Contribution to Bandwidth Matrix Choice for Multivariate Kernel Density Estimate

Horová, Ivanka
Masaryk University, Department of Mathematics and Statistics
Kotlářská 2
Brno 611 37, Czech Republic
E-mail: horova@math.muni.cz

Koláček, Jan
Masaryk University, Department of Mathematics and Statistics
Kotlářská 2
Brno 611 37, Czech Republic
E-mail: kolacek@math.muni.cz

Vopatová, Kamila
University of Defence, Department of Econometrics
Kounicova 65
Brno 662 10, Czech Republic
E-mail: 63985@mail.muni.cz

Zelinka, Jiří
Masaryk University, Department of Mathematics and Statistics
Kotlářská 2
Brno 611 37, Czech Republic
E-mail: zelinka@math.muni.cz

Introduction

The most important factor in the multivariate kernel density estimate is a choice of the bandwidth matrix. Because of its role in controlling both the amount and the direction of multivariate smoothing, this choice is particularly important. Most of popular bandwidth selection methods in a univariate case (see e.g. Scott (1992), Wand and Jones (1995)) can be transferred into multivariate settings (Duong and Hazelton (2005), Chacón and Duong (2010)). Also a special iterative method proposed by Horová and Zelinka (2007) has been extended to bivariate case with diagonal bandwidth matrix (Horová et al. (2008) and Horová et al. (2010)). The advantage of the proposed method consists in the fact that it does not need any pre-transformation of the data. The present paper focuses on a d-variate case and a full bandwidth matrix.

Univariate kernel density estimation

Density derivative estimation

Let $X_1, \ldots, X_n$ be independent real random variables having the same density f. The basic kernel estimate of the ν-th derivative with a kernel K at the point $x \in \mathbf{R}$ can be written as

$$\hat{f}^{(\nu)}(x,h) = \frac{1}{nh^{\nu+1}} \sum_{i=1}^{n} K^{(\nu)} \left(\frac{x - X_i}{h} \right),$$

where K is a kernel and $h > 0$ is a smoothing parameter called also a *bandwidth*.

First, we make some assumptions and notations:

- $\lim_{n\to\infty} h = 0,\ \lim_{n\to\infty} nh^{2\nu+1} = \infty,\ 0 \le \nu$
- $f \in C^{k_0},\ \nu + k \le k_0,\ 0 \le \nu < k,\ \nu,\ k$ are nonnegative integers
- $K \in C^{\nu}[-1,1],\ K^{(j)}(-1) = K^{(j)}(1) = 0,\ j = 0, 1, \ldots, \nu-1,\ V(g) = \int g^2(x)dx$
- $S_{\nu k}$ is a class of real valued functions on $\mathbf{R}$ satisfying conditions

$$(i)\ \ \text{support}(K) = [-1,1]$$

$$(ii)\ \ \int_{-1}^{1} x^j K(x)dx = \begin{cases} 0, & 0 \le j < k, j \ne \nu \\ (-1)^{\nu}\nu!, & j = \nu \\ \beta_k \ne 0, & j = k. \end{cases}$$

Hereinafter, it is assumed $K \in S_{0k} \cap C^{\nu}[-1,1]$, i.e. $K^{(\nu)} \in S_{\nu,k+\nu}$.

We consider Mean Integrated Square Error as a criterion of the quality of the estimate

$$\text{MISE}\{\hat{f}^{(\nu)}(\cdot,h)\} = E\int\{\hat{f}^{(\nu)}(x,h) - f^{(\nu)}(x)\}^2 dx.$$

