AISM (2008) 60:21–44
DOI 10.1007/s10463-006-0077-1

A cross-validation method for data with ties in kernel density estimation

Kamila Żychaluk · Prakash N. Patil

Received: 7 June 2005 / Revised: 9 June 2006 / Published online: 8 November 2006

Abstract Limitation of the cross-validation method of bandwidth selection is well known when applied to data with ties. A method which resolves this problem and which is easy to understand and implement is proposed. We show that the proposed approach is viable in theory, by proving its asymptotic equivalence to the standard cross-validation method. The practical usefulness is shown in simulations and an application to a real data example.

Keywords Cross-validation · Ties · Kernel density estimator

1 Introduction

The choice of smoothing parameter is crucial for nonparametric kernel density estimation and hence, not surprisingly, it has been one of the most intensely discussed and researched topics in statistical literature. This has lead to a wide variety of methods of bandwidth (that is, smoothing parameter) selection for kernel density estimation. Of the various methods proposed, plug-in and least squares cross-validation are the most preferred approaches. A detailed appraisal of these methods is given in Loader (1999). The author concludes that in the detailed analysis plug-in approaches to bandwidth selection fare poorly. This is due to their heavy dependence on the arbitrary specification of pilot bandwidths

K. Żychaluk (✉)
Sensing Imaging and Signal Processing Group, School of Electrical and Electronic Engineering, Faculty of Engineering and Physical Sciences, University of Manchester, Moffat Building, Sackville Street, Manchester M60 1QD, UK
e-mail: Kamila.Zychaluk@manchester.ac.uk

P. N. Patil
The University of Birmingham, The Watson Building, Birmingham B15 2TT, UK

and the fact that they are biased when this specification is wrong. Whereas, careful analysis reveals that classical approach like cross-validation method of bandwidth selection often produces estimates that are far more informative than otherwise.

Unfortunately, the good asymptotic and finite sample performance of the cross-validation can only be guaranteed for a continuous sample and the method fails when the data contain ties (Silverman 1986, p. 51). Ties in the sample occur when the data are discretised to save storage space or to accelerate computations. Moreover, all measurements have only finite accuracy and hence the data are always rounded to the precision of the tool. Considering the good performance revealed by Loader (1999), failing of cross-validation method in the presence of tied observation is surely a disturbing property and, if possible, should be remedied.

The issue of failure of cross-validation method of bandwidth selection in the presence of tied observation was first addressed by Chiu (1991). Noting that all data are rounded or discretised to some degree, he shows how easily such failure of cross-validation method could occur. To overcome the problem, the author considers the estimation of cross-validation score function in the frequency domain and suggests to use the truncated version of the characteristic function when the data contain ties.

Our idea of tackling the problem of ties is a simple and intuitive approach: to add a small "continuous" noise to the given sample with ties and then apply the standard cross-validation method to this "contaminated" sample which now has no ties. This method is very easy to implement and as we show in this paper it has good asymptotic and finite sample properties.

The idea of adding noise to the data on purpose, as we propose, is not new. In a completely different context of simple linear regression when the independent variable is subject to measurement error, Ruppert et al. (1999) propose a method to estimate the slope parameter based on the idea of adding noise to the data on purpose. This idea also has been used by Machado and Santos Silva (2002) to define an estimator of quantile of a discrete random variable and prove its consistency.

2 Definitions

Let $X_1, X_2, \ldots, X_n$ be an independent and identically distributed (i.i.d.) sample with density f_X. Recall the definition of a kernel density estimator (Rosenblatt, 1956)

$$\hat{f}_X(z) = \frac{1}{nh} \sum_{j=1}^{n} K\left(\frac{z - X_j}{h}\right), \tag{1}$$

where the kernel K is a symmetric probability density function which is assumed to be compactly supported and h is the smoothing parameter.

In this article, the bandwidth which minimises the *integrated square error* (ISE) will be referred to as the *optimal bandwidth* $\hat{h}_{X,\mathrm{opt}}$ and the bandwidth which minimises the *mean integrated square error* (MISE $= E\int(f-\hat{f})^2$) the *asymptotically optimal bandwidth* $h_{X,\mathrm{opt}}$.

