Statistic Hydrology and Hydraulics 12 (1998) 33–52 © Springer-Verlag 1998

# Kernel bandwidth selection for a first order nonparametric streamflow simulation model

**A. Sharma, U. Lall, D.G. Tarboton**



**Abstract** A new approach for streamflow simulation using nonparametric methods was described in a recent publication (Sharma et al. 1997). Use of nonparametric methods has the advantage that they avoid the issue of selecting a probability distribution and can represent nonlinear features, such as asymmetry and bimodality that hitherto were difficult to represent, in the probability structure of hydrologic variables such as streamflow and precipitation. The nonparametric method used was kernel density estimation, which requires the selection of bandwidth (smoothing) parameters. This study documents some of the tests that were conduced to evaluate the performance of bandwidth estimation methods for kernel density estimation. Issues related to selection of optimal smoothing parameters for kernel density estimation with small samples (200 or fewer data points) are examined. Both reference to a Gaussian density and data based specifications are applied to estimate bandwidths for samples from bivariate normal mixture densities. The three data based methods studied are Maximum Likelihood Cross Validation (MLCV), Least Square Cross Validation (LSCV) and Biased Cross Validation (BCV2). Modifications for estimating optimal local bandwidths using MLCV and LSCV are also examined. We found that the use of local bandwidths does not necessarily improve the density estimate with small samples. Of the global bandwidth estimators compared, we found that MLCV and LSCV are better because they show lower variability and higher accuracy while Biased Cross Validation suffers from multiple optimal bandwidths for samples from strongly bimodal densities. These results, of particular interest in stochastic hydrology where small samples are common, may have importance in other applications of nonparametric density estimation methods with similar sample sizes and distribution shapes.

**Keywords:** Streamflow, simulation, nonparametric

*Received: November 12, 1997*

A. Sharma
Department of Water Engineering, School of Civil
and Environmental Engineering, University of New South Wales,
Sydney, NSW 2051, Australia

U. Lall, D.G. Tarboton
Utah Water Research Laboratory, Utah State University, Logan,
UT 84322–8200

*Correspondence to:* U. Lall,
Fax (1) 435 797 3663, e-mail: ulall@cc.k.usu.edu.

## 1
## Introduction

Uncertainty in hydrologic inputs is of major concern in the planning, management and operation of water resources systems. One way to account for such uncertainties is to use carefully synthesized inflow sequences to properly plan reservoir operations and system expansions to meet future demands. Synthetic streamflow simulation has been a well researched area in hydrology, with models available for simulating flows at many time scales. Annual and monthly flows have traditionally been simulated using autoregressive moving average type of models (Box and Jenkins 1976). Such models characterize streamflow by an assumed probability distribution which is estimated based on the first two or three moments (mean, covariance, skewness) of the historical record. These moments and the assumed probability distribution are accurately reproduced in the simulated flow sequences.



The authors have recently proposed nonparametric approaches for streamflow simulation (Sharma et al. 1997) and disaggregation (Tarboton et al. 1997). These approaches sidestep the issue of assuming a probability distribution, and use the entire historical sample to estimate the joint and conditional probability densities needed for simulation. This ensures that simulations are similar to the historical record in more aspects than just a few sample moments. Features such as asymmetry or bimodality in the probability density function and nonlinearity or state dependence in the conditional mean are naturally modelled.

The nonparametric method used in the above mentioned studies is kernel density estimation. An important issue regarding use of kernel methods is the selection of an optimal smoothing parameter called bandwidth. Bandwidth estimation for kernel density estimation and regression has been widely studied in the last decade. While several asymptotically optimal methods exist for bandwidth estimation, limited investigations of their small sample performance are available.

The aim of this paper is primarily to share some of our experience with bandwidth selection procedures for small samples (samples less than 200 data points) in a bivariate setting. The choice of the smoothing parameter is important for accurate estimation of joint and conditional probability densities needed in the nonparametric simulation and disaggregation approaches. This study evaluates the performance of bandwidth estimation methods for small samples as are usually encountered in stochastic hydrology. An overview of the three bandwidth selection procedures studied is presented from a practitioners perspective. Their performance is evaluated with samples drawn from bivariate normal mixture densities. The efficiency of local bandwidth based estimators with estimators using a single or global smoothing parameter is also compared. The Mean Integrated Square Error (MISE) between the estimated and parent densities is used as a criterion to evaluate the performance of the bandwidth estimation method.

An introduction to kernel density estimation and a summary of the nonparametric streamflow simulation models is presented first. Kernel density estimation in a multivariate setting and a method for estimation of local bandwidths is described next. Section 4 then describes the bandwidth estimation procedures compared in this paper. Section 5 describes the numerical experiments used to evaluate the performance of the alternative bandwidth selection procedures. Results for the global bandwidth based estimators are followed by results from those that use local bandwidths.

## 2
## Nonparametric approaches for streamflow simulation

A parametric probability density function (PDF) is one that is fully defined by a finite set of parameters. One example is the Gaussian PDF which is completely specified by a location and scale parameter. On the other hand, a nonparametric probability density estimate is based on the entire sample, rather than a few sample moments. A nonparametric estimator is asymptotically local, the effect of distant data points on the estimated probability density vanishing with increasing sample size. Hence, a nonparametric estimator is consistent, whereas a misspecified parametric PDF has a bias that does not reduce with increasing sample size. This has led to the increased use of nonparametric methods in several areas of hydrology including frequency analysis of extreme events (Lall et al. 1993) and hydrologic time series simulation (Sharma et al. 1997; Tarboton et al. 1997; Lall and Sharma 1996). Readers are referred to Lall (1995) for an overview of nonparametric applications in hydrology.



A commonly used nonparametric probability density estimator is the histogram. It suffers from the discrete nature of the estimated probabilities and the specification of the location and width of individual bins. Kernel density estimation is an extension of the histogram and provides a smooth, continuous probability density estimate that is asymptotically consistent. A univariate kernel density estimator is specified as:

$$\hat{f}(x) = \frac{1}{nh}\sum_{i=1}^{n} K\left(\frac{x - x_i}{h}\right) \tag{1}$$

where $K()$ is a kernel function centered at each data point $(x_i)$, and h is the bandwidth which specifies the spread of individual kernel functions. A kernel must be a legitimate PDF for the probability density estimate that integrates to one. The Gaussian kernel function, a popular and practical choice, has been used in this study:

$$K(x) = \frac{1}{(2\pi)^{1/2}} \exp(-x^2/2) \tag{2}$$

The kernel density estimate in (1) can be extended to multiple dimensions, the estimate now being composed of the sum of multivariate kernel functions centered at individual multivariate data points. A bivariate kernel density estimate of variables $x_t$ and $x_{t-1}$ is shown in Figure 1. Also shown are bivariate kernels centered at two sample data points. The bivariate probability density is the cumulative effect of individual bivariate kernels associated with each data point.

The nonparametric order one streamflow simulation model, NP1, (see Sharma et al. 1997) simulates streamflow $x_t$ conditional to $x_{t-1}$. Nonparametric estimates of the bivariate probability density $f(x_t, x_{t-1})$ and the conditional probability density $f(x_t|x_{t-1})$ are used. The conditional density is estimated as a slice through the bivariate probability density at the conditioning value $x_{t-1}$, as illustrated in fig. 1. This is composed of a sum of slices through the individual kernels that form the bivariate density estimate. As Gaussian bivariate kernels are used, the slices are weighted Gaussian PDF's with weights indicating their relative contribution to the conditional probability density estimate. Simulation proceeds by sampling from the conditional density estimate, or, effectively from one of the kernel slices that form the conditional density.

