THE WEIGHTED HELLINGER DISTANCE IN THE MULTIVARIATE KERNEL DENSITY ESTIMATION

A. R. Mugdadi [1]
Jordan University of Science and Technology, Jordan
e-mail: *aamugdadi@just.edu.jo*

Haneef Anver
Southern Illinois University, Carbondale, USA

Key words: Bandwidth Selection, Hellinger distance, Least squares, Multivariate kernel estimation, Plug-in

Abstract: The kernel multivariate density estimation is an important technique to estimate the multivariate density function. In this investigation we will use Hellinger Distance as a measure of error to evaluate the estimator, we will derive the mean weighted Hellinger distance for the estimator, and we obtain the optimal bandwidth based on Hellinger distance. Also, we propose and study a new technique to select the matrix of bandwidths based on Hellinger distance, and compare the new technique with the plug-in and the least squares techniques.

1. Introduction

The univariate kernel density estimators for a random sample $X_1, X_2, \ldots, X_n$, drawn from a probability density function f, is

$$\hat{f}(x,h) = n^{-1}\sum_{i=1}^{n} k_h(x-X_i),$$

where k is chosen to be a unimodal probability density function that is symmetric about zero.

The general form of the d-variate kernel density estimator, for a random sample $\mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_n$ drawn from a density f, is

$$\hat{f}(\mathbf{x},\mathbf{H}) = n^{-1}\sum_{i=1}^{n} \mathbf{k}_{\mathbf{H}}(\mathbf{x}-\mathbf{X}_i),$$

where $\mathbf{x} = (x_1, x_1, \ldots, x_d)^T$ and $\mathbf{X}_i = (\mathbf{X}_{i1}, \mathbf{X}_{i2}, \ldots, \mathbf{X}_{in})^T$,
$i = 1, 2, \ldots, n$. Here $\mathbf{k}$ is the unscaled kernel, $\mathbf{k}_{\mathbf{H}}$ is the scaled kernel and $\mathbf{H}$ is the
$d \times d$ (fixed) bandwidth matrix, which is non-random, symmetric and positive definite. The scaled and unscaled kernels are related by $\mathbf{k}_{\mathbf{H}}(x) = |\mathbf{H}|^{-1/2}\mathbf{k}(\mathbf{H}^{-1/2}\mathbf{x})$. This formulation is a little different from the univariate case since the 1×1 bandwidth matrix is $\mathbf{H} = h^2$ so we are dealing with squared bandwidths here.

[1] Corresponding author.

AMS: 62G05, 62G07, 62H12

We restrict our attention to kernel functions **k** that are spherically symmetric probability density functions. Moreover, we mostly use normal kernels throughout this paper for two reasons: they lead to smooth density estimates and they simplify the mathematical analysis. The bandwidth selector plays a central role in determining the performance of kernel density estimators. Thus we wish to select bandwidths which give the optimal performance which is measured by the closeness of a kernel density estimate to its target density. There are many possible error criteria from which to choose. In this investigation we will concentrate on the mean weighted Hellinger distance as a measure of error. Also, we propose a new technique to select the matrix of bandwidths. A detailed discussion about the multivariate density estimation and the bandwidth selection techniques can be found in Anver (2010), Scott (1992) and Wand and Jones (1995).

2. The asymptotic mean weighted Hellinger distance

The common global error criterion is the mean integrated squared error (MISE), which is defined by:

$$MISE\left\{\hat{f}(\mathbf{x};\mathbf{H})\right\} = E\left\{ISE\hat{f}(.;\mathbf{H})\right\} = E\int\left[\hat{f}(\mathbf{x};\mathbf{H}) - f(\mathbf{x})\right]^2 d\mathbf{x}.$$

Another technique to measure the error is called the mean weighted Hellinger distance, which is defined for the univariate case by:

$$MWHD\left\{(\hat{f}(\mathbf{x};\mathbf{H})\right\} = E\int\left[\hat{f}^{1/2}(\mathbf{x}) - f^{1/2}(\mathbf{x})\right]^2 f(\mathbf{x})d\mathbf{x}. \tag{1}$$

Kanzawa (1995) discussed the relationship between the asymptotic mean Hellinger Distance (AMHD) and the asymptotic mean integrated square error (AMISE) when $f(x)$ defined on a compact set. Ahmad and Mugdadi (2006) discussed the relation between the asymptotic mean weighted Hellinger distance (AMWHD) and the AMISE for both $\hat{f}(x)$ and the estimated kernel distribution function $\hat{F}(x)$. Mugdadi (2004) used Hellinger distance to derive the bandwidth for the kernel density estimation of function of observations. Thus, the MWHD for the multivariate kernel density estimation can be defined by:

$$MWHD(\hat{f}(\mathbf{x};\mathbf{H})) = E\int(\hat{f}^{1/2}(\mathbf{X}) - f^{1/2}(\mathbf{X}))^2 f(\mathbf{X})d\mathbf{X}.$$

Theorem 1 Using the assumptions:

1. Each entry of $\mathscr{H}_f(.)$ is piecewise continuous and square integrable.
2. $\mathbf{H} = \mathbf{H}_n$ is a sequence of bandwidth matrices such that $n^{-1}|\mathbf{H}|^{-1/2}$ and all entries of $\mathbf{H}$ approach zero as $n \to \infty$. Also, we assume that the ratio of the largest and smallest eigenvalues of $\mathbf{H}$ is bounded for all n.
3. **k** is a bounded and compactly supported d-variate kernel.
4. $f^{(4)}(\mathbf{x})$ exists.

The $AMWHD\{\hat{f}(.;\mathbf{H})\}$ is given by:

$$\text{AMWHD}\{\hat{f}(.;\mathbf{H})\} = \frac{1}{4}n^{-1}|\mathbf{H}|^{-1/2}R(\mathbf{k}) + \frac{1}{16}\mu_2(\mathbf{k})^2 \int tr^2\{\mathbf{H}\mathscr{H}_f(\mathbf{x})\}\, d\mathbf{x}.$$

Proof.

Note that

$$\begin{aligned} MWHD(\hat{f}(.;\mathbf{H})) &= E\int (\hat{f}^{1/2} - f^{1/2})^2 f(\mathbf{x})d\mathbf{x} \\ &= \int E\hat{f}(\mathbf{x})f(\mathbf{x})d\mathbf{x} - 2\int E\hat{f}^{1/2}f^{3/2}d\mathbf{x} + \int Ef^2(\mathbf{x})d\mathbf{x} \\ &= I - 2II + R(f). \end{aligned}$$