The cross-validation method approximates the part of ISE which depends on h, $\int\hat{f}^2 - 2\int f\hat{f}$, by the *cross-validation score function*,

$$M_X(h) = \frac{1}{n^2h}\sum_i\sum_j K^*\left(\frac{X_i - X_j}{h}\right) + \frac{2K(0)}{nh}, \tag{2}$$

where $K^*(z) = K*K(z) - 2K(z)$ and $K*K$ denotes convolution; the bandwidth $\hat{h}_{X,\mathrm{cv}}$ which minimises $M_X(h)$ will be referred to as the *cross-validation bandwidth*. For a sample with no ties, this method was shown to be asymptotically consistent (see e.g. Hall 1983, Stone 1984, Bowman 1984).

If the data do contain ties, then consider all pairs (X_i, X_j) for $1 \leqslant i < j \leqslant n$, and let N be the number of these pairs for which $X_i = X_j$; and denote $\zeta = K*K(0)/(4K(0) - 2K*K(0))$. Assume $\zeta > 0$, it can be easily seen that if $N > \zeta n$, then $M_X(h) \to -\infty$, as $h \to 0$ and the cross-validation method chooses $\hat{h}_{X,\mathrm{cv}} = 0$ as the optimal value (see Silverman, 1986, p. 51 for details).

Remark 1 To ensure $N < n\zeta$, a kernel such that $K*K(0) > \frac{4N}{2N+n}K(0)$ is needed but $K*K(0) < K(0)$ for a kernel which has a maximum at 0. Hence, $N < n\zeta$ can only be fulfilled if $N < n/2$. Note also that usually $N \geqslant n$ (at least if there are more than three observations in each bin) and thus the requirement $N \geqslant \zeta n$ often fails to hold and the standard cross-validation cannot be used in many real cases.

We now give a precise interpretation of discretisation and then define our cross-validation method in detail.

2.1 Discretisation and contamination

We consider two cases here:

Case (i) The data are continuous and the ties occur "naturally" (i.e. not as a result of discretisation). We will call such samples *continuous*.

Case (ii) This case arises when the data are artificially discretised, an example here could be the ages given in surveys (obviously, age is a continuous variable but it is discretised to years). Samples of this type will be referred to as *discretised*.

In the first case, we can assume that the data still follow the density f_X, in the latter case, however, the sample has a different distribution.

Definition 1 *Let* $\{x_j\}_{j=-\infty}^{+\infty}$ *be the grid points such that* $\Delta x \equiv x_{j+1} - x_j = Dn^{-\beta}$, $\beta > 0$ *and assume that* $x_0 = 0$. *The* discretised *version of* X_i *is then given by* $\tilde{X}_{i,n} = \sum_{j=-\infty}^{+\infty}\tilde{x}_j I_{[x_j,x_{j+1})}(X_i)$, *where* $\tilde{x}_j = \frac{x_j + x_{j+1}}{2}$.

Note that this implies that the discretised sample is close to the original one, namely $|X_i - \tilde{X}_{i,n}| < Dn^{-\beta}/2$. The density of the discretised sample is then

$$f_{\tilde{X},n}(z) = \sum_{j=-\infty}^{+\infty} \delta_{\tilde{x}_j}(z) \int_{x_j}^{x_{j+1}} f_X(x)\mathrm{d}x,$$

where $\delta_x(z) = \delta_0(x - z)$ and δ_0 denotes the Kronecker delta.

Remark 2 Note that in practice the discretising mechanism may not be known and the only available information is the data. In this case, one would need to choose the grid points arbitrary, e.g. midpoints between the data points.

Define the family of density functions f_{U_n} by the scale condition,

$$f_{U_n}(z) = n^{\gamma} f_U(zn^{\gamma}), \quad \gamma > 0, \tag{3}$$

for an even and compactly supported density f_U. The support of f_{U_n} is $\mathrm{supp}(f_{U_n}) = [-an^{-\gamma}, an^{-\gamma}]$ and the moments of f_{U_n} satisfy $\mu_{U_n,k} \equiv E(U_n^k) = O(n^{-k\gamma})$, where the variable U_n follows the distribution f_{U_n}.