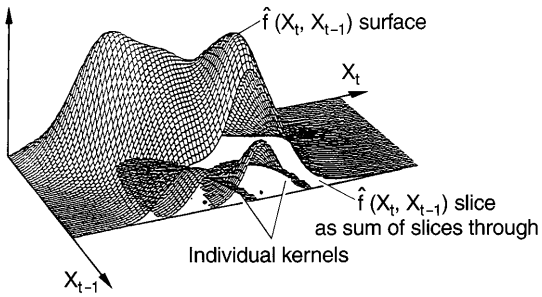


**Fig. 1.** Illustration of the nonparametric streamflow simulation model of *Sharma et al.* (1997)

The nonparametric disaggregation model, NPD (see Tarboton et al. 1997) disaggregates an aggregate streamflow value $z$ (such as the annual streamflow) into $d$ individual disaggregated components $x_1, x_2, \ldots, x_d$. This involves formulation of the nonparametric conditional probability $f(x_1, x_2, \ldots, x_d | x_1 + x_2 + \ldots + x_d = z)$. This can be expressed as the sum of slices through multivariate Gaussian kernels, and $x_1, x_2, \ldots, x_d$ can be simulated using a similar procedure as in the NPl model.

Accurate estimation of the joint and conditional probability densities is important for simulation of representative streamflow sequences in both the nonparametric models discussed above. This requires estimation of the optimal bandwidth that specifies the spread or variability of the kernels that form the joint and conditional probability densities. While a large bandwidth results in an oversmoothed probability density, a small bandwidth results in a rough density estimate. Several methods have been proposed for estimation of the optimal kernel bandwidth. These are described in the next section.

## 3
## Multivariate kernel density estimation

The Gaussian kernel density estimate of a $d$ dimensional probability density function $f(x)$ is written as:

$$\hat{f}(\mathbf{x}) = \frac{1}{n}\sum_{i=1}^{n} \frac{1}{(2\pi)^{d/2}\det(\mathbf{H})^{1/2}} \exp\left(-\frac{(\mathbf{x}-\mathbf{x}_i)^T \mathbf{H}^{-1}(\mathbf{x}-\mathbf{x}_i)}{2}\right) \tag{3}$$

where $n$ is the number of observed vectors $\mathbf{x}_i$ and $\mathbf{H}$ is a bandwidth matrix that must be from the class of symmetric positive definite $d \times d$ matrices (Wand and Jones 1994). The above density estimate is formed by summing Gaussian kernels with a covariance matrix $\mathbf{H}$, centered at each observation $\mathbf{x}_i$. The bandwidth matrix $\mathbf{H}$ here is analogous to the covariance matrix for a multivariate Gaussian density. It can be specified in several ways. Wand and Jones (1993) suggest three ways to parameterize $\mathbf{H}$ for the bivariate case ($d = 2$). These are

$$\mathbf{H} = h^2\mathbf{I}: h > 0 \tag{4a}$$

$$\mathbf{H} = \text{diag}\ (h_1^2, h_2^2): h_1, h_2 > 0 \tag{4b}$$

$$\mathbf{H} = \begin{bmatrix} h_1^2 & h_{12} \\ h_{12} & h_2^2 \end{bmatrix} : h_1, h_2 > 0, |h_{12}| < h_1 h_2 \qquad (4c)$$

where **I** represents the Identity matrix and $h, h_1$ etc. are elements of the bandwidth matrix **H**. The first case (eq. 4a) represents a spherical kernel (with circular contours). The addition of extra smoothing parameters in the second case (eq. 4b) makes the kernel elliptical, though aligned parallel to the coordinate axes. The last case (eq. 4c) represents an elliptical kernel aligned in a direction dictated by the cross diagonal term ($h_{12}$). The number of parameters required in (4c) can be reduced by considering the following parameterization, first proposed by Fukunaga (1972).



$$\mathbf{H} = \lambda^2 \mathbf{S} \qquad (5)$$

Here, **S** is the sample covariance matrix of the data and $\lambda^2$ prescribes the bandwidth relative to this estimate of scale. Use of this parameterization amounts to using a bandwidth $\lambda$ on a transformed sample whose covariance matrix is the identity. This procedure is called "sphering" (Fukunaga 1972) and ensures that all kernels are oriented along the principal components of the covariance matrix. Such a parameterization has significant advantages when the variables are strongly correlated. Wand and Jones (1993) demonstrate the utility of this method for bivariate Gaussian data but note that densities having multiple modes in one of the coordinate directions may be poorly estimated. We have used the estimator in (3) with the bandwidth matrix described in (5) in all our results in sections 4 and 5. This estimator was chosen primarily because it requires the optimization of only one parameter ($\lambda$) while still providing an elliptical kernel oriented as dictated by the sample correlation. In addition, the sample lag 1 correlation for some of the monthly streamflow data sets we have analyzed have ranged from 0.4–0.9 for different month pairs. Use of the sphering approach is an effective way to deal with such highly correlated data sets, even though it may be less efficient for non-Gaussian data-sets. For ease of reference, this factor ($\lambda$) is called the bandwidth in the rest of the paper.

Optimization of the bandwidth is subject to an appropriate criterion. While criteria such as the Integrated Square Error (ISE) and the Mean Integrated Square Error (MISE) give an estimate of the overall or global goodness of fit, the Mean Square Error (MSE) gives a pointwise measure of the error in the kernel density estimate. Minimization of the MSE can be used to estimate optimal pointwise or local bandwidths. Assigning individual bandwidths to each data point (a bandwidth $\lambda_i$ corresponding to data point $x_i$) is an effective specification for local bandwidths. Density estimation based on this specification can proceed using $\mathbf{H}_i = \lambda_i^2 \mathbf{S}$ instead of **H** in Equation (3). Such additional flexibility in the bandwidth may be used to restrict the smoothing imposed by the kernel function in regions of high density or high curvature, where there are a lot of data points, or where the smoothing introduces bias. Conversely, in regions of low density the bandwidth can be increased to average over more points.

Abramson (1982) proposed a method for estimating local bandwidths by considering the Taylor series expansion of the squared bias of $\hat{f}(\mathbf{x})$. The second order term in the Taylor series can be eliminated if the local bandwidth is inversely proportional to the square root of true density. Silverman's (1986, p. 101 and 105) implementation of Abramson's method involves perturbing an appropriate global or fixed bandwidth $\lambda_p$ (denoted as the pilot global bandwidth to

distinguish from $\lambda$) into a sequence of local bandwidths $\lambda_i$ at each observation $x_1$. This implementation (which we shall call the Abramson-Silverman method) can be stated as:

$$\lambda_i = \lambda_p(\hat{f_p}(\mathbf{x}_i)/g)^{-1/2} \tag{6}$$

where $\hat{f_p}(.)$ is a pilot density estimate and $g$ is the geometric mean of $\hat{f_p}(\mathbf{x}_i)$ at data points $\mathbf{x}_i$. The pilot density specifies the amount of perturbation the local bandwidths receive and may be estimated using any acceptable scheme. In results presented in section 5, a kernel density estimate using the global bandwidth $\lambda_p$ as the pilot density in (6) was used. Note that the inverse relationship of a local bandwidth with the estimated density in effect provides a higher bandwidth in regions of low density and a lower bandwidth where the density is high. Several authors (see Scott 1992, p. 187) have noted that local bandwidths need to be "clipped" or restricted to lie within certain upper and lower bounds. We chose not to clip the $\lambda_i$ due to the subjectivity introduced by a prescriptive choice for these upper and lower bounds. Use of local bandwidths for estimating the density, though computationally intensive, can result in reduced errors as demonstrated in Scott (1992, table 6.4).