However by the multivariate version of Taylor's theorem

$$\begin{aligned} E\hat{f}(\mathbf{x};\mathbf{H}) &= \int \mathbf{k}_{\mathbf{H}}(\mathbf{x}-\mathbf{y})f(\mathbf{y})d\mathbf{y} \\ &= \int \mathbf{k}(\mathbf{z})f(\mathbf{x}-\mathbf{H}^{1/2}\mathbf{z})d\mathbf{z} \\ &= \int \{f(\mathbf{x}) - (\mathbf{H}^{1/2}\mathbf{z})^T\mathbf{D}_f(\mathbf{x}) + \frac{1}{2}(\mathbf{H}^{1/2}\mathbf{z})^T\mathscr{H}_f(\mathbf{x})(\mathbf{H}^{1/2}\mathbf{z})\}d\mathbf{z} + o\{tr(\mathbf{H})\} \\ &= f(\mathbf{x}) - \int z^T\mathbf{H}^{1/2}\mathbf{D}_f(\mathbf{x})\mathbf{k}(\mathbf{z})d\mathbf{z} + \frac{1}{2}\int \mathbf{z}^T\mathbf{H}^{1/2}\mathscr{H}_f(\mathbf{x})\mathbf{H}^{1/2}z\mathbf{k}(\mathbf{z})d\mathbf{z} + o\{tr(\mathbf{H})\} \\ &= f(\mathbf{x}) + \frac{1}{2}tr\{\mathbf{H}^{1/2}\mathscr{H}_f(\mathbf{x})\mathbf{H}^{1/2}\int \mathbf{z}\mathbf{z}^T\mathbf{k}(\mathbf{z})d\mathbf{z}\} + o\{tr(\mathbf{H})\} \\ &= f(\mathbf{x}) + \frac{1}{2}\mu_2(\mathbf{k})tr\{\mathbf{H}\mathscr{H}_f(\mathbf{x})\} + o\{tr(\mathbf{H})\}. \end{aligned}$$

Thus

$$\begin{aligned} I &= \int f(\mathbf{x})^2 d\mathbf{x} + \frac{1}{2}\mu_2(\mathbf{k})\int tr\{\mathbf{H}\mathscr{H}_f(\mathbf{x})\}f(\mathbf{x})d\mathbf{x} \\ &= R(f) + \frac{1}{2}\mu_2(\mathbf{k})\int tr\{\mathbf{H}\mathscr{H}_f(\mathbf{x})\}f(\mathbf{x})d\mathbf{x}. \end{aligned}$$

Next, from the standard analysis of $\hat{f}(\mathbf{x})$, we have that

$$\begin{aligned} \hat{f}(\mathbf{x}) &= f(\mathbf{x}) + \frac{1}{2}\mu_2(\mathbf{k})tr\{\mathbf{H}\mathscr{H}_f(\mathbf{x})\} + o\{tr(\mathbf{H})\} + \mathbf{Z}[n^{-1}|\mathbf{H}|^{-1/2}R(\mathbf{k})f(\mathbf{x})]^{1/2} \\ \hat{f}(\mathbf{x})^{1/2} &= f(\mathbf{x})^{1/2}\left\{1 + \frac{1}{2f(\mathbf{x})}\mu_2(\mathbf{k})tr\{\mathbf{H}\mathscr{H}_f(\mathbf{x})\} + o\{tr(\mathbf{H})\}\right. \\ &\quad \left. + \mathbf{Z}\left[\frac{n^{-1}|\mathbf{H}|^{-1/2}R(\mathbf{k})}{f(\mathbf{x})}\right]^{1/2}\right\}^{1/2}. \end{aligned}$$

Thus,

$$\begin{aligned} \hat{f}^{1/2}(\mathbf{x}) \cong f(\mathbf{x})^{1/2}\left\{1 + \frac{1}{4f(\mathbf{x})}\mu_2(\mathbf{k})tr\{\mathbf{H}\mathscr{H}_f(\mathbf{x})\} + \frac{o\{tr(\mathbf{H})\}}{2}\right. \\ \left. + \frac{\mathbf{Z}}{2}\left[\frac{n^{-1}|\mathbf{H}|^{-1/2}R(\mathbf{k})}{f(\mathbf{x})}\right]^{1/2}\right\} - \frac{w^2}{8}, \end{aligned}$$

where

$$\mathbf{w}^2 = \left[\frac{1}{2}f(\mathbf{x})\mu_2(\mathbf{k})tr\{\mathbf{H}\mathscr{H}_f(\mathbf{x})\}\right]^2 + \mathbf{z}^2\left[\frac{n^{-1}|\mathbf{H}|^{-1/2}R(\mathbf{k})}{f(\mathbf{x})}\right] + \mathbf{z}\frac{1}{f(\mathbf{x})}\mu_2(\mathbf{k})tr\{\mathbf{H}\mathscr{H}_f(\mathbf{x})\}\left[\frac{n^{-1}|\mathbf{H}|^{-1/2}R(\mathbf{k})}{f(\mathbf{x})}\right]^{1/2}.$$

Thus

$$E\hat{f}^{1/2}(\mathbf{x};\mathbf{H}) = f(\mathbf{x})^{1/2}\left\{1 + \frac{1}{4f(\mathbf{x})}\mu_2(\mathbf{k})tr\{\mathbf{H}\mathscr{H}_f(\mathbf{x})\} + \frac{1}{2}\left[\frac{n^{-1}|\mathbf{H}|^{-1/2}R(\mathbf{k})}{f(\mathbf{x})}\right]^{1/2} - \frac{1}{32f(\mathbf{x})^2}\mu_2(\mathbf{k})^2tr^2\{\mathbf{H}\mathscr{H}_f(\mathbf{x})\} - \frac{1}{8}\left[\frac{n^{-1}|\mathbf{H}|^{-1/2}R(\mathbf{k})}{f(\mathbf{x})}\right]\right\}.$$

$$\begin{aligned}
2\text{II} &= 2\int E\hat{f}^{1/2}(\mathbf{x})f^{\frac{3}{2}}(\mathbf{x})d\mathbf{x} \\
&\quad + 2\int f^2(\mathbf{x})d\mathbf{x} + \frac{\mu_2(\mathbf{k})}{2}\int f(\mathbf{x})tr\{\mathbf{H}\mathscr{H}_f(\mathbf{x})\}d\mathbf{x} + \int\left[\frac{n^{-1}|\mathbf{H}|^{-1/2}R(\mathbf{k})}{f(\mathbf{x})^{1/2}f^2(\mathbf{x})}\right]d\mathbf{x} \\
&\quad - \int\frac{1}{16}\mu_2(\mathbf{k})^2tr^2\{\mathbf{H}\mathscr{H}_f(\mathbf{x})\}d\mathbf{x} - \frac{1}{4}\int\left[\frac{n^{-1}|\mathbf{H}|^{-1/2}R(\mathbf{k})}{f(\mathbf{x})}\right]f^2(\mathbf{x})d\mathbf{x} \\
&= 2\int f^2(\mathbf{x})d\mathbf{x} + \frac{\mu_2(\mathbf{k})}{2}\int f(\mathbf{x})tr\{\mathbf{H}\mathscr{H}_f(\mathbf{x})\}d\mathbf{x} + \left[n^{-1}|\mathbf{H}|^{-1/2}R(\mathbf{k})\right]^{1/2}\int f(\mathbf{x})d\mathbf{x} \\
&\quad - \frac{1}{16}\mu_2(\mathbf{k})^2\int tr^2\{\mathbf{H}\mathscr{H}_f(\mathbf{x})\}d\mathbf{x} - \frac{1}{4}\left[n^{-1}|\mathbf{H}|^{-1/2}R(\mathbf{k})\right]\int f(\mathbf{x})d\mathbf{x}.
\end{aligned}$$