Since MISE is not mathematically tractable, we employ an asymptotic mean integrated square error (AMISE) which can be written as a sum of an asymptotic integrated variance and asymptotic integrated square bias

$$\text{AMISE}\{\hat{f}^{(\nu)}(\cdot,h)\} = \underbrace{\frac{V(K^{(\nu)})}{nh^{2\nu+1}}}_{AIvar\, f^{(\nu)}} + \underbrace{h^{2k}\frac{\beta_{k+\nu}^2}{(k+\nu)!^2}V(f^{(k+\nu)})}_{AIbias^2\, f^{(\nu)}} \tag{1}$$

and the optimal bandwidth minimizing AMISE, $h_{opt,\nu,k} = \arg\min \text{AMISE}\{\hat{f}^{(\nu)}(\cdot,h)\}$, takes the form

$$h_{opt,\nu,k}^{2(k+\nu)+1} = \frac{(2\nu+1)V(K^{(\nu)})}{2(k-\nu)nV(f^{(k+\nu)})}\frac{(k+\nu)!^2}{\beta_{k+\nu}^2}. \tag{2}$$

Thus $h_{opt,\nu,k} = O\left(n^{-1/2(k+\nu)+1}\right)$ and $\text{AMISE}\{\hat{f}^{(\nu)}(\cdot,h)\} = O\left(n^{-4/2(k+\nu)+1}\right)$.

Choosing of the optimal bandwidth

The optimal bandwidth minimizing AMISE depends on the unknown density f and thus we turn our attention to data driven bandwidth matrix selectors. Härdle et al. (1990) have proposed the modified cross-validation for the first derivative estimate. The objective function is defined as

$$CV_{(\nu)}(h) = \int(\hat{f}^{(\nu)}(x,h))^2 dx - 2\frac{(-1)^{\nu}}{n}\sum_{i=1}^{n}\hat{f}_{-i}^{(2\nu)}(X_i,h),$$

where $\hat{f}_{-i}^{(2\nu)}$ is the estimate of the 2ν-th derivative of f at the point X_i without using this point.
We are going to extend the iterative method proposed in paper Horová and Zelinka (2007) to the estimate of ν-th derivative. This method is also based on a suitable estimate of $\text{MISE}\{\hat{f}^{(\nu)}(\cdot,h)\}$ and on the fact that

$$\frac{2\nu+1}{2k}AIvar\hat{f}^{(\nu)}(\cdot,h_{opt,\nu,k}) = AIbias^2\hat{f}^{(\nu)}(\cdot,h_{opt,\nu,k}) \tag{3}$$

Further, we use the estimates of quantities at this equality

$$AI\widehat{var}\,\hat{f}^{(\nu)} = \frac{1}{nh^{2\nu+1}}V(K^{(\nu)})$$

and

$$\begin{aligned} AI\widehat{bias}^{\,2}\,\hat{f}^{(\nu)} &= \int\left(\int K(x)\hat{f}^{(\nu)}(x-hy,h)dy - \hat{f}^{(\nu)}(x,h)\right)^2 dx \\ &= \frac{1}{n^2h^{2\nu+1}}\sum_{i,j=1}^{n}\Lambda^{(2\nu)}\left(\frac{X_i-X_j}{h}\right), \end{aligned}$$

where

$$\Lambda^{(2\nu)}(z) = \frac{\partial^{2\nu}\Lambda^{(0)}(z)}{\partial z^{2\nu}}$$

and

$$\Lambda^{(0)}(z) = \Lambda(z) = (K*K*K*K - 2K*K*K + K*K)(z)$$

Consider equation (3) in the form

$$\frac{2\nu+1}{nh^{2\nu+1}}V(K^{(\nu)}) - 2k\frac{1}{n^2h^{2\nu+1}}\sum_{i,j=1}^{n}\Lambda^{(2\nu)}\left(\frac{X_i-X_j}{h}\right) = 0. \tag{4}$$

Let $\hat{h}_{IT,\nu,k}$ be a solution of this equation. The statistical properties of this estimate are given in the following theorem.