## 4
## Bandwidth estimation methods

This section describes some methods for estimation of the optimal bandwidth. These are:

1) Reference to a standard distribution
2) Maximum Likelihood Cross Validation (MLCV)
3) Unbiased or Least Square Cross Validation (LSCV)
4) Biased Cross Validation (BCV2)

Each of these methods and their associated advantages/disadvantages are discussed, followed by modifications needed to estimate optimal local bandwidths using the Abramson–Silverman method in eq. (6).

### 4.1
### Gaussian reference bandwidth (GREF)

The simplest automated choice for the bandwidth $\lambda$ is the reference bandwidth. A reference bandwidth is optimal for an assumed (reference) distribution using an appropriate criterion. A Taylor series expansion of the MISE is used to develop expressions for the optimal reference bandwidth. Scott (1992, p. 131) gives an expression for the first order Taylor series approximation of the univariate MISE. This expression, using a Gaussian kernel, can be stated as:

$$\mathrm{AMISE}(h) = \frac{1}{2\sqrt{\pi}nh} + \frac{h^4}{4}R(f'') \tag{7}$$

where AMISE stands for the Asymptotic Mean Integrated Square Error (Scott 1992, p. 54), $R(g(x)) = \int g(x)^2 dx$ (in this case $g(x) = f''(x)$) and $f''(x)$ is the second derivative of the true density. Minimization of the multivariate version of (7) with the true density assumed to be Gaussian, results in the following expression for the AMISE optimal Gaussian reference bandwidth (Scott 1992, p. 152).

$$\lambda_{\mathrm{GREF}} = \left(\frac{4}{d+2}\right)^{1/(d+4)} n^{-1/(d+4)} \tag{8}$$

Here $d$ and $n$ denote the dimension and sample size respectively. Although simple, this choice suffers from the obvious disadvantage of being optimal only for a Gaussian density. It should, however, be noted that the use of a Gaussian reference bandwidth may not be very detrimental for data-sets that originate from near-Gaussian densities, as is illustrated in the results presented in section 5. The following methods avoid assuming an underlying density and instead choose bandwidth by optimizing data based estimates of likelihood or square error.



### 4.2
### Maximum likelihood cross validation (MLCV)

This method is a natural development of the idea of using likelihood to judge the goodness of fit of any statistical model. The use of MLCV to choose the bandwidth for kernel density estimation was proposed by Duin (1976) and Habbema et al. (1974). The rationale behind this method is to estimate the log-likelihood of the density at observation $\mathbf{x}_i$ based on all observations except $\mathbf{x}_i$. Averaging this log-likelihood over all observations results in the following MLCV score:

$$\mathrm{MLCV}(\mathbf{H}) = \frac{1}{n}\sum_{i=1}^{n} \log \hat{f}_{-i}(\mathbf{x}_i) \tag{9}$$

where $\hat{f}_{-i}(\mathbf{x}_i)$ denotes the density estimated from all the data points except $\mathbf{x}_i$. Using the estimator in (3), the MLCV score can be stated as:

$$\mathrm{MLCV}(\mathbf{H}) = \frac{1}{n}\sum_{i=1}^{n} \log\left(\frac{\sum_{j\neq i}^{n} \exp(-L_{ij}/2)}{(2\pi)^{d/2}(n-1)\det(\mathbf{H})^{1/2}}\right) \tag{10}$$

where

$$L_{ij} = (\mathbf{x}_i - \mathbf{x}_j)^T \mathbf{H}^{-1} (\mathbf{x}_i - \mathbf{x}_j) \tag{11}$$

and $\mathbf{H}$ is the bandwidth matrix specified in (5). Maximizing the score in (10) results in an MLCV optimal bandwidth $\lambda_{\mathrm{MLCV}}$.

For long tailed densities the MLCV criterion can lead to degenerate bandwidth choices and inconsistent density estimates where a global bandwidth is used (Silverman 1986, p. 54). Schuster (1985) recommended optimizing the MLCV function over $\mathbf{x}\varepsilon\mathbf{X}$ where $\mathbf{X}$ is an appropriate subset of the sample space that excludes the tails.

### 4.3.
### Unbiased or least square cross validation (LSCV)

This method was first proposed by Bowman (1984) and Rudemo (1982). LSCV is based on the direct minimization of the Integrated Square Error (ISE). The ISE for a multivariate density $f$ can be expanded as:

$$\mathrm{ISE} = \int (\hat{f}(\mathbf{x}) - f(\mathbf{x}))^2 d\mathbf{x} = R(\hat{f}(\mathbf{x})) - 2\int \hat{f}(\mathbf{x})\, f(\mathbf{x}) d\mathbf{x} + R(f(\mathbf{x})) \tag{12}$$

The first term in (12) depends solely on the data, bandwidth and kernel used. The last term, $R(f(\mathbf{x}))$ is independent of the bandwidth and does not need to be considered. The middle term involving the product of the true and estimated densities may be recognized as $\mathrm{E}[\hat{f}(\mathbf{X})]$ and estimated using leave one out cross validation. The LSCV criterion can then be stated as:

$$\mathrm{LSCV}(\mathbf{H}) = R(\hat{f}(\mathbf{x})) - \frac{2}{n}\sum_{i=1}^{n}\hat{f}_{-i}(\mathbf{x}_i) \tag{13}$$



Sain et al. (1994) provide an expression for LSCV in any dimension with multivariate Gaussian product kernels (Gaussian kernels with a diagonal $\mathbf{H}$ matrix). The LSCV score for a generalized $\mathbf{H}$ matrix (one with cross diagonal terms) can be stated as:

$$\mathrm{LSCV}(\mathbf{H}) = \frac{1 + \sum_{i=1}^{n}\sum_{j\neq i}\left[\frac{\exp(-L_{ij}/4)}{n} - \frac{2^{d/2+1}\exp(-L_{ij}/2)}{n-1}\right]}{n(2\sqrt{\pi})^{d}\det(\mathbf{H})^{1/2}} \tag{14}$$

where $L_{ij}$ follows the definition in (11). Minimizing the LSCV score in (14) results in an LSCV optimal bandwidth $\lambda_{\mathrm{LSCV}}$, or a bandwidth that minimizes the LSCV score in (13) or the ISE in (12).

Though unbiased, bandwidth estimated using the LSCV score function has been reported to suffer from the disadvantages of a tendency to undersmooth (Chiu 1990, Chiu 1991) and a high variance of the estimated bandwidths as compared to the BCV estimator of Scott and Terrell (1987). The higher variance corresponds to a tendency for the LSCV score function to have multiple local minima and hence a tendency to undersmooth (Chiu 1990). Solutions to this problem are suggested among others by Chiu (1990, 1991). These solutions, however, were not implemented in our analysis. Compared to the BCV2 score function that is described next, Sain et al. (1994) report that LSCV has a marginally smaller asymptotic variance when applied to estimate a univariate Gaussian density.

### 4.4
### Biased cross validation (BCV2)

Proposed by Sain et al. (1994), biased cross validation provides a data based estimate of the AMISE in (7). We shall refer to this as BCV2, the notation used by Sain et al. in their paper. The term $R(f'')$ in eq. (7) is estimated using leave one out cross validation. For a vector $\mathbf{x}$, this term can be stated as

$$R(f'') = \int f''(\mathbf{x})^2 d\mathbf{x} = \int f^{iv}(\mathbf{x}) f(\mathbf{x}) d\mathbf{x} = E[f^{iv}(\mathbf{x})] \tag{15}$$

or,

$$R(f'') \approx \frac{1}{n}\sum_{i=1}^{n}\hat{f}^{iv}_{-i}(\mathbf{x}_i) \tag{16}$$

where $\hat{f}^{iv}_{-i}(\mathbf{x}_i)$ is the fourth derivative of the kernel density estimate at $\mathbf{x}_i$ formed by leaving out data point $\mathbf{x}_i$. Sain et al. (1994) provide an expression for BCV2 using

Gaussian product kernels. Extending their result to the case of a general bandwidth matrix **H** we obtain

$$BCV2(\mathbf{H}) = \frac{1}{(2\sqrt{\pi})^d n \det(\mathbf{H})^{1/2}} + \frac{\sum_{i=1}^{n}\sum_{j\neq i}\left[L_{ij} - (2d+4)L_{ij} + d^2 + 2d\right]}{4n(n-1)\det(\mathbf{H})^{1/2}(\sqrt{2}\pi)^d}\exp(-L_{ij}/2) \tag{17}$$

with $L_{ij}$ as defined in (11). A global bandwidth $\lambda_{BCV2}$ can be optimized by minimizing the score in (17).



