ACTA UNIVERSITATIS MATTHIAE BELII, series MATHEMATICS 18 (2011), 47–53
Also available at http://actamath.savbb.sk
ISBN 978-80-557-0160-8

Kernel choice with respect to the bandwidth in kernel density estimates*

Kamila Vopatová
Department of Mathematics and Statistics,
Faculty of Science, Masaryk University, Kotlářská 2
637 11 Brno, Czech Republic
E-mail: vopatova@mail.muni.cz

Abstract

Kernel density estimates belong to the most popular nonparametric density estimates. It is well known that these estimates depend on a bandwidth, which controls the smoothness of the estimate, and on a kernel, which plays a role of weight function.
We focus on the kernel function choice, especially on kernels with bounded supports. Our aim is to study the kernel optimality with respect to the bandwidth choice. In a simulation we show a comparison of the kernels. We propose that the cosine kernel may be a good alternative to the frequently used Epanechnikov kernel.

Keywords Kernel, bounded support, density estimation, cross-validation.

MSC (2010) 62G07, 30C40

Received: July 29, 2010; Revised: August 28, 2010; Accepted: August 30, 2010.

1 Kernel density estimation

The concept of nonparametric estimates was introduced in the fifties and sixties, see e.g. [3, 5] and references therein. After years, there are still problems to be solved or to be improved.

Let a d-variate random sample $\mathbf{X}_1, \dots, \mathbf{X}_n$ come from distribution with a density f. The kernel density estimator $\hat{f}$ is defined as a weighted average of observations

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

K is a d-variate kernel function, which is often taken to be a probability density function satisfying $\int_{\mathbb{R}^d} K(\mathbf{x})\,\mathrm{d}\mathbf{x} = 1$, where we omit the subscript $\mathbb{R}^d$ in the rest of the text. H is a symmetric positive definite $d \times d$ matrix called a bandwidth matrix and $\mathbf{x} = (x_1, \dots, x_d)^T \in \mathbb{R}^d$ is a generic vector.

Transfer of the kernel estimation from univariate settings to the multivariate settings brings a few new issues. The first problem is to find a kernel and the second one is to find an optimal bandwidth. There are two common ways to create the multivariate kernel K from the univariate kernel k: the product kernel $K^P(\mathbf{x}) = \prod_{i=1}^{d} k(x_i)$

*This research was supported by Masaryk University under the Student Project Grant MUNI/A/1001/2009.

and the spherically symmetric kernel $K^S(\mathbf{x}) = c_k^{-1}k((\mathbf{x}^T\mathbf{x})^{1/2})$, $c_k = \int k(\sqrt{\mathbf{x}^T\mathbf{x}})\ \mathrm{d}\mathbf{x}$. Let us denote the class of symmetric positive definite $d \times d$ matrices as $\mathcal{H}_\mathcal{F}$. One needs to choose $d(d+1)/2$ distinct entries of matrix $H \in \mathcal{H}_\mathcal{F}$, which is computationally intensive. On the other hand, using a single parameter simplification, i.e. $H = h^2 \cdot I_d$ (I_d is a $d \times d$ identity matrix), is not advised for data which have different dispersions in the co-ordinate directions, see [4]. Thus, the diagonal matrix class $\mathcal{H}_\mathcal{D}$ ($H = \mathrm{diag}(h_1^2, \ldots, h_d^2)$) seems to be a compromise between computational speed and sufficient flexibility.

Mean integrated square error (MISE) quantifies the performance of a multivariate kernel density estimator. MISE can be rewritten as a sum of an integrated variance and an integrated square bias

$$\mathrm{MISE}(H) = E\int \left[\hat{f}(\mathbf{x}, H) - f(\mathbf{x})\right]^2\ \mathrm{d}\mathbf{x} = \int \mathrm{Var}\,\hat{f}(\mathbf{x}, H)\ \mathrm{d}\mathbf{x} + \int \mathrm{Bias}^2\,\hat{f}(\mathbf{x}, H)\ \mathrm{d}\mathbf{x}.$$