Theorem 1.

$$\begin{aligned} &E(I\widehat{bias}^{\,2}\,\hat{f}^{(\nu)}) = Ibias^2\hat{f}^{(\nu)} + \tfrac{1}{nh^{2\nu+1}}\Lambda^{(2\nu)}(0) + o(h^{2k}) \\ &var(I\widehat{bias}^{\,2}\,\hat{f}^{(\nu)}) = \tfrac{8k^2}{n^2h^{4\nu+1}}V\left(\Lambda^{(2\nu)}\right)V(f) + o(h^{4k}+n^{-2}h^{-(4\nu+1)}). \end{aligned}$$

Corollary.

$$\frac{\hat{h}_{IT,\nu,k}}{h_{opt,\nu,k}} \xrightarrow{P} 1.$$

Multivariate kernel density estimation

Let a d-variable random sample $\mathbf{X}_1,\ldots,\mathbf{X}_n$ come from distribution with a density f. The kernel density estimator $\hat{f}$ is defined

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

H is a symmetric positive definite $d\times d$ matrix called the bandwidth matrix, where $|H|$ stands for the determinant of H. The kernel function K is often taken to be a d-variable probability density function satisfying $\int\limits_{\mathbf{R}^d}K(\mathbf{x})d\mathbf{x}=1$, $\int\limits_{\mathbf{R}^d}\mathbf{x}K(\mathbf{x})d\mathbf{x}=0$, $\int\limits_{\mathbf{R}^d}\mathbf{x}\mathbf{x}^TK(\mathbf{x})d\mathbf{x}=\beta_2I_d$, I_d is an identity matrix and $\mathbf{x}=(x_1,\ldots,x_d)^T$ is a generic vector.
We make some additional assumptions:

- $H=H_n$ is a sequence of bandwidth matrices such that $n^{-1}|H|^{-1/2}$ and all entries of H approach zero as $n\to\infty$.

- $D^{\otimes r} f(\mathbf{x})$ is the vector containing all partial derivatives of the order r of f at $\mathbf{x}$, i.e. if $f : \mathbf{R}^d \to \mathbf{R} \Rightarrow D^{\otimes r} f(\mathbf{x}) \in \mathbf{R}^{d^r}$, $D^{\otimes 1} f(\mathbf{x}) = Df(\mathbf{x})$.
- Every component of $D^{\otimes 2} f(\mathbf{x})$ is bounded, continuous and square integrable.
- $V(g) = \int g(\mathbf{x}) g(\mathbf{x})^T d\mathbf{x}$ for any square integrable vector valued function g.
- vec$H$ is $d^2 \times 1$ vector obtained by stacking from left to right columns of H.

The quality of the estimate can be expressed by means of the asymptotic mean integrated error

$$\mathrm{AMISE}\{\hat{f}(\cdot, H)\} = \underbrace{n^{-1}|H|^{-1/2} V(K)}_{AIvar\, f} + \underbrace{\frac{\beta_2^2}{4} \mathrm{vec}H^T V(D^{\otimes 2} f) \mathrm{vec}H}_{AIbias^2\, f}.$$

The last equation can be rewritten as

$$\mathrm{AMISE}\{\hat{f}(\cdot, H)\} = n^{-1}|H|^{-1/2} V(K) + \frac{\beta_2^2}{4} \int \mathrm{tr}^2(HD^2 f(\mathbf{x})) d\mathbf{x},$$

where tr$A$ is the trace of A and $D^2 f(\mathbf{x}) = Df(\mathbf{x}) D^T f(\mathbf{x})$.

Choice of optimal bandwidth

The optimal bandwidth matrix is defined as

$$H_{AMISE} = \arg\min \mathrm{AMISE}\{\hat{f}(\cdot, H)\}.$$

Unfortunately, it does not exist any explicit solution of the equation $\frac{\partial \mathrm{AMISE}(H)}{\partial \mathrm{vec}H} = \mathbf{0}$. But the following relation holds

$$AIvar(H_{AMISE}) = \frac{4}{d} AIbias^2(H_{AMISE}) \tag{5}$$

Remark. $H_{AMISE} = O(J_d\, n^{-2/(d+4)})$, where J_d is $d \times d$ matrix of ones and $AMISE(H_{AMISE}) = O(n^{-4/(d+4)})$ (see e.g. Chacón and Duong (2010)).