The *contaminated cross-validation* is the standard cross-validation method applied to the sample contaminated by noise. The sample contamination is defined below; note that the definition slightly differs in the continuous and in the discrete case.

Case (i) Let $U_{n,1}, U_{n,2}, \ldots, U_{n,n}$ be an i.i.d. sample from a known density f_{U_n} for an *arbitrary* value of $\gamma > 0$ (subject to the assumptions stated in the next section) and assume that $U_{n,k}$ is orthogonal to X_j for all k, j. Then the *contaminated* sample $Y_{n,i} = X_i + U_{n,i}$, $i = 1, \ldots, n$, has the density

$$f_{Y_n}(z) = f_X(z) + \frac{\mu_{U_n,2}}{2} f_X''(z) + \frac{\mu_{U_n,4}}{4!} f_X^{(4)}(z) + \int f_{U_n}(v) \frac{v^4}{4!} \Delta(f, z, v, \theta_v)\mathrm{d}v, \tag{4}$$

where $\Delta(f, z, v, \theta_v) = f_X^{(4)}(z - \theta_v v) - f_X^{(4)}(z)$; here and in what follows $\theta_v \in (0, 1)$.

Expansion (4) implies that the contaminated density f_{Y_n} converges pointwise to f_X at the rate $n^{-2\gamma}$.

Case (ii) In this case, the density of the noise $U_{n,1}, U_{n,2}, \ldots, U_{n,n}$ is uniform with $a = D/2$ and $\gamma = \beta$, which guarantees that the noise fills in the "gaps" between the data. Define $\tilde{Y}_{n,i} = \tilde{X}_{n,i} + U_{n,i}$, the density of $\tilde{Y}_{n,i}$ is

$$\begin{aligned} f_{\tilde{Y}_n}(z) = {} & f_X(z) D n^{-\gamma} \sum_{j=-\infty}^{+\infty} f_{U_n}\left(z - \tilde{x}_j\right) + f_X'(z) D n^{-\gamma} \sum_{j=-\infty}^{+\infty} \left(\tilde{x}_j - z\right) f_{U_n}\left(z - \tilde{x}_j\right) \\ & + \sum_{j=-\infty}^{+\infty} f_{U_n}\left(z - \tilde{x}_j\right) \int_{x_j}^{x_{j+1}} \frac{(x-z)^2}{2} f_X''(z_x^*)\mathrm{d}x, \end{aligned} \tag{5}$$

where z_x^* lies between z and x.

Note that the first term on the right-hand side of (5) $Dn^{-\gamma}\sum_{j=-\infty}^{+\infty} f_{U_n}(z-\tilde{x}_j) \equiv 1$ for all z, if and only if f_{U_n} is the uniform density over $[-Dn^{-\gamma}/2, Dn^{-\gamma}/2]$. For the uniform distribution, the contaminated density $f_{\tilde{Y}_n}$ tends to f_X at the rate $n^{-\gamma}$, slower than in the continuous case. If the noise density is not uniform then the contaminated density does *not* converge to f_X.

The convergence of the contaminated density to f_X has different rate in the continuous and discrete case; this results in different rates of convergence for the cross-validation bandwidth in these two cases. Let us denote this rate of convergence by γ_0, which will be equal to 1/5 for the continuous case and to 2/5 for the discretised case.

2.2 Cross-validation with contaminated sample

The cross-validation score function $M_Y(h)$ (based on either $\tilde{Y}_n$ or Y_n) and its minimum $\hat{h}_{Y,\text{cv}}$ depend on the noise sample and hence for each sample of noise, one obtains a different function $M_Y(h)$ possibly with a minimum at a different point. Note that the standard cross-validation bandwidth $\hat{h}_{X,\text{cv}}$ is also a random variable; however, the dependence on noise sample introduces additional variability in the cross-validation bandwidth $\hat{h}_{Y,\text{cv}}$ and the contaminated cross-validation gives results which vary even for a given sample $X_1, X_2, \ldots, X_n$, while the standard cross-validation bandwidth is fixed for this sample.