It is easy to see, that finding the bandwidth matrix H_{MISE}, which minimizes this error, is very difficult. Wand and Jones [5] derived, under some assumptions on the density f, the kernel function K, and the bandwidth matrix H, the asymptotic mean integrated square error

$$\mathrm{AMISE}(H) = n^{-1}|H|^{-1/2}V(K) + \tfrac{1}{4}\beta_2(K)^2(\mathrm{vech}\,H)^T\Psi_4(\mathrm{vech}\,H).$$

$V(K) = \int K^2(\mathbf{x})\ \mathrm{d}\mathbf{x}$, $\beta_2(K) = \int x_i^2K(\mathbf{x})\ \mathrm{d}\mathbf{x}$ is independent of i and vech is a vector half operator, i.e. for a matrix M, $\mathrm{vech}\,M$ is a $d(d+1)/2 \times 1$ vector of stacked columns of the lower triangular matrix of M. The matrix Ψ_4 includes entries depending on the unknown density f. For a d-variate function g and for a vector $\mathbf{r} = (r_1, \ldots, r_d)$ of non-negative integers, $g^{(\mathbf{r})}$ is defined

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

assuming that the derivative exists. The notation $|\mathbf{r}|$ is for the sum of the components of the vector $\mathbf{r}$. Each entry of Ψ_4 can be written, under sufficient conditions, in the form $\psi_\mathbf{r} = \int f^{(\mathbf{r})}(\mathbf{x})f(\mathbf{x})\ \mathrm{d}\mathbf{x}$, where $|\mathbf{r}|$ is even (see Section 4.3 in [5]).

2 Simulation study

2.1 Cross-validation methods

There is a wide range of methods for the optimal bandwidth choice, thus we aimed our study to cross-validation (CV) methods.

The most widely used cross-validation method is the least square cross-validation method (LSCV) [5].

$$\mathrm{LSCV}(H) = \int \left(\hat{f}(\mathbf{x}, H)\right)^2\ \mathrm{d}\mathbf{x} - \frac{2}{n}\sum_{i=1}^{n}\hat{f}_{-i}(\mathbf{X}_i, H),$$

is the LSCV objective function, where $\hat{f}_{-i}(\mathbf{X}_i, H) = (n-1)^{-1}\sum_{j=1, j\neq i}^{n} K_H(\mathbf{X}_i - \mathbf{X}_j)$ is a leave-one-out estimator of f. LSCV function can be written in terms of convolutions $(f * g)(x) = \int_\mathbb{R} f(t)g(x-t)\ \mathrm{d}t$ (see e.g. [1])

$$\mathrm{LSCV}(H) = n^{-1}(n-1)^{-1}\sum_{i=1}^{n}\sum_{\substack{j=1\\ i\neq j}}^{n}(K_H * K_H - 2K_H)(\mathbf{X}_i - \mathbf{X}_j) + n^{-1}V(K)|H|^{-1/2}.$$

LSCV is also called the unbiased cross-validation, because the LSCV(H) function is unbiased in the sense that $E[\mathrm{LSCV}(H)] = \mathrm{MISE}(H) - \int f^2(\mathbf{x})\,\mathrm{d}\mathbf{x}$.

Biased cross-validation (BCV) method estimates AMISE, i.e. BCV is a biased estimate of MISE. There are two types of BCV, depending on the way of estimating functionals $\psi_{\mathbf{r}}$ [1]:

$$\mathrm{BCV}_1(H) = n^{-1}V(K)|H|^{-1/2} + \frac{1}{4}\beta_2(K)^2(\operatorname{vech} H)^T\hat{\Psi}_4(\operatorname{vech} H),$$

where

$$\hat{\psi}_{\mathbf{r}} = n^{-2}\sum_{i=1}^{n}\sum_{\substack{j=1,\\ j\neq i}}^{n}\big(K_H^{(\mathbf{r})} * K_H\big)(\mathbf{X}_i - \mathbf{X}_j).$$