BCV2 suffers from the problem of being a biased estimator of the optimal bandwidth. An earlier version of biased cross validation (denoted BCV by Scott and Terrell 1987 and BCV1 by Sain et al. 1994) had the advantage of having smaller variance though being more heavily biased than BCV2. Apart from the reduction in bias, another justification for BCV2 over BCV is the relative ease with which it can be implemented in multivariate settings. Sain et al. (1994) conducted a simulation study to compare the performance of BCV2 with LSCV. They found that LSCV tends to have higher variance than BCV2 for a standard Gaussian density, although their derivations for the asymptotic standard deviations (Sain et al. 1994, table 2) indicate otherwise. Sain et al. support their results by noting that LSCV tends to undersmooth which may increase the variance unnecessarily.

### 4.5
### Estimation of optimal local bandwidths

Results in the next section use the Gaussian reference, MLCV and LSCV as criteria for estimation of local bandwidths. The Abramson–Silverman method (Equation 6) is used. Details of our implementation of the MLCV and LSCV score functions for local bandwidths are given here.

As mentioned in section 3, local bandwidths may be estimated based on a pilot density estimate (see eq. 6). We have used a kernel density estimate based on a global bandwidth $\lambda_p$ as the pilot density. Local bandwidths can then be estimated by perturbing $\lambda_p$ as given in eq. (6). These steps amount to the following algorithm:

1) Estimate the pilot density using the pilot bandwidth $\lambda_p$. Call this pilot density $\hat{f}_p(.)$.
2) Estimate the corresponding local bandwidths $\lambda_i$ using the Abramson–Silverman method in eq. (6).
3) Calculate the criterion function (MLCV or LSCV) for local bandwidths $\lambda_i$.

The only specification used here is that of the global bandwidth $\lambda_p$. Silverman's suggestion can be taken to optimize the global bandwidth $\lambda_p$ and then simply perturb it to a vector of local bandwidths. In this study the target score or optimization function was computed using the density estimates based on the local bandwidths $\lambda_i$ obtained upon perturbing the global $\lambda_p$. The optimal $\lambda_p$ is then selected as the optimizer of the score function computed using the local bandwidths. This procedure was followed in all cases except when the Gaussian reference bandwidth is used, where that bandwidth was used directly to determine corresponding local bandwidths. The global Gaussian reference bandwidth (GREF) is denoted as $\lambda_{p_{GREF}}$.

The MLCV criterion using local bandwidths is a natural extension of eq. (10).

$$\text{MLCV} = \frac{1}{n}\sum_{i=1}^{n} \log\left(\sum_{j\neq i}^{n} \frac{\exp(-U_1/2)}{(2\pi)^{d/2}(n-1)\det(\mathbf{H}_j)^{1/2}}\right) \tag{18}$$

where

$$U_1 = (\mathbf{x}_i - \mathbf{x}_j)^T \mathbf{H}_j^{-1}(\mathbf{x}_i - \mathbf{x}_j) \tag{19}$$



$\mathbf{H}_j$ equals $\lambda_j^2\mathbf{S}$ in this specification. Optimization proceeds by maximizing the score in (18) with respect to the pilot bandwidth $\lambda_p$. The MLCV optimal pilot bandwidth is denoted $\lambda_{p_{\text{MLCV}}}$.

An equivalent of the LSCV in (14) for local bandwidths is:

$$\begin{aligned}\text{LSCV} = \frac{1}{(2\sqrt{\pi})^d n}\sum_{i=1}^{n}\Bigg\{ & \frac{1}{n\det(\mathbf{H}_i)^{1/2}} \\ & + 2^{d/2}\sum_{j\neq i}\left[\frac{\exp(-U_2/2)}{n\det(\mathbf{H}_i+\mathbf{H}_j)^{1/2}}\right. \\ & \left. - \frac{2\exp(-U_1/2)}{(n-1)\det(\mathbf{H}_j)^{1/2}}\right]\Bigg\}\end{aligned} \tag{20}$$

where $U_1$ follows the definition in (17) and $U_2$ equals

$$U_2 = (\mathbf{x}_i - \mathbf{x}_j)^T(\mathbf{H}_i + \mathbf{H}_j)^{-1}(\mathbf{x}_i - \mathbf{x}_j) \tag{21}$$

As before, $\mathbf{H}_i$ equals $\lambda_i^2\mathbf{S}$ for this specification. Note that Equation (20) reduces to Equation (14) when $\mathbf{H}_i = \mathbf{H}_j = \mathbf{H}$. Optimization proceeds by minimizing Equation (20) with respect to the pilot bandwidth $\lambda_p$. The LSCV optimal pilot bandwidth is referred to as $\lambda_{p_{\text{LSCV}}}$.

Four global bandwidth selectors, GREF, MLCV, LSCV and BCV2, were discussed in this section. Extensions for local bandwidths were developed for MLCV and LSCV. Results using these procedures applied to samples from selected Gaussian mixture densities are provided in the next section.

## 5
## Application to samples from Gaussian mixture densities

Here we evaluate the performance of bandwidth estimation methods described in the previous section. These methods were tested with samples drawn from mixtures of bivariate Gaussian densities. Small samples (less than 200 data points for each bivariate sample) were considered. We chose the class of Gaussian mixture densities since exact results for the MISE and AMISE given a bandwidth matrix (or matrices if local bandwidths are used) are known (see for instance Wand and Jones 1995, p. 101–103). We used the Numerical Recipes function BRENT (Press et al. 1989, p. 283–286) to optimize the bandwidth ($\lambda$ or $\lambda_p$ in the range $\lambda_{\text{GREF}}/10$ to $2\lambda_{\text{GREF}}$. This optimization function is based on inverse parabolic interpolation that ensures convergence given an initial range in which the minimum can be

found. Several spot checks indicated that the minima found by this routine was indeed a global minima. The performance of each estimator was measured by recording the ISE between the true and estimated densities for each sample. The ISE was then averaged over the samples to get a measure of the MISE. The fact that the integral of the product of two Gaussian densities can be represented as a single Gaussian density, as reported in Wand and Jones (1993, appendix 1), reduced the ISE calculations considerably. It should be noted that since our samples are drawn from Gaussian mixture densities, each term in the ISE in (12) can be represented as a sum of Gaussian p.d.f.'s.

Results for three sample sizes (50, 100 and 200) from the four distributions listed in table 1 are evaluated. These distributions were chosen based on a list of bivariate distributions in Wand and Jones (1993, table 1). Contour plots of these p.d.f.'s are illustrated in fig. 2. Our testing procedure involved drawing 50 samples of sizes 50, 100 and 200 from each test distribution. The main conclusions from this study were found to be insensitive to the use of a greater number of samples.



