators of a density and its derivatives. SIAM J. Appl. Math. 42, 390–399.

Hall, P., 1983. On near neighbor estimates of a multivariate density. J. Multivariate Anal. 13, 24–39.

Hall, P., 1992. On global properties of variable bandwidth density estimators. Ann. Statist. 20, 762–778.

Hall, P., Wand, M.P., 1996. On the accuracy of binned kernel density estimators. J. Multivariate Anal. 56, 165–184.

Hall, P., Marron, J.S., 1988. Variable window width kernel estimates. Probability Theory and Related Fields 80, 37–49.

Hall, P., Hu, T.C., Marron, J.S., 1994. Improved variable window kernel estimates of probability densities. Ann. Statist. 23, 1–10.

Härdle, W.K., Scott, D.W., 1992. Smoothing by weighted averaging of rounded points. Comput. Statist. 7, 97–128.

Hazelton, M.L., 1998. Bias annihilating bandwidths for local density estimates. Statist. Probab. Lett. 38, 305–309.

Jones, M.C., 1990. Variable kernel density estimates and variable kernel density estimates. Austr. J. Statist. 32, 361–371.

Loftsgaarden, D.O., Quesenberry, C.P., 1965. A nonparametric estimate of a multivariate density function. Ann. Math. Statist. 36, 1049–1051.

McKay, I.J., 1993. A note on the bias reduction in variable kernel density estimates. Can. J. Statist. 21, 367–375.

McLachlan, G., Peel, D., 2000. Finite Mixture Models. Wiley, New York.

Mack, Y., Rosenblatt, M., 1979. Multivariate K-nearest neighbor density estimates. J. Multivariate Anal. 9, 1–15.

Marchette, D.J., Priebe, C.E., Rogers, G.W., Solka, J.L., 1996. Filtered kernel density estimation. Comput. Statist. 11, 95–112.

Marron, J.S., Wand, M.P., 1992. Exact mean integrated squared error. Ann. Statist. 20, 712–736.

Minnotte, M.C., Scott, D.W., 1993. The mode tree: a tool for visualization of nonparametric density features. J. Comput. Graph. Statist. 2, 51–68.

Priebe, C.E., 1994. Adaptive mixtures. J. Am. Statist. Assoc. 89, 796–806.

Rudemo, M., 1982. Empirical choice of histogram and kernel density estimators. Scand. J. Statist. 9, 65–78.
Sain, S.R., 1994. Adaptive kernel density estimation. Unpublished Ph.D. Dissertation, Department of Statistics, Rice University.
Sain, S.R., Baggerly, K.A., Scott, D.W., 1994. Cross-validation of multivariate densities. J. Am. Statist. Assoc. 89, 807–817.
Sain, S.R., Gray, H.L., Woodward, W.A., Fisk, M.D., 1999. Outlier detection from a mixture distribution when training data are unlabeled. Bull. Seismol. Soc. Am. 89, 294–304.
Sain, S.R., Scott, D.W., 1996. On locally adaptive density estimation. J. Am. Statist. Assoc. 91, 1525–1534.
Sain, S.R., Scott, D.W., 2001. Zero-bias locally adaptive density estimators. Scand. J. Statist., to appear.
Scott, D.W., 1992. Multivariate density estimation: Theory, Practice, and Visualization. Wiley, New York.
Scott, D.W., Gotto, A.M., Cole, J.S., Gorry, G.A., 1978. Plasma lipids as collateral risk factors in coronary heart disease—a study of 371 males with chest pain. J. Chronic Dis. 31, 337–345.
Scott, D.W., Sheather, S.J., 1985. Kernel density estimation with binned data. Communications in Statistics. Theory and Methods 14, 1353–1359.
Scott, D.W., Wand, M.P., 1991. Feasibility of multivariate density estimates. Biometrika 78, 197–205.
Sheather, S.J., Jones, M.C., 1991. A reliable data-based bandwidth selection method for kernel density estimation. J. R. Statist. Soc. Ser. B 53, 683–690.
Silverman, B.W., 1982. Kernel density estimation using the fast fourier transform. Appl. Statist. 31, 93–97.
Silverman, B.W., 1986. Density Estimation for Statistics and Data Analysis. Chapman & Hall, London.
Statistical Sciences, ., 1995. S-PLUS Guide to Statistics and Mathematical Analysis. StatSci, a division of MathSoft, Inc., Seattle.
Terrell, G.R., Scott, D.W., 1992. Variable kernel density estimation. Ann. Statist. 20, 1236–1265.
Tukey, P.A., Tukey, J.W., 1981. Data-driven view selection: agglomeration and sharpening. In: Burnett, V. (Ed.), Interpreting Multivariate Data. Wiley, Chichester.
Wand, M.P., 1994. Fast computation of multivariate kernel estimators. J. Comput. Graph. Statist. 3, 433–445.
Wand, M.P., Jones, M.C., 1993. Comparison of smoothing parameterizations in bivariate kernel density estimation. JASA 88, 520–528.
Wand, M.P., Jones, M.C., 1994. Multivariate plug-in bandwidth selection. Comput. Statist. 9, 97–116.
Wand, M.P., Jones, M.C., 1995. Kernel Smoothing. Chapman & Hall, London.
Worton, B.J., 1989. Optimal smoothing parameters for multivariate fixed and adaptive kernel methods. J. Statist. Comput. Simul. 32, 45–57.