Therefore I - 2II + R(f)

$$\begin{aligned}
&= R(f) + \frac{1}{2}\mu_2(\mathbf{k})\int tr\{\mathbf{H}\mathscr{H}_f(\mathbf{x})\}f(\mathbf{x})d\mathbf{x} - 2R(f) - \frac{\mu_2(\mathbf{k})}{2}\int tr\{\mathbf{H}\mathscr{H}_f(\mathbf{x})\}f(\mathbf{x})d\mathbf{x} \\
&\quad - [n^{-1}|\mathbf{H}|^{-1/2}R(\mathbf{k})]^{-1/2} + \frac{1}{16}\mu_2(\mathbf{k})^2\int tr^2\{\mathbf{H}\mathscr{H}_f(\mathbf{x})\}d\mathbf{x} + \frac{1}{4}[n^{-1}|\mathbf{H}|^{-1/2}R(\mathbf{k})] + R(f) \\
&= \frac{1}{8}\mu_2(\mathbf{k})^2\int tr^2\{\mathbf{H}\mathscr{H}_f(\mathbf{x})\}d\mathbf{x} - [n^{-1}|H|^{-1/2}R(\mathbf{k})]^{-1/2} + \frac{1}{4}[n^{-1}|\mathbf{H}|^{-1/2}R(\mathbf{k})] \\
&= \frac{1}{4}[n^{-1}|\mathbf{H}|^{-1/2}R(\mathbf{k})] + \frac{1}{16}\mu_2(\mathbf{k})^2\int tr^2\{\mathbf{H}\mathscr{H}_f(\mathbf{x})\}d\mathbf{x} - [n^{-1}|H|^{-1/2}R(\mathbf{k})]^{1/2}.
\end{aligned}$$

The theorem is now proved. ■

Under the above conditions and in the case where $\mathbf{H} = h^2\mathbf{I}$ we obtain the following corollary:

Corollary 1

$$\text{AMWHD}\{\hat{f}(.;\mathbf{H})\} = \frac{n^{-1}h^{-d}R(\mathbf{k})}{4} + \frac{1}{16}h^4\mu_2(\mathbf{k})^2\int\{\nabla^2 f(\mathbf{x})\}^2d\mathbf{x},$$

where

$$\nabla^2 f(\mathbf{x}) = \sum_{i=1}^{d}(\partial^2/\partial x_i^2)f(\mathbf{x}).$$

In this case the optimal bandwidth has an explicit formula given by the following corollary.

Corollary 2 Under the conditions of Theorem 1 and in the case of $\mathbf{H} = h^2\mathbf{I}$, the optimal bandwidth is given as

$$h_{AMWHD} = \left[\frac{dR(\mathbf{k})}{\mu_2(\mathbf{k})^2 \int \{\nabla^2 f(\mathbf{x})\}^2 d\mathbf{x}\, n}\right]^{1/(d+4)}.$$

This can be easily proved by differentiating (1) with respect to h and setting the derivative equal to zero and solving for h.

3. Data-dependent bandwidth choices

The problem of selecting the scalar bandwidth in univariate kernel density estimation is quite well understood. A number of methods exist that combine good theoretical properties with strong practical performance. Many of these techniques can be extended to the multivariate case in a relatively straight forward fashion if $\mathbf{H}$ is constrained to be a diagonal matrix. However, imposing such a constraint on the bandwidth matrix can result in markedly suboptimal density estimates.

The preceding discussion indicates a need for data-dependent methods for choosing full (i.e unconstrained) bandwidth matrices. The development of selectors for full $\mathbf{H}$ is rather more challenging than that for diagonal $\mathbf{H}$. In particular, the need to consider the orientation of kernel functions to the coordinate axes in the former case introduces a problem without a univariate analogue. The additional difficulties in selecting full (as opposed to diagonal) bandwidth matrices partly explain the relatively slow progress in this area of the two major approaches to bandwidth selection, namely the plug-in method and cross validation (CV) techniques. Only the former has received attention to date in the context of full bandwidth matrices. Stone (1984) looks at the multivariate least squares cross validation (LSCV) criterion for the multivariate kernel density estimator. Wand and Jones (1994) outlined a plug-in selector that can be applied to full bandwidth matrices, but they concentrated on a diagonal $\mathbf{H}$ when presenting methodological particulars. A more detailed account of plug-in selectors for full $\mathbf{H}$ was provided by Duong and Hazelton (2003).

All CV bandwidth matrix selectors aim to estimate MISE, or AMISE, and some combine the two (modulo a constant) and then minimize the resulting function. The unbiased cross-validation (UCV) targets MISE and employs the objective function

$$UCV(H) = \int \hat{f}(\mathbf{x};\mathbf{H})^2 d\mathbf{x} - 2n^{-1}\sum_{i=1}^{n} \hat{f}_{-i}(\mathbf{X}_i;\mathbf{H}),$$

where

$$\hat{f}_{-i}(\mathbf{x}_i;\mathbf{H}) = (n-1)^{-1}\sum_{j\neq i}^{n} \mathbf{k}_{\mathbf{H}}(\mathbf{x}-\mathbf{X}_j),$$

is a leave-one-out estimator of f. The function $UCV(H)$ is unbiased in the sense that $E[UCV(H)]$= $MISE[\hat{f}(.;H)]$ - $R(f)$. It can be expanded to give

$$UCV(H) = n^{-2}\sum_{i=1}^{n}\sum_{j=1}^{n}(\mathbf{k}_{\mathbf{H}} * \mathbf{k}_{\mathbf{H}})(\mathbf{X}_i-\mathbf{X}_j) - n^{-1}(n-1)^{-1}\sum_{i=1}^{n}\sum_{j=1}^{n}\mathbf{k}_{\mathbf{H}}(\mathbf{X}_i-\mathbf{X}_j),$$

where * denotes the convolution. The UCV bandwidth matrix selector $\hat{H}_{UCV}$ is the minimiser of $UCV(H)$. It is usual to judge the performance of a bandwidth matrix $\mathbf{H}$ according to a global error

criterion for $\hat{f}(\mathbf{x})$. Next we derive the estimate for the bandwidth matrix $\mathbf{H}$ using MWHD as the error criteria and call it WHCV($\mathbf{H}$).

$$\begin{aligned}\text{WHCV}(\mathbf{H}) =& E\int\left\{\hat{f}^{1/2}(\mathbf{x};\mathbf{H})-f^{1/2}(\mathbf{x})\right\}^2 f(\mathbf{x})d\mathbf{x}\\ =& E\left\{\int\hat{f}(\mathbf{x};\mathbf{H})f(\mathbf{x})d\mathbf{x}-2\int\hat{f}^{1/2}(\mathbf{x};\mathbf{H})f^{3/2}(\mathbf{x})d\mathbf{x}+R(f)\right\}\\ =& E\left\{\int\hat{f}(\mathbf{x};\mathbf{H})f(\mathbf{x})d\mathbf{x}-2\sum_{i=1}^{n}\hat{f}^{1/2}(\mathbf{x}_i;\mathbf{H})\hat{f}_i^{1/2}(\mathbf{x}_i)\int f(\mathbf{x})d\mathbf{x}\right\}+R(f).\end{aligned}$$

Thus the $WHC(\mathbf{H})$ can be estimated by

$$W\hat{H}CV(\mathbf{H})=\frac{\sum_{i=1}^{n}\hat{f}(\mathbf{x}_i)}{n}-\frac{2}{n}\sum_{i=1}^{n}\hat{f}_i^{1/2}(\mathbf{x}_i;\mathbf{H})\hat{f}^{1/2}(\mathbf{x}_i;\mathbf{H})+R(f).$$

The WHCV bandwidth matrix selector $\hat{H}_{WHCV}$ is the minimiser of $W\hat{HCV}(\mathbf{H})$. Minimising of such function can be done using the Powell Optimization method — an efficient method for finding the minimum of a function of several variables without calculating derivatives (Powell, 1964).