To reduce the variability introduced by noise in the cross-validation bandwidth $\hat{h}_{Y,\text{cv}}$, we consider a mean bandwidth. Let $\{U_i^l\}_{i=1,\ldots,n}^{l=1,\ldots,L}$ be L independent samples of noise. For each sample, find the cross-validation score function $M_{Y,l}(h)$. Define the *mean cross-validation score function*

$$\bar{M}_{Y,L}(h) = \frac{1}{L}\sum_{l=1}^{L} M_{Y,l}(h),$$

and denote by $\hat{H}_{\text{cv},L}$ the bandwidth which minimises this function.

Another approach to averaging is to find the cross-validation bandwidth $\hat{h}_{\text{cv},Y}$ for each of the functions $M_{Y,l}(h)$ and then consider the mean of these bandwidths. Thus we denote $\hat{h}_{\text{cv},Y}$ calculated for lth sample of noise by $\hat{h}_{\text{cv},l}$ and then

$$\bar{h}_{\text{cv},L} = \frac{1}{L}\sum_{l=1}^{L} \hat{h}_{\text{cv},l}.$$

The two approaches do not necessarily give the same result.

3 Main results

In the next two subsections, we present the asymptotic results for the contaminated cross-validation method. The first subsection consists of the results for the criterion functions while in the second one, we deal with the asymptotic behaviour of the cross-validatory bandwidth. The proofs are given in the Appendix.

To simplify the notation, we use the subscript Y for either Y_n or $\tilde{Y}_n$. The following assumptions will be used in the theorems:

(A1) f_X has two continuous derivatives;
(A2) f_X has six continuous derivatives;
(A3) $h_n \in I_n \equiv [b_1 n^{-\frac{1}{5}}, b_2 n^{-\frac{1}{5}}]$;
(A4) the symmetric kernel K is a Hölder continuous function, i.e. $\forall_{u,v} |K(u) - K(v)| \leqslant M|u-v|^{\alpha}$, for some $\alpha \in (0,1]$ and a constant $M > 0$;
(A5) the symmetric kernel K has two continuous derivatives.

3.1 Criterion functions

Theorem 1 *Let the densities f_{Y_n}, $f_{\tilde{Y}_n}$ be defined as above and assume that the assumptions* A1, A3 *and* A4 *hold. If $\gamma > \gamma_0$, then*

$$\sup_{h \in I_n} \left| \frac{\mathrm{ISE}_X(h) - \mathrm{ISE}_Y(h)}{\mathrm{MISE}_X(h)} \right| \to 0 \tag{6}$$

almost surely as $n \to \infty$.

It can be easily shown that $\sup_h |M_X(h) + \int f_X^2 - \mathrm{ISE}_X(h)| = o(\mathrm{MISE}(f_X))$ (see Lemma 1 in the Appendix). Immediate consequence of this fact and Theorem 1 is the following corollary, which we state without proof.

Corollary 1 *Under the assumptions* A1, A3 *and* A4*,*

$$\sup_{h \in I_n} \left| \frac{M_X(h) - M_Y(h)}{\mathrm{MISE}(f_X)} \right| \to 0 \tag{7}$$

almost surely as $n \to \infty$.

A similar result holds for the mean cross-validation score function which we again state without proof.

Theorem 2 *For any fixed integer $L > 1$, assuming* A1, A3 *and* A4*,*

$$\sup_{h \in I_n} \left| \frac{M_X(h) - \bar{M}_{Y,L}(h)}{\mathrm{MISE}(f_X)} \right| \to 0 \tag{8}$$

almost surely as $n \to \infty$, if $\gamma > \gamma_0$.

3.2 Cross-validation bandwidth

Denote $R(g) = \int g^2(x)\mathrm{d}x$, $K_1(x) = xK'(x)$, $K_1^*(x) = xK^{*''}(x)$ and let $S_j(x) = ((x_j - x)/h, (x_{j+1} - x)/h)$ and $S_{1j} = (x_j, x_{j+1})$.