**Table 1.** Description of the bivariate normal mixture densities studied

| Density | $w_1 N(\mu_1, \mu_2, \sigma_{11}^2, \sigma_{12}^2, \rho_1) + \ldots . w_k N(\mu_k, \mu_k, \sigma_{k1}^2, \sigma_{k2}^2, \rho_k)$ |
|---|---|
| (A) Standard | $N(0,\ 0,\ 1,\ 1,\ 0)$ |
| (B) Bimodal | $1/2\ N(-1,\ 0, (2/3)^2, (2/3)^2,\ 0) + 1/2\ N(1,\ 0,\ (2/3)^2,\ (2/3)^2, 0)$ |
| (C) Bimodal | $1/2\ N(1,\ -1, (2/3)^2, (2/3)^2, 7/10) + 1/2\ N(-1,\ 1,\ (2/3)^2,\ (2/3)^2, 0)$ |
| (D) Bimodal | $0.4\ N(-1.2,\ 0, (3/5)^2, (3/5)^2, 7/10) + 0.4\ N(1.2,\ 0,\ (3/5)^2,$ $(3/5)^2, 7/10) + 0.2N(0,\ 0,\ (3/5)^2,\ (3/5)^2,\ -7/10)$ |

In the notation used, N(.) refers to a bivariate Gaussian distribution with mean $(\mu_{j1}, \mu_{j2})$, variance $(\sigma_{j1}^2, \sigma_{j2}^2)$ and correlation $\rho_j$ where j ranges from 1 to k, k being the number of mixtures used. $w_1, \ldots, w_k$ denote the weights for individual Gaussian p.d.f.'s.

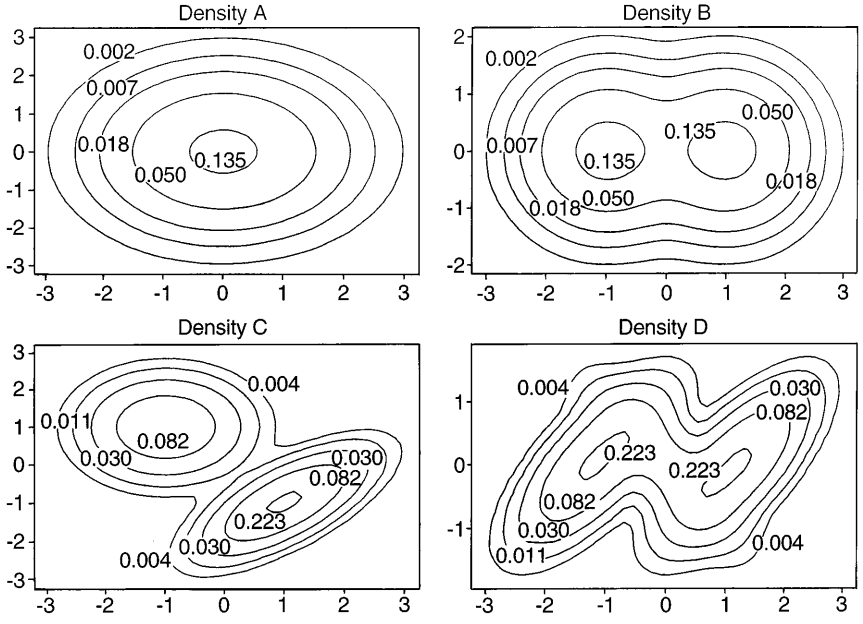


**Fig. 2.** Contour plots of the bivariate test densities listed in table 1

The Integrated Square Error (ISE) was used to evaluate the performance of various methods. Also presented is the bandwidth that minimizes the exact MISE for each of the test densities. Estimates of this exact MISE are based on relations in Wand and Jones (1993, Theorem 1). Results for global bandwidth estimators are followed by those using local bandwidths as described in section 4.5.

### 5.1
### Results using a global $\lambda$



Integrated Square Error's were computed as a function of the optimal global bandwidth for each sample. Table 2 shows the average of these ISE scores for all test densities and sample sizes. Histograms of the optimal bandwidths for each of the test densities are illustrated in figs. 3 (density A), 4 (density B), 5 (density C) and 6 (density D). Also shown are the MISE optimal bandwidth and the Gaussian reference bandwidth for each density and sample size. The bias and standard deviation of the optimal bandwidths (from MLCV, LSCV and BCV2) are presented in table 3.

As expected, the reference Gaussian bandwidth results in the minimum MISE (see table 2) for density A (the standard normal p.d.f.) for all sample sizes, with MLCV coming a close second. Higher MISE's for BCV2 and LSCV are partly due to the greater variability in optimal bandwidths from both methods (see fig. 3 and table 3). It is notable that MLCV (a method that is not based on minimizing the sum of the squared error terms) gives a lower variance and a smaller MISE than other cross validation based methods. A low bias in all methods is apparent from table 3.

Density B is bimodal with the modes aligned along one coordinate axis. Note how close the reference bandwidth lies to the MISE optimal bandwidth for samples of size 50 from this density. This reflects in the ISE scores too with the Gaussian reference bandwidth resulting in the best MISE amongst all methods. Poor performance for BCV2 is apparent and is due to the high variability in its optimal bandwidths (see fig. 4. and table 3). MLCV again gives the best results amongst all cross validation methods.

**Table 2.** Mean ISE for global bandwidth-based estimators

| Density | Method | $n = 50$ | $n = 100$ | $n = 200$ |
|---|---|---|---|---|
| | GREF | 0.0075 | 0.0048 | 0.0031 |
| (A) | MLCV | 0.0077 | 0.0050 | 0.0032 |
| | LSCV | 0.0089 | 0.0054 | 0.0038 |
| | BCV2 | 0.0083 | 0.0057 | 0.0036 |
| | GREF | 0.0109 | 0.0077 | 0.0049 |
| (B) | MLCV | 0.0119 | 0.0082 | 0.0050 |
| | LSCV | 0.0134 | 0.0089 | 0.0053 |
| | BCV2 | 0.0145 | 0.0095 | 0.0054 |
| | GREF | 0.0331 | 0.0263 | 0.0208 |
| (C) | MLCV | 0.0259 | 0.0175 | 0.0109 |
| | LSCV | 0.0261 | 0.0173 | 0.0107 |
| | BCV2 | 0.0467 | 0.0293 | 0.0107 |
| | GREF | 0.0274 | 0.0215 | 0.0158 |
| (D) | MLCV | 0.0253 | 0.0180 | 0.0116 |
| | LSCV | 0.0286 | 0.0186 | 0.0119 |
| | BCV2 | 0.0369 | 0.0227 | 0.0122 |

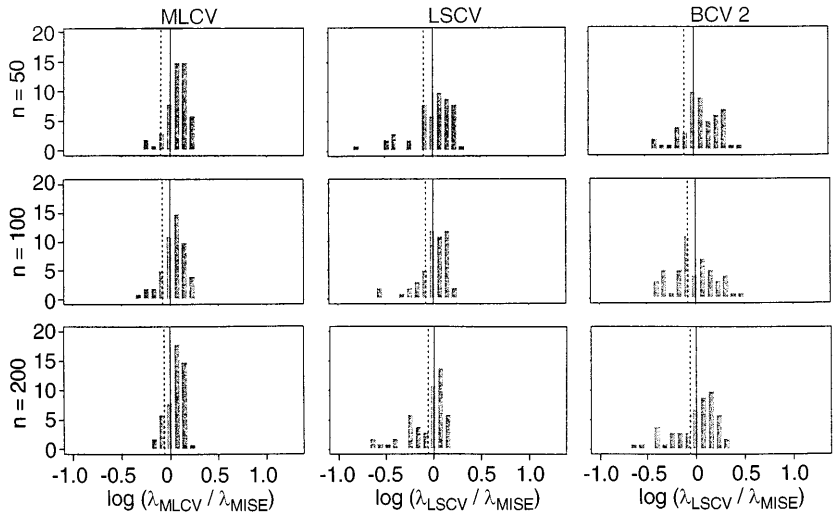


**Fig. 3.** Histograms of optimal global bandwidths from MLCV, BCV2, and LSCV for the test density A. The bandwidths are divided by the MISE optimal bandwidth and plotted on a log scale. The continuous and dotted lines denote the MISE optimal bandwidth and the Gaussian reference bandwidth in (8), respectively