3.1. Comparing the performance of the WHCV matrix selector with other bandwidth selectors for normal mixture densities

In order to asses the performance of the new technique to select the matrix of bandwidths we will simulate from the densities that are displayed in Table 1. These densities represent a wide range of multivariate densities as shown in Figure 1.

Table 1: Formulas for target densities A – E.

Target Density	Formula
A	$N\left(\begin{bmatrix}0\\0\end{bmatrix},\begin{bmatrix}0.25 & 0\\0 & 1\end{bmatrix}\right)$
B	$\frac{1}{2}N\left(\begin{bmatrix}1\\0\end{bmatrix},\begin{bmatrix}4/9 & 0\\0 & 4/9\end{bmatrix}\right)+1/2N\left(\begin{bmatrix}-1\\0\end{bmatrix},\begin{bmatrix}4/9 & 0\\0 & 4/9\end{bmatrix}\right)$
C	$1/2N\left(\begin{bmatrix}1\\-0.9\end{bmatrix},\begin{bmatrix}1 & 0.9\\0.9 & 1\end{bmatrix}\right)+1/2N\left(\begin{bmatrix}-1\\0.9\end{bmatrix},\begin{bmatrix}1 & 0.9\\0.9 & 1\end{bmatrix}\right)$
D	$1/2N\left(\begin{bmatrix}1\\-1\end{bmatrix},\begin{bmatrix}4/9 & 14/45\\14/45 & 4/9\end{bmatrix}\right)+1/2N\left(\begin{bmatrix}-1\\1\end{bmatrix},\begin{bmatrix}4/9 & 0\\0 & 4/9\end{bmatrix}\right)$
E	$N\left(\begin{bmatrix}0\\0\end{bmatrix},\begin{bmatrix}1 & 9/10\\9/10 & 1\end{bmatrix}\right)$

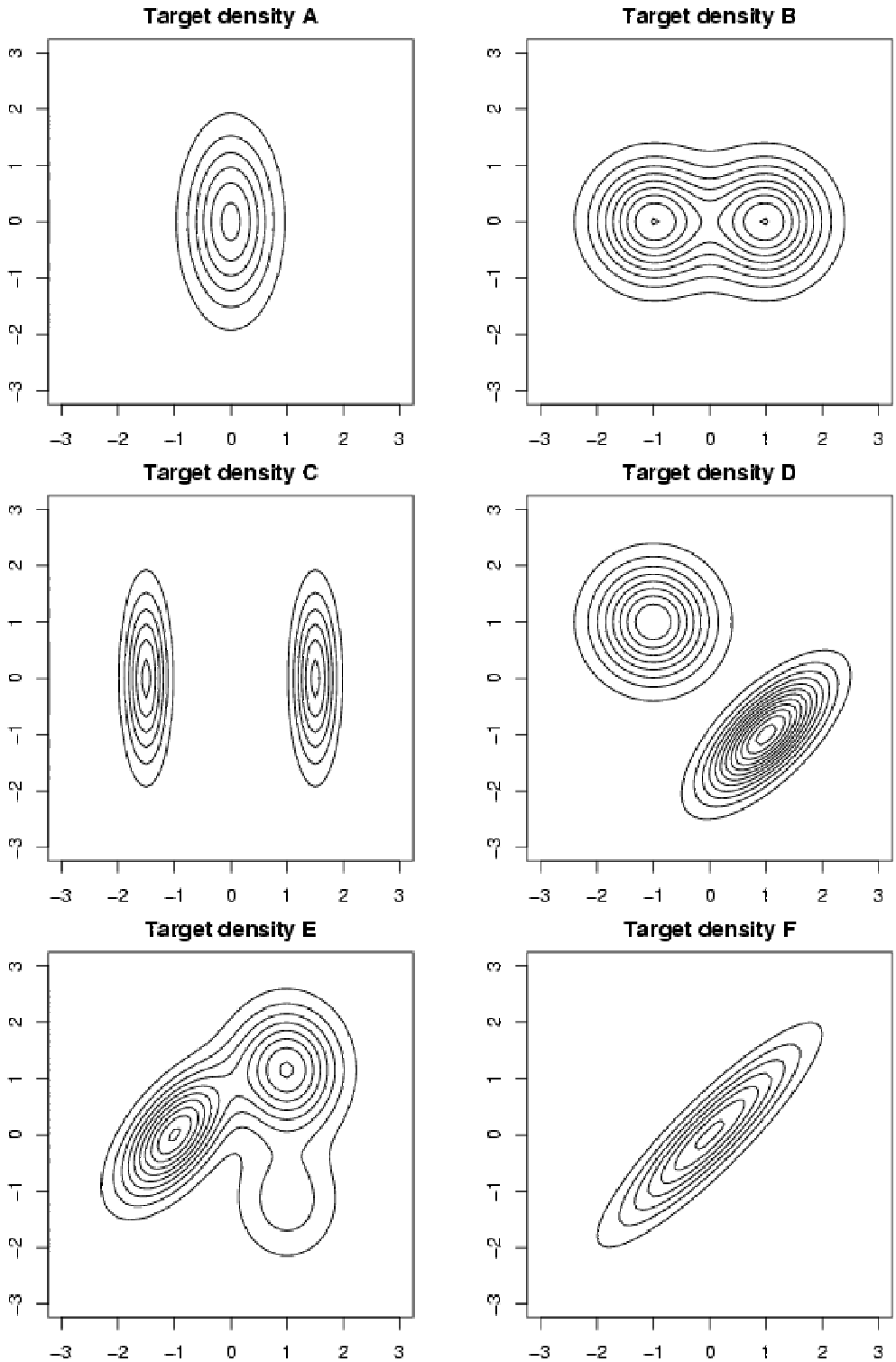

Figure 1: Contour plots for target densities A – E.

Bandwidth matrices were selected for 100 random sample generated from the densities in Table 1. The integrated square error (ISE) for each resulting density estimate was recorded. They are displayed in Figure 2 using box plots with a log scale. The performance of a selector depends largely on the target density shape. We see that the plug-in bandwidth selector performs well for densities A, B, D and E. Both CV selectors perform adequately for densities B, D and E, but cannot compete with the plug-in bandwidth selector on density B and E and to a lesser extent on density D. For density C one noteworthy aspect is that the performance of WHCV is better than the plug-in bandwidth selector. LSCV suffers in comparison to WHCV for densities A, B and C, and giving similar results for D and E.

Performance of bandwidth selectors are more visible and provide a variety of interpretations of the structure of the data when we plot contour plots. To illustrate this we simulated random samples for densities A – E in two cases. The first case considers small sample simulations of size 10 for each density for two instances and the results are given in Figures 3, 5, 7, 9 and 11. For the second case we looked at comparatively large samples of sizes 200 for each density and results are given in Figures 4, 6, 8, 10 and 12. Various bandwidth selectors provide a variety of interpretations of the structure of the data.