The latter is defined by the function

$$\mathrm{BCV}_2(H) = n^{-1}V(K)|H|^{-1/2} + \frac{1}{4}\beta_2(K)^2(\operatorname{vech} H)^T\tilde{\Psi}_4(\operatorname{vech} H),$$

with

$$\tilde{\psi}_{\mathbf{r}} = n^{-1}\sum_{i=1}^{n}\hat{f}_{-i}^{(\mathbf{r})}(\mathbf{X}_i, H) = n^{-1}(n-1)^{-1}\sum_{i=1}^{n}\sum_{\substack{j=1,\\ j\neq i}}^{n}K_H^{(\mathbf{r})}(\mathbf{X}_i - \mathbf{X}_j).$$

2.2 Kernels

We focused on studying the product kernels with bounded supports. Due to a wide range of the class of the kernels with bounded supports $\mathcal{D}$: $\{[x_1, x_2] \in \mathbb{R}^2 : |x_1| \leq 1 \wedge |x_2| \leq 1\}$, we selected the easiest five two-dimensional kernels listed in Table 1. Their one-dimensional representations are displayed in Figure 1.

	$K(x_1,x_2)$	$V(K)$	$\beta_2(K)$
Uniform	$1/4$	$1/4$	$1/3$
Epanechnikov	$9/16(1-x_1^2)(1-x_2^2)$	$9/25$	$1/5$
Biweight	$225/256(1-x_1^2)^2(1-x_2^2)^2$	$25/49$	$1/7$
Cosine	$\pi^2/16\cos(x_1\pi/2)\cos(x_2\pi/2)$	$(\pi/4)^4$	$1-8/\pi^2$
Triangular	$(1-|x_1|)(1-|x_2|)$	$4/9$	$1/6$

Table 1: Two-dimensional product kernels with bounded supports.

Remark 2.1. Epanechnikov kernel is the optimal kernel in the AMISE sense, i.e. Epanechnikov kernel minimizes the functional $T(K) = [V(K)\beta_2(K)]^{2/3}$ (see [5]). From $\mathrm{eff}(K) = [T(K)/T(K_{Epan})]^{3/2}$ is apparent that the cosine kernel is a convenient alternative to Epanechnikov kernel.

2.3 Densities

We drew samples of the size $n = 50$ and $n = 100$ from densities listed in Table 2. Contour plots of target densities are displayed in Figure 2. One hundred replications for each of the sample sizes and for each of the densities were generated.

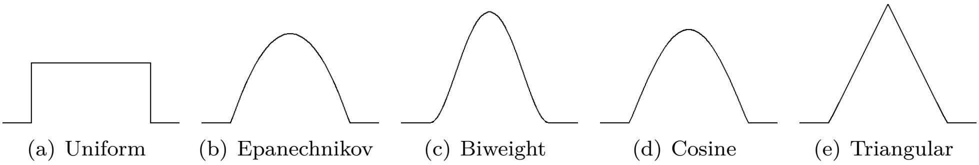

(a) Uniform (b) Epanechnikov (c) Biweight (d) Cosine (e) Triangular

Figure 1: One-dimensional kernels with bounded supports.

(A)	Normal	$N_2(0,0;4,1,0)$
(B)	Student	$t_2(4)$
(C)	Weibull	$W(2,2)\cdot W(2,4)$
(D)	Exponential	$Exp(2)\cdot Exp(1)$
(E)	Gamma Student	$Gamma(2,1)\cdot t(5)$
(F)	Lognormal	$LN_2(0,0;1,1,0)$

Table 2: Target densities.

2.4 Comparative criteria

In the case of the diagonal AMISE-optimal bandwidth matrix $H_{\mathrm{AMISE}} = \operatorname{diag}(h^2_{1,\mathrm{A}}, h^2_{2,\mathrm{A}})$, we can express its entries [5]

$$h_{1,\mathrm{A}} = \left[\frac{\psi_{04}^{3/4} V(K)}{\beta_2(K)^2 \psi_{40}^{3/4}(\psi_{22} + \psi_{04}^{1/2}\psi_{40}^{1/2})n}\right]^{1/6},$$

$$h_{2,\mathrm{A}} = \left[\frac{\psi_{40}^{3/4} V(K)}{\beta_2(K)^2 \psi_{04}^{3/4}(\psi_{22} + \psi_{04}^{1/2}\psi_{40}^{1/2})n}\right]^{1/6}.$$