Consider an estimate $\widehat{Ibias}^2$ in the form

$$\widehat{Ibias}^2 = \frac{1}{n^2} \sum_{i,j=1}^{n} \Lambda_H(\mathbf{X}_i - \mathbf{X}_j),$$

where

$$\Lambda_H(\mathbf{z}) = (K_H * K_H * K_H * K_H - 2K_H * K_H * K_H + K_H * K_H)(\mathbf{z}).$$

Now, instead of solving equation (5) we are dealing with equation

$$n^{-1}|H|^{-1/2} V(K) d - 4n^{-2} \sum_{i,j=1}^{n} \Lambda_H(\mathbf{X}_i - \mathbf{X}_j) = 0. \tag{6}$$

Let us assume that the kernel K is standard normal density, i.e. $K = \Phi_I$ (i.e. $\beta_2 = 1$).
Let H_{IT} stand for solution of the equation (6). If we proceed the similar way in univariate case, we can show that the following theorem holds:

Theorem 2.

$$E(\widehat{Ibias}^2 \hat{f}) = Ibias^2 \hat{f} + n^{-1}|H|^{-1/2} \Lambda_H(0) + o(||\mathrm{vec}H||^2)$$
$$var(\widehat{Ibias}^2 \hat{f}^{(\nu)}) = 2n^{-2}|H|^{-1/2} V(\Lambda) V(f) + o(||\mathrm{vec}H||^2 + n^{-2}|H|^{-1/2}).$$

Computations

The equation (6) can be rewritten as

$$|H|^{1/2} = \frac{n^2 V(K) d}{4 \sum_{i,j=1}^{n} \Lambda_H(\mathbf{X}_i - \mathbf{X}_j)}.$$

This is a nonlinear equation for $d' = \frac{1}{2}d(d+1)$ unknown entries of H. We can adopt Scott's idea and assume

$$\hat{h}_{ij} = \hat{\sigma}_{ij} n^{-1/(d+4)},\ i,j = 1,\dots,d \Rightarrow \hat{h}_{ij} = \frac{\hat{\sigma}_{ij}}{\hat{\sigma}_{11}} \hat{h}_{11},\ i = 2,\dots,d,\ j = 1,\dots,i,$$

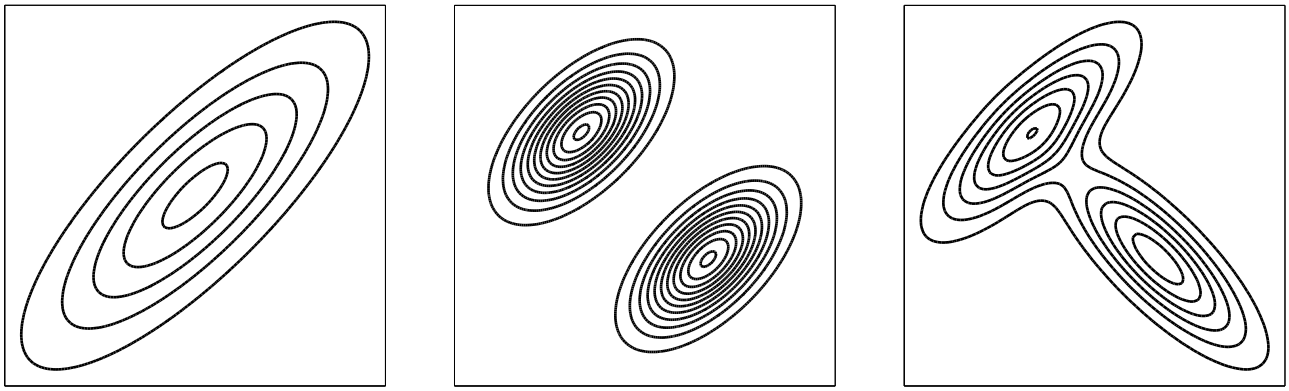
then we obtain one nonlinear equation for $\hat{h}_{11}$ which can be solved by means of a suitable method.