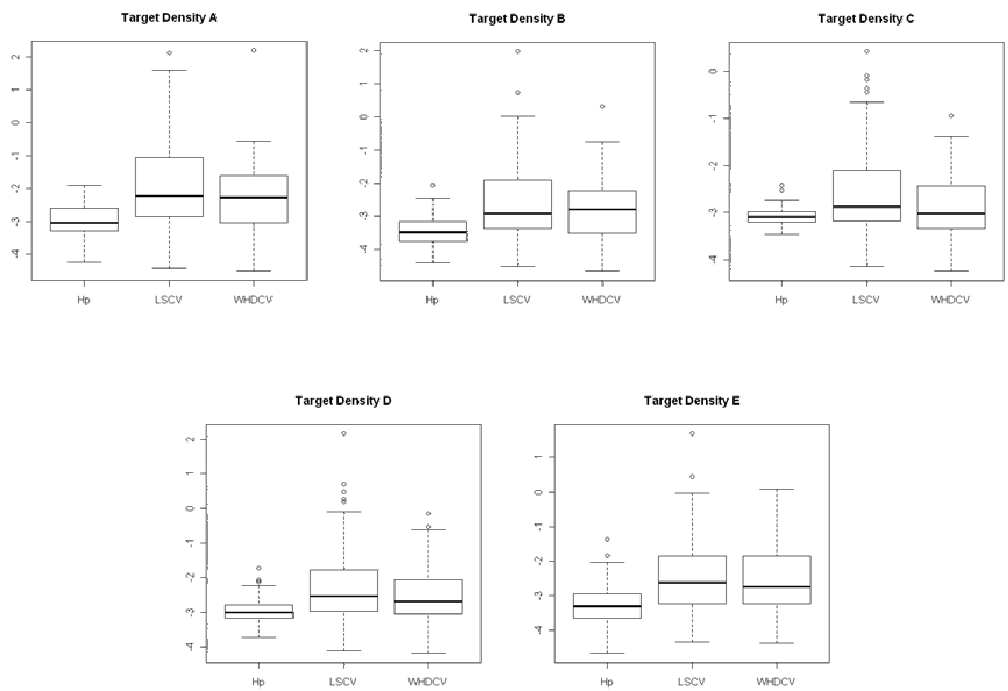

Figure 2: Box plots of log(ISE) for bivariate bandwidth selectors.

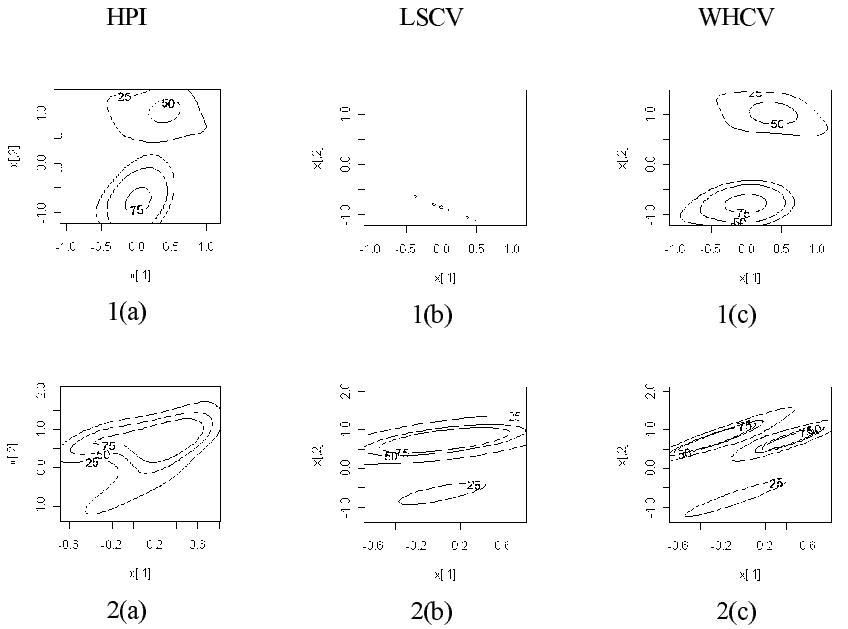

Figure 3: Density A: small sample.

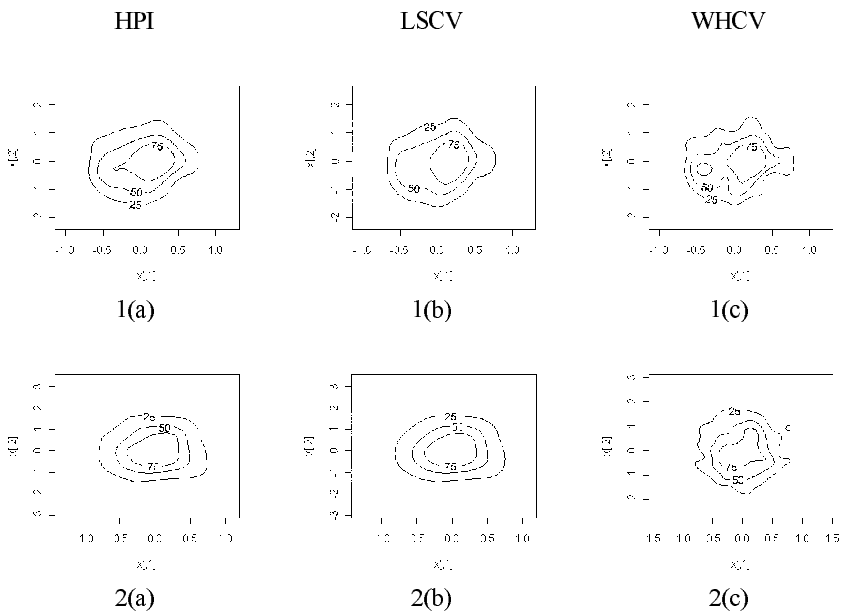

Figure 4: Density A: large sample.

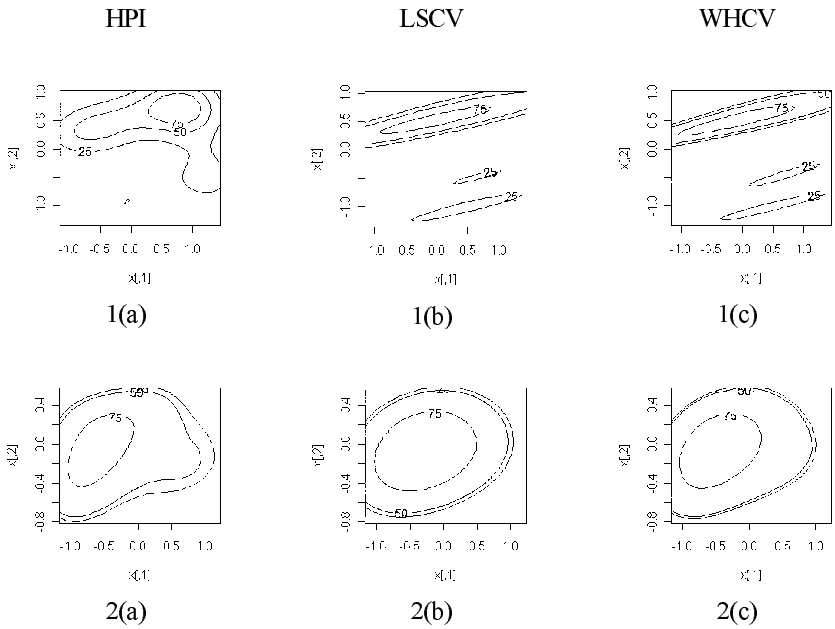

Figure 5: Density B: small sample.