Theorem 3 *Assume that* A2, A3 *and* A5 *hold, Then*

$$n^{3/10}(\hat{h}_{Y,\mathrm{cv}} - \hat{h}_{X,\mathrm{opt}}) \to N(0, \sigma^2_{\mathrm{cv}}), \tag{9}$$

in distribution as $n \to \infty$, with $\gamma \geqslant \gamma_0$. The variance σ^2_{cv} is equal to $\sigma^2_{X,\mathrm{cv}}$, if $\gamma > \gamma_0$, to $\sigma^2_{Y,\mathrm{cv}}$, if $\gamma = 1/5$ in the continuous case, or to $\sigma^2_{\tilde{Y},\mathrm{cv}}$, if $\gamma = 2/5$ in the discretised case, where

$$\sigma^2_{X,\mathrm{cv}} = \left(\frac{2}{C_h}\right)^3 R(f_X)\cdot R(K_1) + \left(2\sigma_K^2 C_h\right)^2 \left(\int (f_X''(x))^2 f_X(x)\mathrm{d}x - R^2(f_X')\right),$$

$$\sigma^2_{Y,\mathrm{cv}} = \frac{2\mu_{U,2}}{C_h^5}\left(2R(K^{*'}) + R(K_1^*)\right) R(f_X) + \sigma^2_{X,\mathrm{cv}},$$

$$\sigma^2_{\tilde{Y},\mathrm{cv}} = \frac{C(K, f_x, D)}{n^4 h^3} + \sigma^2_{Y,\mathrm{cv}},$$

with $C_h = \left(\frac{R(K)}{4\sigma_K^2 R(f_X'')}\right)^{1/5}$ and

$$C(K, f_x, D) = \int\left(z - \frac{D}{C_h}\sum_{j=-\infty}^{+\infty}\frac{2j+1}{2}\left(I_{S_j(x)}(z) - I_{S_{1j}}(x)\right)\right)^2 \times \left(2K^{*'}(z) + zK^{*''}(z)\right)^2 f_X^2(x)\mathrm{d}x\,\mathrm{d}z.$$

Theorem 3 and the fact that $\hat{h}_{X,\mathrm{opt}}/h_{X,\mathrm{opt}} \to 1$ almost surely imply immediately the asymptotic normality of the ratio, as stated in the corollary below.

Corollary 2 *Under the assumptions* A2, A3 *and* A5,

$$n^{1/10}\frac{\hat{h}_{Y,\mathrm{cv}} - \hat{h}_{X,\mathrm{opt}}}{\hat{h}_{X,\mathrm{opt}}} \to N\left(0, \left(\frac{\sigma_{\mathrm{cv}}}{C_h}\right)^2\right), \quad \textit{in distribution, as } n \to \infty, \tag{10}$$

with $\gamma \geqslant \gamma_0$ and C_h defined as in Theorem 3.

Theorem 3 implies also the following corollary, which we state without proof.

Corollary 3 *Under the assumptions* A2, A3 *and* A5,

$$n\left(\mathrm{ISE}_X(\hat{h}_{Y,\mathrm{cv}}) - \mathrm{ISE}_X(\hat{h}_{X,\mathrm{opt}})\right) \to \frac{\sigma^2_{\mathrm{cv}}}{2}\chi_1^2, \tag{11}$$

in distribution, as $n \to \infty$, with $\gamma \geqslant \gamma_0$.

The asymptotic results for the cross-validation bandwidth $\hat{h}_{Y,\mathrm{cv}}$ have their analogues for the bandwidths $\hat{H}_{\mathrm{cv},L}$ and $\bar{h}_{\mathrm{cv},L}$. We summarise them in the following proposition.

Proposition 1 *Under the assumptions* A2, A3 *and* A5, *the bandwidths* $\hat{H}_{\mathrm{cv},L}$ *and* $\hat{h}_{\mathrm{cv},L}$ *are asymptotically equivalent to the optimal bandwidth*

$$\frac{\hat{H}_{L,\mathrm{cv}}}{\hat{h}_{X,\mathrm{opt}}} \to 1 \quad \textit{and} \quad \frac{\bar{h}_{L,\mathrm{cv}}}{\hat{h}_{X,\mathrm{opt}}} \to 1,$$

almost surely as $n \to \infty$, when $\gamma > \gamma_0$. Furthermore, if we assume that $f_X \in C^6$, then