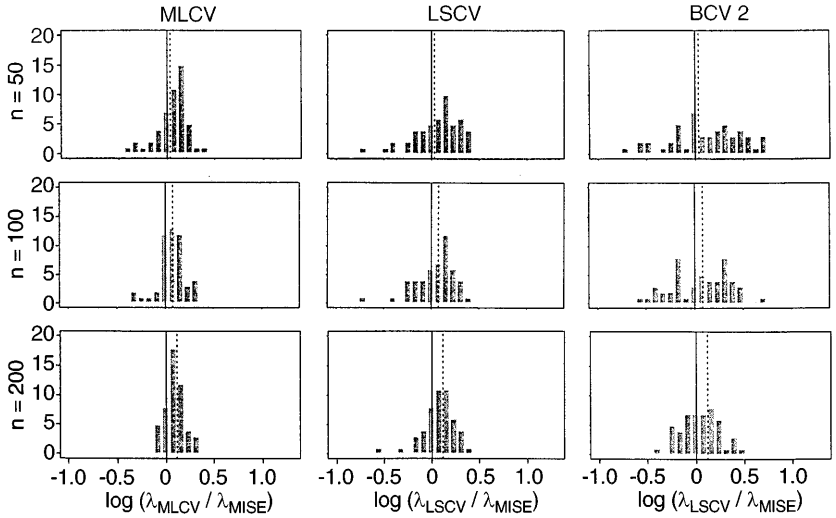


**Fig. 4.** Histograms of optimal global bandwidths from MLCV, BCV2, and LSCV for the test density B. The bandwidths are divided by the MISE optimal bandwidth and plotted on a log scale. The continuous and dotted lines denote the MISE optimal bandwidth and the Gaussian reference bandwidth in (8), respectively

Density *C* is more distinctly bimodal than density *B*. Gaussian reference is a poor choice for the bandwidth for this density (see fig. 5). Both MLCV and LSCV prove comparable and perform better than other methods. Both present comparable standard deviations and ISE scores, though LSCV shows a lower bias than MLCV. BCV2 proves disappointing for this density. The histogram showing BCV2

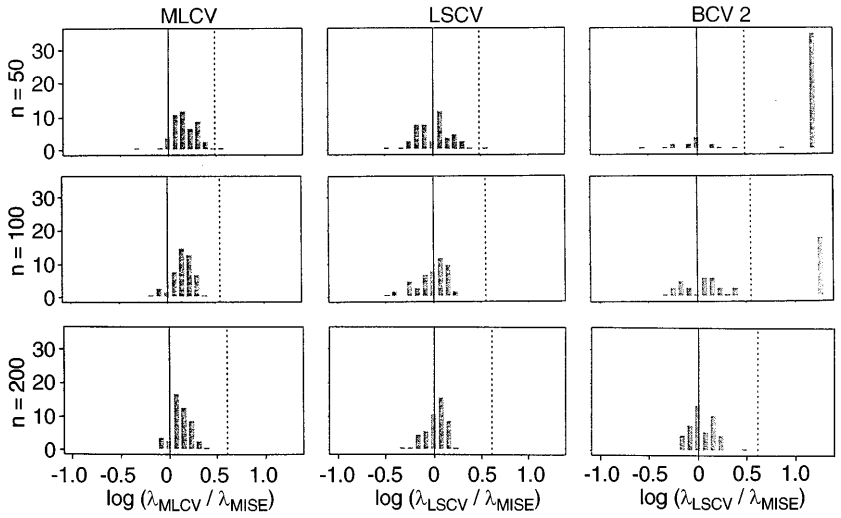


**Fig. 5.** Histograms of optimal global bandwidths from MLCV, BCV2, and LSCV for the test density C. The bandwidths are divided by the MISE optimal bandwidth and plotted on a log scale. The continuous and dotted lines denote the MISE optimal bandwidth and the Gaussian reference bandwidth in (8), respectively

**Table 3.** Sample bias and standard deviation of estimated optimal global bandwidths

| Density | Method | $n = 50$ | | $n = 100$ | | $n = 200$ | |
|---|---|---|---|---|---|---|---|
| | | $E(\lambda-\lambda_{MISE})$ | $\sigma(\lambda)$ | $E(\lambda-\lambda_{MISE})$ | $\sigma(\lambda)$ | $E(\lambda-\lambda_{MISE})$ | $\sigma(\lambda)$ |
| | MLCV | 0.0493 | 0.0681 | 0.0192 | 0.0613 | 0.0273 | 0.0404 |
| (A) | LSCV | 0.0107 | 0.1151 | 0.0081 | 0.0771 | −0.0216 | 0.0774 |
| | BCV2 | 0.0510 | 0.1163 | 0.0023 | 0.1115 | 0.0057 | 0.0909 |
| | MLCV | 0.0362 | 0.0793 | 0.0319 | 0.0596 | 0.0364 | 0.0415 |
| (B) | LSCV | 0.0435 | 0.1205 | 0.0271 | 0.0836 | 0.0326 | 0.0619 |
| | BCV2 | 0.0979 | 0.2034 | 0.0489 | 0.1370 | 0.0214 | 0.0741 |
| | MLCV | 0.0664 | 0.0586 | 0.0472 | 0.0359 | 0.0315 | 0.0277 |
| (C) | LSCV | 0.0125 | 0.0687 | −0.0012 | 0.0411 | 0.0036 | 0.0259 |
| | BCV2 | 0.5116 | 0.3318 | 0.2488 | 0.3153 | 0.0113 | 0.0329 |
| | MLCV | 0.0273 | 0.0652 | 0.0224 | 0.0407 | 0.0298 | 0.0335 |
| (D) | LSCV | 0.0186 | 0.0959 | 0.0067 | 0.0483 | 0.0016 | 0.0362 |
| | BCV2 | 0.2604 | 0.3175 | 0.0619 | 0.1866 | 0.0002 | 0.0501 |

optimal bandwidths in fig. 5 indicates the presence of two distinct modes for BCV2 optimal bandwidths for sample sizes 50 and 100. Considering the fact that our bandwidth search procedure is not permitted to extend beyond the limit of the plot, the standard deviation for this method would actually be higher than what is reported in table 3. It is notable, though, that the BCV2 results for sample size 200 show a distinct improvement (and a distinct mode in the histogram) over other sample sizes.

Figure 6 shows a large difference in the Gaussian reference and the MISE optimal bandwidths for density D. Much like the earlier case, the Gaussian reference choice compares poorly with the other methods. Multiple modes are again evident for BCV2 (for $n = 50$ and 100), hence the higher standard deviation (see

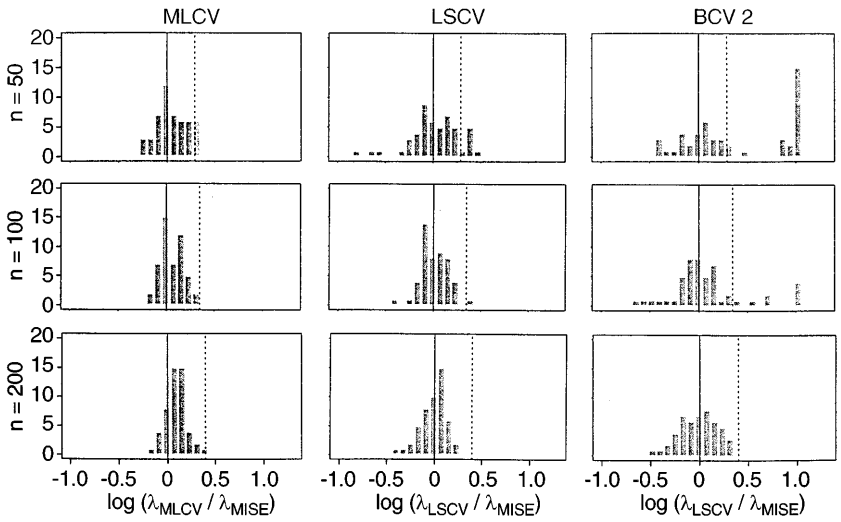


**Fig. 6.** Histograms of optimal global bandwidths from MLCV, BCV2, and LSCV for the test density D. The bandwidths are divided by the MISE optimal bandwidth and plotted on a log scale. The continuous and dotted lines denote the MISE optimal bandwidth and the Gaussian reference bandwidth in (8), respectively