We used two criteria to decide which kernel fits best the chosen method: the average of squared Euclidean norm of difference vectors

$$\mathrm{ED} = \operatorname{avg}_H \|(\hat{h}_1 - h_{1,\mathrm{A}}, \hat{h}_2 - h_{2,\mathrm{A}})^T\|_2^2$$

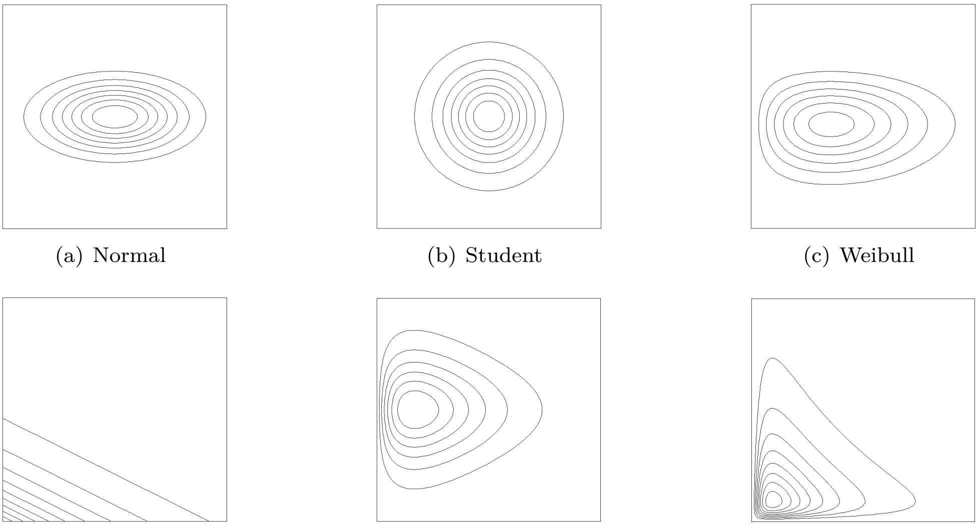
and the average of integrated square errors

$$\mathrm{ISE} = \operatorname{avg}_H \int \left[\hat{f}(\mathbf{x}, H) - f(\mathbf{x})\right]^2 \,\mathrm{d}\mathbf{x},$$

where the average is taken over simulated realizations. ED can be considered as a visual criterion and ISE, which was computed numerically, can be viewed as a numerical criterion.

3 Results

We compared a performance of kernels within each of the mentioned cross-validation methods. We selected bandwidth matrices — we computed values of $\min_{H\in\mathcal{H}_\mathcal{D}} \mathrm{LSCV}(H)$, $\min_{H\in\mathcal{H}_\mathcal{D}} \mathrm{BCV}_1(H)$, and $\min_{H\in\mathcal{H}_\mathcal{D}} \mathrm{BCV}_2(H)$ on a dense grid numerically. For a faster computation, we used ideas by Horová et al. [2]. Tables 3 and 4 summarize results of ED criterion and Tables 5 and 6 summarize results of ISE criterion.



(a) Normal (b) Student (c) Weibull

(d) Exponential (e) Gamma Student (f) Lognormal

Figure 2: Contour plots of target densities.

ED		LSCV				
density	n	Uni	Epan	Biw	Cos	Tri
(A)	50	0.76(0.07)	0.69(0.05)	0.67(0.06)	1.82(0.16)	4.24(0.28)
	100	0.85(0.05)	0.76(0.04)	0.58(0.04)	2.01(0.12)	4.97(0.21)
(B)	50	0.86(0.04)	0.88(0.05)	0.82(0.05)	1.21(0.08)	2.75(0.17)
	100	0.92(0.02)	0.90(0.02)	0.81(0.02)	1.16(0.05)	2.61(0.09)
(C)	50	0.80(0.02)	0.81(0.02)	0.75(0.02)	1.62(0.07)	3.32(0.12)
	100	0.95(0.02)	1.00(0.02)	0.95(0.02)	1.84(0.05)	3.28(0.10)
(D)	50	0.69(0.08)	0.69(0.09)	0.63(0.09)	1.62(0.21)	4.32(0.42)
	100	0.62(0.05)	0.42(0.04)	0.31(0.03)	1.34(0.14)	3.82(0.27)
(E)	50	1.05(0.05)	1.04(0.05)	1.04(0.06)	1.85(0.12)	3.78(0.21)
	100	1.13(0.03)	1.13(0.02)	1.07(0.02)	1.66(0.07)	3.55(0.12)
(F)	50	2.05(0.06)	2.26(0.05)	2.43(0.06)	2.55(0.09)	5.10(0.23)
	100	2.11(0.05)	2.19(0.04)	2.35(0.04)	2.56(0.08)	4.96(0.14)