$$n^{1/10}\frac{\hat{H}_{L,\mathrm{cv}} - \hat{h}_{X,\mathrm{opt}}}{\hat{h}_{X,\mathrm{opt}}} \to N\left(0, \sigma^2_{L,\mathrm{cv}}\right), \tag{12}$$

$$n^{1/10}\frac{\bar{h}_{L,\mathrm{cv}} - \hat{h}_{X,\mathrm{opt}}}{\hat{h}_{X,\mathrm{opt}}} \to N\left(0, \sigma^2_{L,\mathrm{cv}}\right), \tag{13}$$

in distribution as $n \to \infty$, with $\gamma \geqslant \gamma_0$. The variance $\sigma^2_{L,\mathrm{cv}}$ is equal to $\sigma^2_{X,\mathrm{cv}}$ if $\gamma > \gamma_0$ and to $\sigma^2_{L,c}$ if $\gamma = \gamma_0 = 1/5$ in the continuous case or to $\sigma^2_{L,d}$ if $\gamma = \gamma_0 = 2/5$ in the discretised case, where

$$\sigma^2_{L,c} = \frac{2\mu_{U,2}}{LC_h^5}\left(2R(K^{*'}) + R(K_1^*)\right)R(f_X) + \frac{\sigma^2_{X,\mathrm{cv}}}{C_h^2},$$

$$\sigma^2_{L,d} = \frac{C(K, f_x, D)}{n^4 h^3} + \sigma^2_{L,c}.$$

Remark 3 Notice that the asymptotic variances for h_L and H_L are the same and that averaging decreases asymptotic variances of these bandwidths if $\gamma = \gamma_0$.

4 Examples

4.1 Simulations

To illustrate how the cross-validation score function behave for the continuous, discretised and contaminated samples, we compare its plots for three simulated samples drawn from densities: standard normal, f_1, mixture of two normal distributions, $f_2 = 0.3\frac{5}{\sqrt{2\pi}}\mathrm{e}^{-25x^2/2} + 0.7\frac{1}{3\sqrt{2\pi}}\mathrm{e}^{-x^2/18}$ and gamma, $f_3 = \mathrm{Gam}(2, 1.5)$. Two discretised (as defined in Sect. 2.1) versions for each of the three simulated samples are created, one with a ratio of the number of ties to the sample size

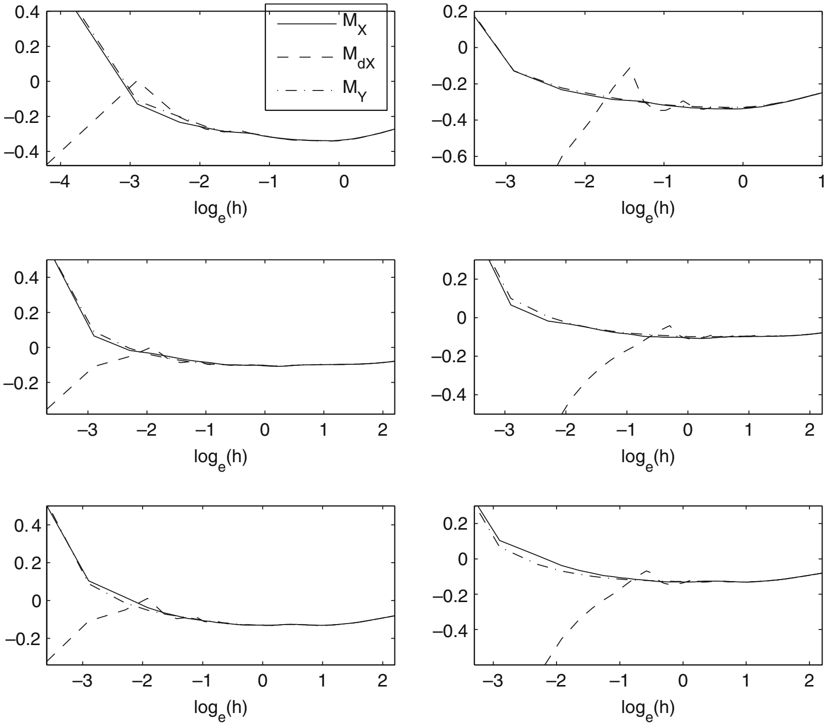

Fig. 1 The cross-validation score functions for samples of size 50 from densities $f_1 - f_3$, with $L = 50$ in three cases: for original continuous data (*solid*), for discretised data (*dashed*) and for discretised data with noise (*dashdot*). The value of N/n was near 0.5 for the left-hand side plots and 2 for the right-hand side plots. Plots are in $\log_e$ scale