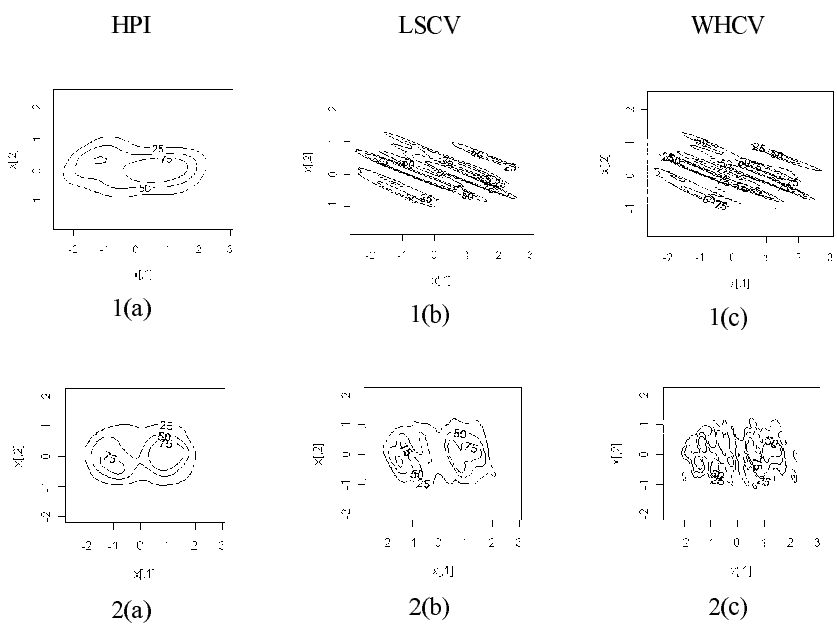

Figure 6: Density B: large sample.

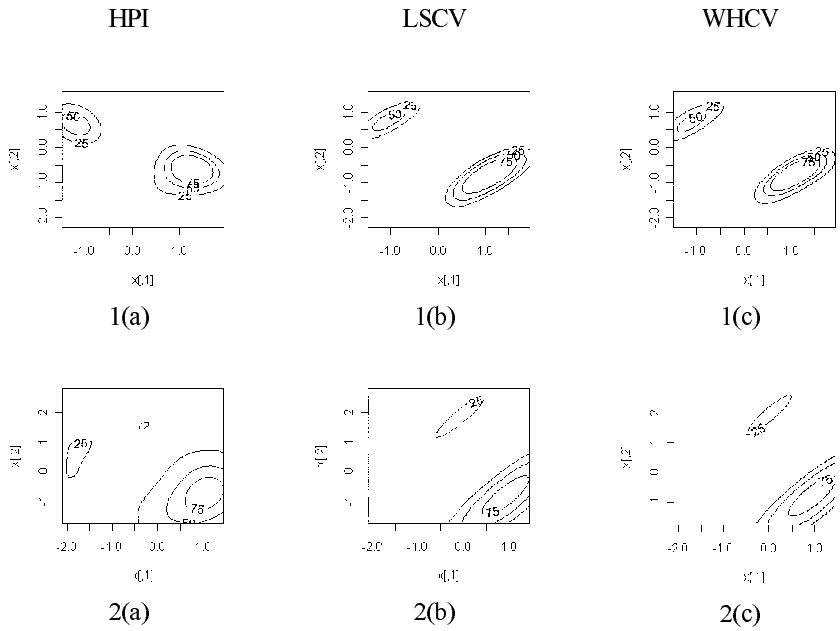

Figure 7: Density C: small sample.

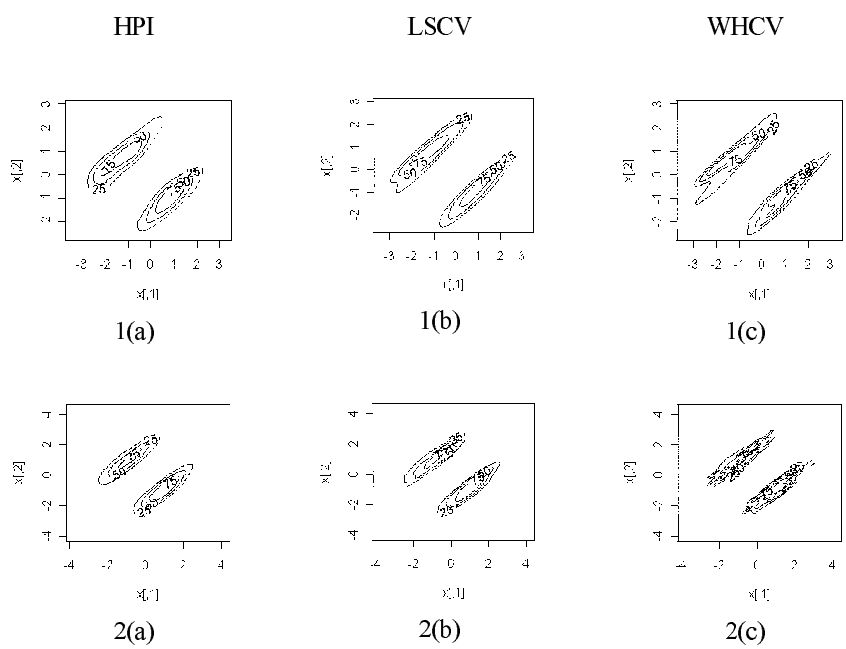

Figure 8: Density C: large sample.

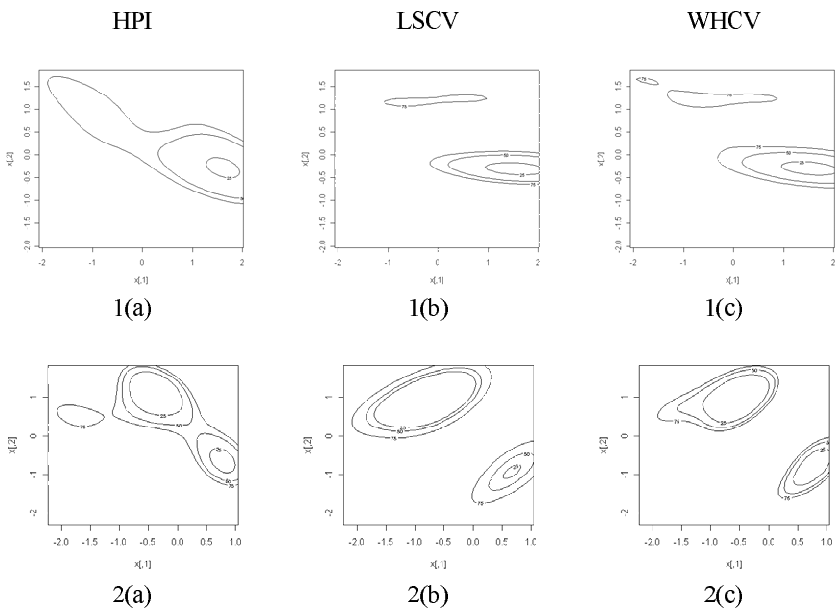

Figure 9: Density D: small sample.