Table 3: LSCV method: an average of Euclidean norm with a standard error.

One can use each of the kernels for the least square cross-validation method. The bias cross-validation methods require some smoothness conditions to be satisfied. In this case, we can use only Epanechnikov, the biweight, and the cosine kernel for BCV_1 and the biweight and the cosine kernel for BCV_2.

The biweight kernel seems to be the best choice for the least square cross-validation method. According to ED-criterion, Epanechnikov and the uniform kernel are also good choice. By contrast, the uniform kernel is the least suitable choice according to ISE-criterion. The second best choice is Epanechnikov and the cosine kernel. On the other hand, LSCV-optimal bandwidths suffer from a large variability (see [2]). For all kernels this variability is approximately the same.

In the case of the BCV_1, we can choose between using Epanechnikov and the

ED		BCV$_1$			BCV$_2$	
density	*n*	Epan	Biw	Cos	Biw	Cos
(A)	50	3.79(0.02)	6.29(0.03)	4.15(0.02)	2.62(0.06)	1.12(0.03)
	100	3.43(0.01)	5.44(0.01)	3.74(0.01)	2.31(0.03)	1.07(0.02)
(B)	50	1.19(0.01)	2.12(0.01)	1.34(0.01)	0.54(0.01)	0.15($*$)
	100	1.14($*$)	1.91(0.01)	1.26($*$)	0.53(0.01)	0.16($*$)
(C)	50	0.63($*$)	1.16(0.01)	0.71(0.01)	0.22(0.01)	0.04($*$)
	100	0.62($*$)	1.05($*$)	0.69($*$)	0.23(0.01)	0.05($*$)
(D)	50	1.90(0.01)	3.27(0.02)	2.11(0.02)	1.05(0.03)	0.35(0.01)
	100	1.79(0.01)	2.88(0.01)	1.97(0.01)	1.02(0.02)	0.40(0.01)
(E)	50	1.14(0.01)	1.98(0.01)	1.26(0.01)	0.66(0.02)	0.34(0.02)
	100	1.06($*$)	1.75(0.01)	1.17($*$)	0.56(0.01)	0.25(0.01)
(F)	50	0.01($**$)	0.04($*$)	0.01($**$)	0.32(0.02)	0.57(0.02)
	100	0.02($**$)	0.06($**$)	0.02($**$)	0.15($*$)	0.31($*$)

Table 4: BCV$_1$ and BCV$_2$ methods: an average of Euclidean norm with a standard error ($*$ stands for the standard error less than 0.005 and $**$ for the standard error less than 0.001).

$100 \times$ ISE		LSCV				
density	*n*	Uni	Epan	Biw	Cos	Tri
(A)	50	0.48(0.02)	0.39(0.02)	0.38(0.02)	0.43(0.02)	0.48(0.02)
	100	0.37(0.01)	0.27(0.01)	0.24(0.01)	0.30(0.01)	0.38(0.01)
(B)	50	1.11(0.04)	0.86(0.03)	0.77(0.03)	0.90(0.04)	1.10(0.04)
	100	0.97(0.02)	0.69(0.02)	0.58(0.02)	0.70(0.02)	0.91(0.03)
(C)	50	3.54(0.07)	2.47(0.07)	2.03(0.07)	2.48(0.07)	3.06(0.07)
	100	3.52(0.05)	2.47(0.05)	1.97(0.05)	2.37(0.05)	2.84(0.06)
(D)	50	5.23(0.07)	4.43(0.08)	4.05(0.08)	4.28(0.08)	4.62(0.08)
	100	5.04(0.06)	4.12(0.06)	3.67(0.06)	3.92(0.06)	4.30(0.06)
(E)	50	1.03(0.03)	0.79(0.03)	0.73(0.03)	0.82(0.04)	0.94(0.04)
	100	0.97(0.02)	0.71(0.02)	0.61(0.02)	0.68(0.02)	0.83(0.02)
(F)	50	5.32(0.08)	4.25(0.08)	3.76(0.08)	4.10(0.08)	4.57(0.08)
	100	5.32(0.06)	4.12(0.06)	3.59(0.06)	3.97(0.06)	4.47(0.06)