**Table 4.** Mean ISE for local bandwidth-based estimators

| Density | Method | $n = 50$ | $n = 100$ | $n = 200$ |
|---|---|---|---|---|
| | GREF | 0.0089 | 0.0054 | 0.0039 |
| (A) | MLCV | 0.0090 | 0.0067 | 0.0062 |
| | LSCV | 0.0150 | 0.0058 | 0.0060 |
| | GREF | 0.0129 | 0.0086 | 0.0051 |
| (B) | MLCV | 0.0163 | 0.0117 | 0.0072 |
| | LSCV | 0.0181 | 0.0115 | 0.0059 |
| | GREF | 0.0303 | 0.0223 | 0.0158 |
| (C) | MLCV | 0.0268 | 0.0190 | 0.0123 |
| | LSCV | 0.0327 | 0.0219 | 0.0117 |
| | GREF | 0.0277 | 0.0203 | 0.0136 |
| (D) | MLCV | 0.0319 | 0.0231 | 0.0152 |
| | LSCV | 0.0372 | 0.0233 | 0.0147 |

table 3). Variability in the optimal bandwidth and the associated ISE are lowest for MLCV followed closely by LSCV.

On the whole, BCV2 does not appear to be a stable estimator of the optimal bandwidth. Amongst other choices, while Gaussian reference is a fairly good guess for Gaussian or near Gaussian data, MLCV and LSCV are the two most consistent bandwidth estimators amongst the ones studied.

## 5.2
## Results using local bandwidths

Results from using the optimal local bandwidth estimators developed in section 4.5 based on the Abramson-Silverman method (eq. 6), are given here. Table 4 shows the mean ISE for all test densities and sample sizes using the Gaussian

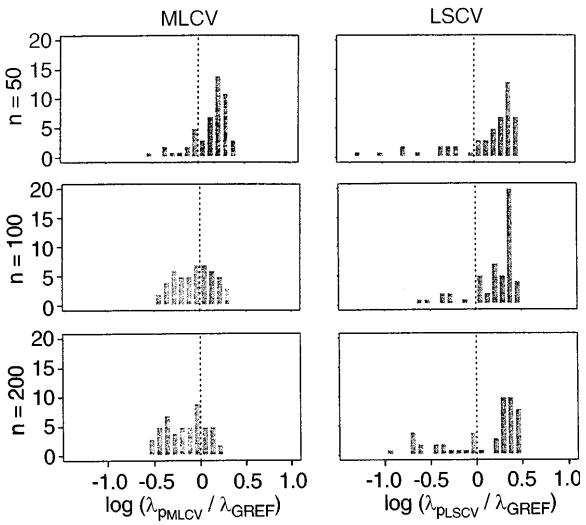


**Fig. 7.** Histograms of the optimal pilot bandwidths from MLCV and LSCV for the test density A. The bandwidths are scaled by the Gaussian reference bandwidth (shown dotted in figure) in (8) and plotted on a log scale. The scaling is done merely to provide a common scale for the different sample sizes for which the results are portrayed

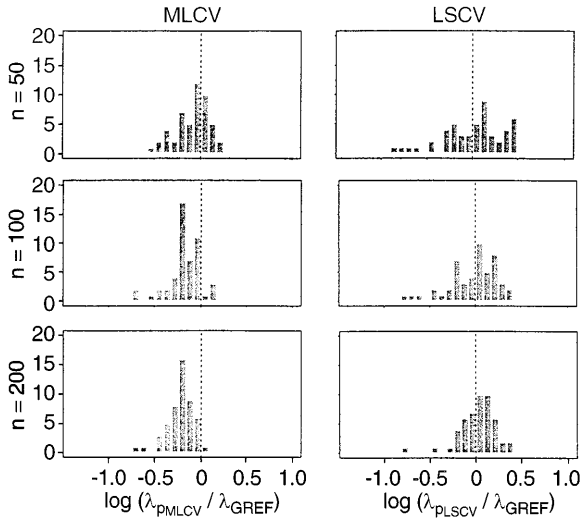


**Fig. 8.** Histograms of the optimal pilot bandwidths from MLCV and LSCV for the test density B. The bandwidths are divided by the Gaussian reference bandwidth (shown dotted in figure) in (8) and plotted on a log scale

reference, MLCV and LSCV criteria for selecting local bandwidths. Although there are n local bandwidths $\lambda_i$, they are all keyed to a single global pilot bandwidth $\lambda_p$ through eq. (6). We present results for these pilot bandwidths in our comparisons. Histograms of optimal pilot bandwidths for each of the four test densities are illustrated in figs. 7 (density A), 8 (density B), 9 (density C) and 10 (density

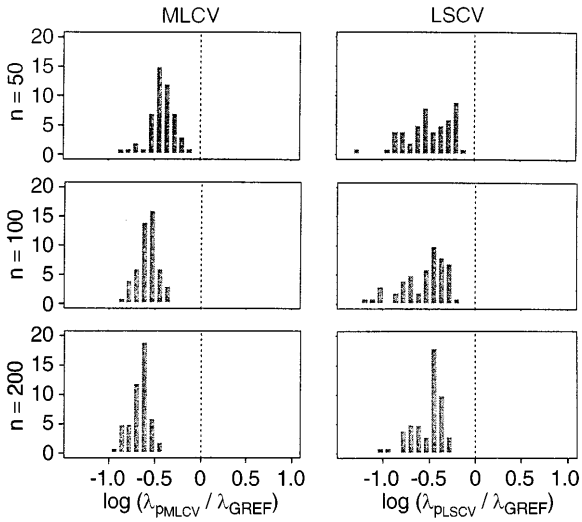


**Fig. 9.** Histograms of the optimal pilot bandwidths from MLCV and LSCV for the test density C. The bandwidths are divided by the Gaussian reference bandwidth (shown dotted in figure) in (8) and plotted on a log scale

**Table 5.** Sample standard deviations of estimated optimal $\lambda_p$

| Density | Method | $n = 50$ | $n = 100$ | $n = 200$ |
|---|---|---|---|---|
| (A) | MLCV | 0.1086 | 0.0926 | 0.0759 |
| | LSCV | 0.1892 | 0.1324 | 0.1572 |
| (B) | MLCV | 0.0843 | 0.0667 | 0.0471 |
| | LSCV | 0.1648 | 0.1117 | 0.0785 |
| (C) | MLCV | 0.0470 | 0.0295 | 0.0231 |
| | LSCV | 0.0771 | 0.0606 | 0.0389 |
| (D) | MLCV | 0.0583 | 0.0327 | 0.0268 |
| | LSCV | 0.1082 | 0.0635 | 0.0501 |

D). The Gaussian reference bandwidth (eq. 8) is also shown. Standard deviations of the optimal pilot bandwidths (using MLCV and LSCV) are presented in table 5.

The ISE scores in table 4 indicate no improvement over global bandwidth based estimators. Density A shows that LSCV performs marginally better than MLCV for sample size 100 and 200, though both are much worse than the global bandwidth results presented in table 2. This could be due in part to the higher variability of the optimal pilot density (see figs. 3 and 7).