(N/n) approximately equal to 0.5 and the other with $N/n \approx 2$. The values of the cross-validation bandwidth are chosen in the interval $[0.02, 20] \cdot n^{-1/5}$. In each case three cross-validation functions are plotted: the function based on the original continuous sample (M_X), the function based on the discretised sample (M_{dX}) and the function based on the contaminated (discretised plus noise) sample (M_Y). The plots are in Fig. 1 and the values of the bandwidths chosen in each case are in Table 1. As expected, the cross-validation score function tends to minus infinity as $h \to 0$ for the discretised sample and therefore the bandwidth $\hat{h}_{\tilde{X},\mathrm{cv}}$ is close to 0 (here, $\hat{h}_{\tilde{X},\mathrm{cv}} = 0.0091$). But the contaminated cross-validation function has a minimum and the minimum is close to the minimum for the original sample. Note that, even when the number of ties to the sample size ratio is as high as 2, the contaminated cross-validation chose the bandwidths close to those chosen for the original sample in most cases. Also, the mean cross-validation functions based on the contaminated samples do not differ much from the cross-validation score function for the original sample.

Table 1 The values of bandwidth h chosen by the cross-validation method; the bandwidths $\hat{\bar{H}}_{\mathrm{cv},L}$, $\bar{\bar{h}}_{\mathrm{cv},L}$ are for a contaminated discretised sample

Distri-bution	$\hat{h}_{X,\mathrm{cv}}$	$\hat{h}_{\tilde{X},\mathrm{cv}}$	$\hat{\bar{H}}_{\mathrm{cv},L}$	$\bar{\bar{h}}_{\mathrm{cv},L}$	$\frac{N}{n}$	Bin width
f_1	0.9274	0.0091	0.8815	0.8760	0.5000	0.0600
f_1	0.9274	0.0091	0.8815	0.7869	2.0200	0.2400
f_2	1.2488	0.0091	1.2947	1.2938	0.5000	0.1600
f_2	1.2488	0.0091	1.3407	1.4114	2.0600	0.7450
f_3	2.6722	0.0091	2.6722	2.2140	0.5000	0.1600
f_3	2.6722	0.0091	2.6722	1.8228	2.1800	0.5850

All parameters as in Fig. 1

4.2 Comparison of different methods of averaging

We proposed two methods of reducing the variability in the estimates of optimal bandwidth obtained by contaminated cross-validation and showed that their asymptotic variances are the same. To compare their performance for a finite sample, we have estimated conditional (given the sample $X_1, X_2, \ldots, X_n$) variances of both bandwidths $\hat{H}_{\mathrm{cv},L}$ and $\bar{h}_{\mathrm{cv},L}$ for different values of L and plotted them as a function of L. The variances were estimated via Monte Carlo method.

Our simulations show that the bandwidth $\bar{\bar{h}}_{\mathrm{cv},L}$ performs better (i.e. has smaller variance), if the uniform kernel is used, nearly for all samples $X_1, X_2, \ldots, X_n$. The bandwidth $\bar{\bar{h}}_{\mathrm{cv},L}$ is also preferred for most samples $X_1, X_2, \ldots, X_n$ in the case of the Epanechnikov kernel but the evidence is not as strong in this case (results not shown here). If the normal kernel is used, then the results are hardly conclusive: the bandwidth $\hat{\bar{H}}_{\mathrm{cv},L}$ has lower conditional variance for nearly as many samples as the bandwidth $\bar{\bar{h}}_{\mathrm{cv},L}$.

Therefore, we have also estimated the (unconditional) variances for the bandwidths in the case of normal kernel. The bandwidth $\bar{\bar{h}}_{\mathrm{cv},L}$ seems to give better results than its competitor this time as well.