Table 5: LSCV method: an average of the integrated square error with a standard error.

cosine kernel, because their performance is comparable regarding both critera. But the BCV$_1$ method gives quite underestimated values of the optimal bandwidth H.

Concerning BCV$_2$, it is obvious that the cosine kernel performs better than the biweight kernel. Density (F) is an exception, but there is influence of boundary effects. The main advantage of the cosine kernel over the biweight kernel is that its fourth derivative, needed for calculation of BCV$_2$, is not a constant function, as it is the case for the biweight kernel.

4 Conclusion

In this paper we compared several two-dimensional kernels with a bounded support with respect to the method for finding optimal bandwidth. We propose that the biweight kernel is the best choice for the LSCV method and the cosine kernel is the best choice for the BCV$_2$. In the case of BCV$_1$, one can decide whether to use the

$100 \times$ ISE		BCV$_1$			BCV$_2$	
density	n	Epan	Biw	Cos	Biw	Cos
(A)	50	4.79(0.12)	9.48(0.32)	5.51(0.15)	1.24(0.05)	0.73(0.03)
	100	4.80(0.12)	9.08(0.28)	5.54(0.15)	0.90(0.02)	0.54(0.01)
(B)	50	5.82(0.28)	11.35(0.58)	6.81(0.33)	1.48(0.07)	0.94(0.04)
	100	5.81(0.23)	11.17(0.51)	6.63(0.28)	1.07(0.04)	0.67(0.02)
(C)	50	9.87(0.22)	18.90(0.42)	11.18(0.25)	2.44(0.10)	1.55(0.07)
	100	9.81(0.19)	18.80(0.45)	11.26(0.23)	1.72(0.05)	1.10(0.04)
(D)	50	7.43(0.22)	14.10(0.41)	8.43(0.25)	3.37(0.10)	3.34(0.08)
	100	7.65(0.17)	14.38(0.37)	8.98(0.23)	2.55(0.06)	2.55(0.06)
(E)	50	5.81(0.18)	11.42(0.40)	6.68(0.21)	1.47(0.05)	0.94(0.03)
	100	6.16(0.17)	11.63(0.42)	7.04(0.22)	1.12(0.03)	0.71(0.02)
(F)	50	7.90(0.25)	15.94(0.73)	9.09(0.32)	2.97(0.08)	2.96(0.08)
	100	8.95(0.28)	17.51(0.66)	10.48(0.35)	2.37(0.05)	2.31(0.05)

Table 6: BCV$_1$ and BCV$_2$ methods: an average of the integrated square error with a standard error.

cosine kernel or Epanechnikov kernel.

In the future, we want to extend this study to a larger group of kernels and also to a higher dimension.

References

[1] T. Duong and M. L. Hazelton, *Cross-validation Bandwidth Matrices for Multivariate Kernel Density Estimation*, Scand. J. Statist. **32** (2005) 485–506.

[2] I. Horová, J. Koláček, J. Zelinka and K. Vopatová, *Bandwidth choice for kernel density estimates*, Proc. IASC (2008) 542–551.

[3] D.W. Scott, “Multivariate Density Estimation: Theory, Practice, and Visualization”, John Wiley & Sons, New York, 1992.

[4] M.P. Wand and M.C. Jones, *Comparison of Smoothing Parameterizations in Bivariate Kernel Density Estimation*, J. Amer. Statist. Assoc. **88** (1993), 520–528.

[5] M.P. Wand and M.C. Jones: “Kernel Smoothing”, Chapman & Hall, London, 1995.