GREF is picked over MLCV and LSCV for Density B, though no improvement over global bandwidth estimators is visible. A higher variability in the pilot bandwidths (see figs. 4 and 8) could again be the cause for this performance.

Results from density C are slightly reassuring. The Gaussian reference bandwidth when used as the pilot bandwidth shows better performance than its global counterpart. This is however countered by the fact that GREF is the worst choice (except BCV2) for density C in table 2 and is significantly higher than the optimal bandwidths picked by MLCV or LSCV (see fig. 9). MLCV and LSCV pick the best local bandwidths though there are no improvements over global MLCV or LSCV choices.

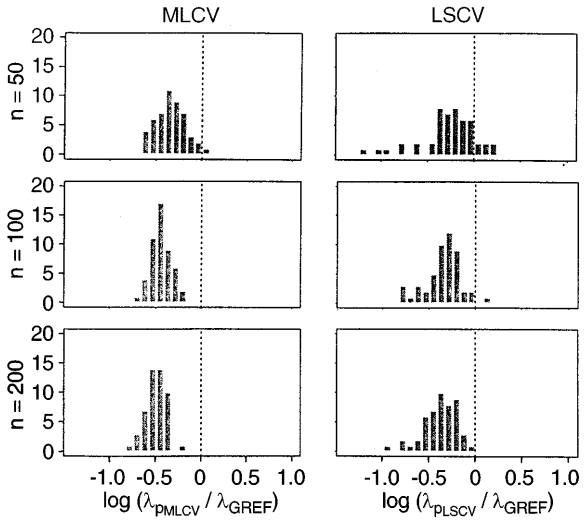


**Fig. 10.** Histograms of the optimal pilot bandwidths from MLCV and LSCV for the test density D. The bandwidths are divided by the Gaussian reference bandwidth (shown dotted in figure) in (8) and plotted on a log scale

Comparing table 4 to table 2, improvements in the ISE with GREF are evident for density D for sample sizes 100 and 200. It, however, heavily oversmoothes (see fig. 10) as compared to the MLCV or LSCV optimal bandwidths which are again inferior to their global counterparts.

On the whole, using local bandwidths did not result in any improvements over global bandwidth estimators. This was contrary to our expectations.

## 6
## Conclusions and Recommendations

Four methods for bandwidth estimation were evaluated. Results using both global and local bandwidths were compared. Although the list of test distributions over which these methods were compared is limited, we feel that these results are representative of plausible practical cases where limited data are available, and do provide guidance and insight for the practitioner.

The conclusions can be summarized as follows:

1) Local bandwidth estimation methods did not improve MISE performance relative to methods using a global smoothing parameter for the sample sizes and test distributions used.
2) Gaussian reference can be a good choice when the densities are weakly Gaussian (as in density A and B in table 1). Use of the Gaussian reference in other cases can lead to significant oversmoothing of the density estimate.
3) Biased Cross Validation (BCV2) does a poor job in estimating the optimal bandwidth, more so for sample size 50 or 100 than for size 200. One disturbing aspect of this method is the high variability of BCV2 optimal bandwidths. Two distinct peaks were visible in the histograms for BCV2 optimal bandwidths for densities C and D. This small sample performance of BCV2 is disappointing and contrary to the results in Sain et al. (1994). An increase in the variance of LSCV has been demonstrated in the past due to occasional extreme under-

smoothing. While BCV2 appears to avoid such extreme undersmoothing, its choice of $\lambda$ for $n < 200$ appears to be more diffuse than that from other cross validation criteria. The high variance and bias lead to a poor performance overall in the simulations presented here.

4) LSCV and MLCV both perform well. However, given the computational simplicity of maximum likelihood cross validation, we recommend MLCV over LSCV for use on small samples. Cautionary notes on the consistency of MLCV with fixed bandwidths and long tailed densities do however apply.

The bandwidth estimation methods that were selected from this study for use in the nonparametric streamflow simulation (Sharma et al. 1997) or disaggregation (Tarboton et al. 1997) models that were described before, were maximum likelihood cross validation and least squares cross validation. Both of these methods were found to work well for hydrologic time series with record lengths of up to 80 years. For reasons of conciseness, results using only the least square cross validation approach are presented in Sharma et al. (1997) and Tarboton et al. (1997).



## References

**Abramson IS** (1982) On bandwidth variation in kernel estimates – a square root law, The Annals of Statistics, 10(4), 1217–1223

**Bowman AW** (1984) An alternative method of cross-validation for the smoothing of density estimates, Biometrika, 71(2), 353–360

**Box GEP, Jenkins GM** (1976) Time Series Analysis: Forecasting and Control, Holden-Day, San Francisco, CA

**Chiu S-T** (1990) Why bandwidth selectors tend to choose smaller bandwidths, and a remedy, Biometrica, 77(1), 222–6

**Chiu S-T** (1991) Bandwidth selection for kernel density estimation, The Annals of Statistics, 19(4), 1883–1905

**Duin RPW** (1976) On the choice of smoothing parameters for parzen estimators of probability density functions, IEEE Transactions on Computers, 1175–1179

**Fukunaga K** (1972) Introduction to Statistical Pattern Recognition, Academic Press, New York

**Habbema JDF, Hermans J, Vandenbroek K** (1974) "A stepwise discriminant analysis program using density estimation," COMPSTAT 1974, ed. G. Bruckmann, Wien: Physica Verlag., 101–110

**Lall U, Moon YI, Bosworth K** (1993) Kernel Flood Frequency Estimators: Bandwidth Selection and Kernel Choice, Water Resources Research, 29 (4), 1005–1015

**Lall U** (1995) Recent Advances in Nonparametric Function Estimation, pp. 1093–1102, U.S. National Report to International Union of Geodesy and Geophysics 1991–1994, Reviews of Geophysics

**Lall U, Sharma A** (1996) A nearest neighbor bootstrap for time series resampling, Water Resources Research, 32 (3), 679–693

**Press WH, Flannery BP, Teukolsky SA, Vetterling WT** (1989) Numerical recipes: The art of scientific computing (Fortran Version), Cambridge University Press, New York

**Rudemo M** (1982) Empirical choice of histograms and kernel density estimators, Scandinavian Journal of Statistics, 9, 65–78

**Sain SR, Baggerly KA, Scott DW** (1994) Cross-Validation of Multivariate Densities, Journal of the American Statistical Association, 89 (427), 807–817

**Schuster EF** (1985) Incorporating support constraints into nonparametric estimators of densities, Commun. Statist.-Theor. Meth., 14(5), 1123–1136

**Scott DW** (1992) Multivariate Density Estimation: Theory, Practice and Visualization, John Wiley & Sons Inc

**Scott DW, Terrell GR** (1987) Biased and unbiased cross-validation in density estimation, Journal of the American Association, 82 (400), 1131–1146

**Silverman BW** (1986) Density estimation for statistics and data analysis, Chapman and Hall, 175 p

**Sharma A, Tarboton DG, Lall U** (1997) Streamflow simulation: A nonparametric approach, Water Resources Research, 33(2), 291–308
**Tarboton DG, Sharma A, Lall U** (1998) Disaggregation Procedures for Stochastic Hydrology based on Nonparametric Density Estimation, Water Resources Research 34(1), 107–120
**Wand MP, Jones MC** (1993) Comparison of Smoothing Parameterizations in Bivariate Kernel Density Estimation, Journal of the American Statistical Association, 88(422), 520–528
**Wand MP, Jones MC** (1994) Multivariate Plug-in Bandwidth Selection, Journal of Computational Statistics, 9, 97–116
**Wand MP, Jones MC** (1995) Kernel Smoothing, Chapman and Hall, London, 212 p