Simulation

We run a short simulation study to verify the quality of the proposed method. A performance of the bandwidth matrix can be easily measured by the integrated square error (ISE)

$$ISE(H) = \int_{\mathbf{R}^2} \left[\hat{f}(\mathbf{x}, H) - f(\mathbf{x})\right]^2 d\mathbf{x}.$$

We drew samples of the size $n = 100$ from each density and selected bandwidth matrices for 100 random samples generated from each density.



Contour plot of training densities.

Density	Formula
I	$N_2([0,0],[1,4/5,4/5,1])$
II	$0.5 \cdot N_2([-1,1],[4/9,12/45,12/45,4/9])$
	$+0.5 \cdot N_2([1,-1],[4/9,12/45,12/45,4/9])$
III	$0.5 \cdot N_2([0,0],[1/5,4/25,4/25,1/5])$
	$+0.5 \cdot N_2([1,-1],[1/5,-4/25,-4/25,1/5])$

Training densities.

AMISE-optimal bandwidth matrices:

$$\begin{aligned}
\operatorname{vec} H_I &= (0.2171, 0.1736, 0.1736, 0.2171)^T, \\
\operatorname{vec} H_{II} &= (0.1219, 0.0736, 0.0736, 0.1219)^T, \\
\operatorname{vec} H_{III} &= (0.0490, 0, 0, 0.0490)^T.
\end{aligned}$$

The following table summarizes the average of the ISE, where the average was taken over simulated realizations.

Density	$ISE(H_{IT})$	$ISE(H_{AMISE})$
I	0.0093 (0.0042)	0.0512 (0.0034)
II	0.0174 (0.0038)	0.0256 (0.0036)
III	0.0583 (0.0111)	0.0738 (0.0084)

ISE – The average with a standard deviation in parentheses.

Acknowledgements

The research was supported by The Jaroslav Hájek Center for Theoretical and Applied Statistics (MŠMT LC 06024). K. Vopatová has been supported by the University of Defence through an "Institutional development project UO FEM – Economics Laboratory" project.

REFERENCES

Chacón, J.E. and Duong, T. (2010). Multivariate plug-in bandwidth selection with unconstrained pilot bandwidth matrices. *Test*, **19**, pp. 375–398.

Duong, T. and Hazelton, M.L (2005). Cross-validation bandwidth matrices for multivariate kernel density estimation. *Scandinavian Journal of Statistics*, **32**, 3, pp. 485–506.

Härdle, W., Marron, J. S. and Wand, M. P. (1990). Bandwidth choice for density derivatives. *Journal of the Royal Statistical Society. Series B (Methodological)*, **52**, 1, pp. 223–232.

Horová, I., Koláček, J., Zelinka, J. and Vopatová, K. (2008). Bandwidth Choice for Kernel Density Estimates. In *Proceedings IASC, Yokohama : IASC, 2008*, pp. 542–551.

Horová, I., Koláček, J. and Vopatová, K. (2010). Visualization and Bandwidth Matrix Choice. To appear in *Communications in Statistics, Theory and Methods.*

Horová, I. and Zelinka, J. (2007). Contribution to the bandwidth choice for kernel density estimates, *Computational Statistics*, **22**, 1, pp. 31–47.

Scott, D.W. (1992). *Multivariate density estimation: Theory, practice, and visualization.* Wiley series in probability and mathematical statistics: Applied probability and statistics (Wiley & sons).

Wand, M.P. and Jones, M.C. (1995). *Kernel smoothing.* Chapman and Hall, London.

Keywords

multivariate density, bandwidth matrix, kernel estimate