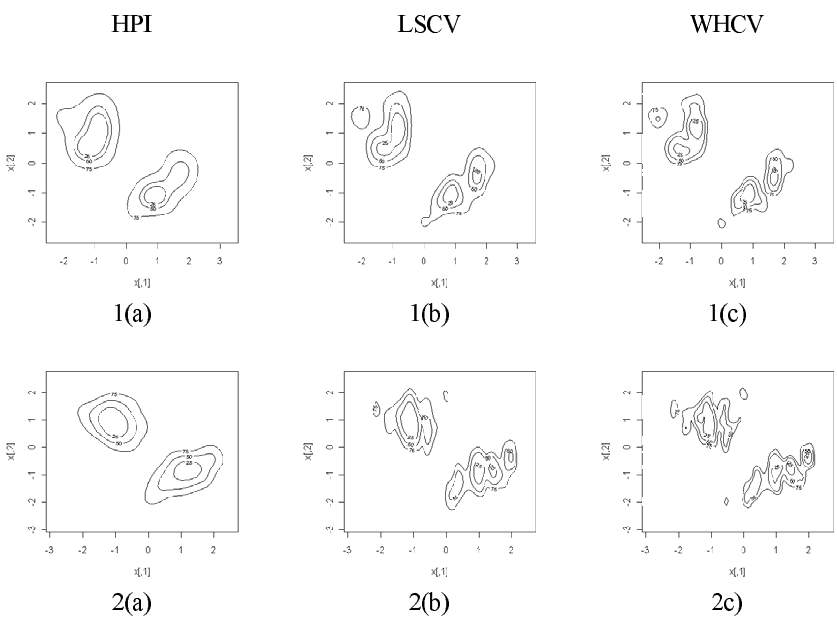

Figure 10: Density D: large sample.

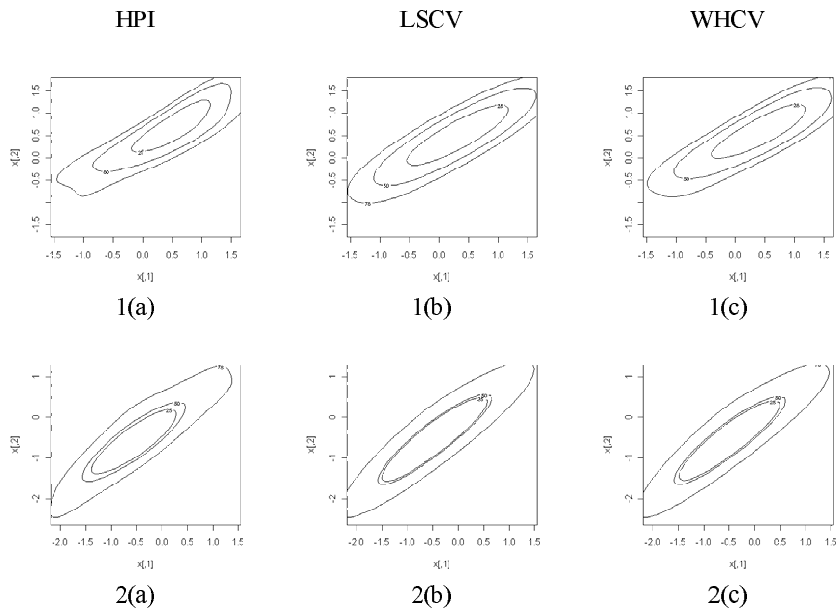

Figure 11: Density E: small sample.

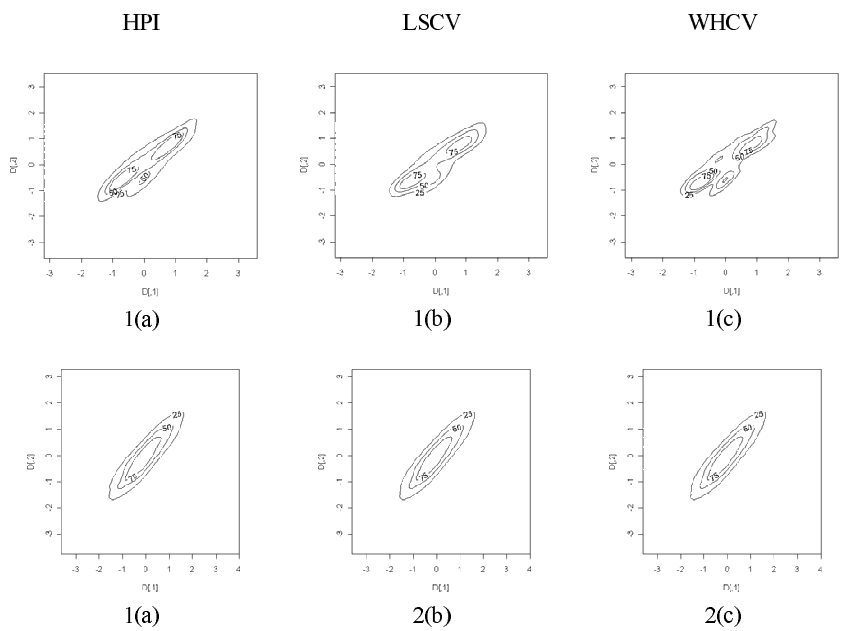

Figure 12: Density E: large sample.

Figures 3 to 12 give the kernel density estimates using the bandwidth selectors HPI, LSCV and WHCV. As might be expected, no method is uniformly best for all the densities. Both LSCV and WHCV cannot compete with the performance of HPI for increasing sample sizes. HPI gives a better estimate for all the densities for large samples. We see that both CV selectors give very poor estimate of density B. It is known that CV techniques do not perform well for duplicated data values, which cannot be the case for poor density B estimate. For small sample size we can see that for densities A, B and D signs of oversmoothing is displayed with all three bandwidth selectors while for densities E and C the bimodality is preserved for all three selectors. In the case of density E all three bandwidth selectors give a better estimate for small sample whereas for density A, the performance is poor. Computation times were not taken in to consideration because of computing resources. However HPI running time was much lesser than the other two CV selectors.

3.2. Bivariate real-life data analysis

We now turn to the analysis of data on under-5 mortality (per 1000 live births) and average life expectancy (in years) for 73 countries with Gross National Income $<$ \$1000 per person per year. The data were obtained from Unicef (United Nations Children's Fund). A scatter plot of the Unicef data is given in Figure 13. This dataset has probability mass oriented to the axes. Since the dataset contains duplicates, LSCV's estimate was very poor. Therefore we used the smooth cross validation bandwidth selectors (SCV) as our second choice and the density estimates obtained using Plug-in, SCV and WHCV are displayed in Figure 14. The plot for SCV has spurious features and the most noisy whereas the plots for HPI and WHCV are smoother, hence giving better estimates.

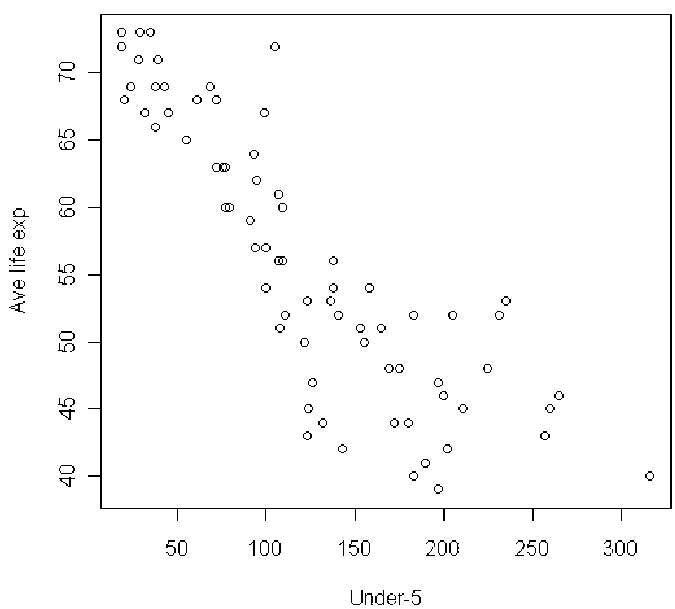

Figure 13: Unicef-data Scatterplot.

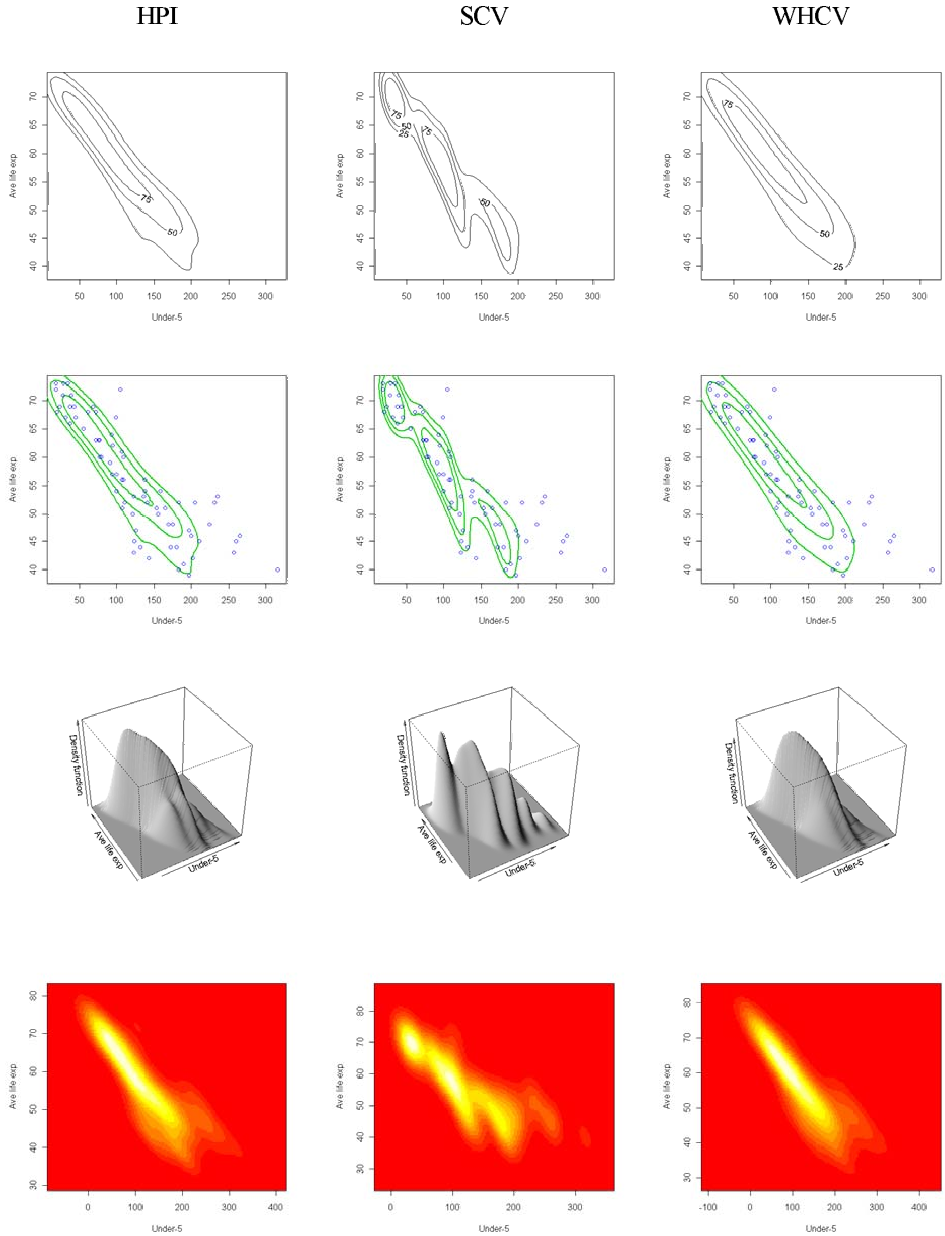

Figure 14: Unicef Plots for Bandwidth Selectors.

4. Conclusion

The selection of full bandwidth matrices for multivariate density estimation raises issues that have no univariate counterpart. In particular, the orientation of the kernel functions to the coordinate axes must be determined. Furthermore, intuition drawn from the univariate setting cannot necessarily be transformed to the multivariate case. The additional difficulties involved in the multivariate case generate significant scope for further research in full bandwidth matrix selection. Looking further afield, the use of adaptive kernel density estimators has great potential for multivariate data. While some progress has been made, this remains a challenging research direction.

References

AHMAD, I. A. AND MUGDADI, A. R. (2006). Weighted Hellinger distance as an error criterion for bandwidth selection in kernel estimation. *Journal of Nonparametric Statistics*, **18** (2), 215–226.

ANVER, H. (2010). *Mean Hellinger Distance as an Error Criterion in Univariate and Multivariate Kernel Density Estimation*. Ph.D. dissertation, Southern Illinois University: Carbondale, U.S.A.

DUONG, T. AND HAZELTON, M. (2003). Plug-in bandwidth matrices for bivariate kernel density estimation. *Journal of Nonparametric Statistics*, **15** (1), 17–30.

KANZAWA, Y. (1995). Hellinger distance and Kullback-Leibler loss for the kernel density estimator. *Statistics and Probability Letters*, **18**, 315–321.

MUGDADI, A. R. (2004). Bandwidth selection for the function of observations using Hellinger distance. *Journal of Applied Statistical Science*, **13** (3), 231–240.

POWELL, M. J. D. (1964). An efficient method for finding the minimum of a function of several variables without calculating derivatives. *The Computer Journal*, **7** (2), 155–162.

SCOTT, D. W. (1992). *Multivariate Kernel Density Estimation: Theory, Practice and Visualization*. John Wiley & Sons Inc.: New York.

STONE, C. J. (1984). An asymptotically optimal window selection rule for kernel density estimates. *The Annals of Statistics*, **12**, 1285–1297.

WAND, M. P. AND JONES, M. C. (1994). Multivariate plug-in bandwidth selection. *Computational Statistics*, **9**, 97–116.

WAND, M. P. AND JONES, M. C. (1995). *Kernel Smoothing*. Chapman and Hall: New York.

Manuscript received, 2015-11-23, revised, 2016-01-29, accepted, 2016-05-31.