The examples of the estimated variances as functions of L are in Fig. 2. The left hand side shows the estimates of the conditional variances of the mean bandwidths for the uniform kernel and a sample drawn from distribution f_3; clearly, $\mathrm{Var}(\bar{\bar{h}}_{\mathrm{cv},L}|\underline{X})$ is lower than its competitor. The plots in the right-hand side of Fig. 2 are the estimated (unconditional) variances for the normal kernel and the distribution f_3. Similarly here, $\mathrm{Var}(\bar{\bar{h}}_{\mathrm{cv},L})$ is lower. It is important to notice that the differences between the two types of averaging become small for larger values of L and thus the choice of the method is not as influential in this case.

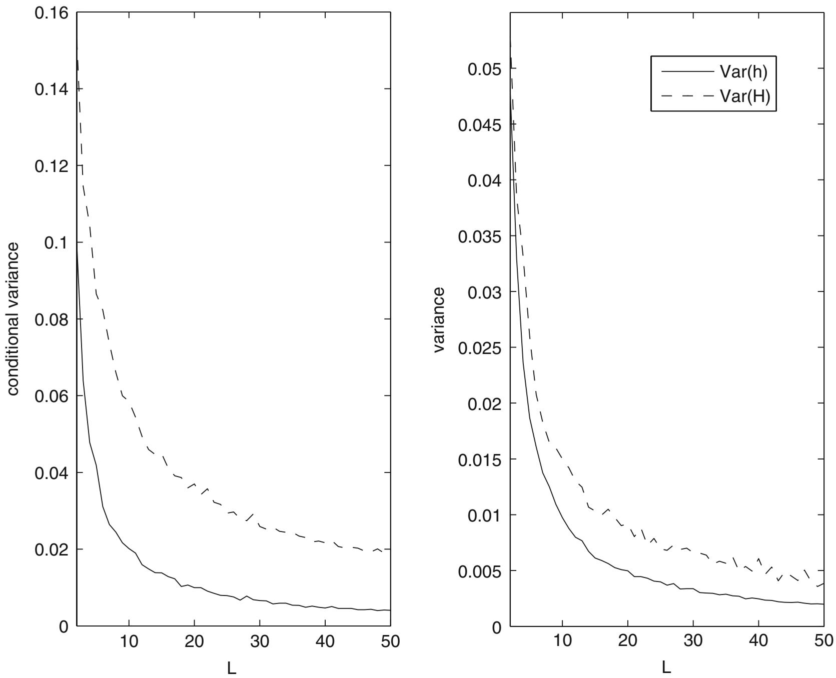

Fig. 2 The estimated variances of the cross-validation bandwidths $\hat{\bar{H}}_{\mathrm{cv},L}$ (Var(H)) and $\bar{\bar{h}}_{\mathrm{cv},L}$ (Var(h)) plotted as functions of L. The left-hand side panel shows the conditional variances estimated for a sample of size 30 from the distribution f_3 and uniform kernel, in the right hand side panel, there are the unconditional variances based on 50 samples of size 30 from the distribution f_3 and normal kernel

4.3 Real data example – the Stanford Heart Transplant data

We applied our method to a real data set. This example is the sample of ages of the patients in the Stanford Heart Transplant data (Andrews and Herzberg 1985). The ages of patients are given in years hence, the additional noise in contaminated cross-validation is (uniformly) distributed in $[0, 1]$.

The ratio of ties in the data for the analysed sample is $N/n = 2.6522$, which is more than ζ for Epanechnikov kernel and hence the method fails. The cross-validation score function for the original data drops off around 0 but the function for the contaminated data has a minimum in $\hat{H}_{\mathrm{cv},L} = 4.5988$ (Fig. 3, left-hand side); the value of the other average cross-validation bandwidth $\bar{h}_{50,\mathrm{cv}}$ was 4.4882. We have plotted the kernel estimates for the density function with bandwidth $\bar{h}_{50,\mathrm{cv}}$ based on the original sample and on the conatinated sample (Fig. 3, right-hand side). The estimate for the other bandwidth was virtually the same; hence, it is not presented here. The result is a unimodal density with mode around 49 and two "bumps" in 20 and 30. The estimate based on the contaminated sample is nearly the same as the one based on the original sample, but slightly smoother.

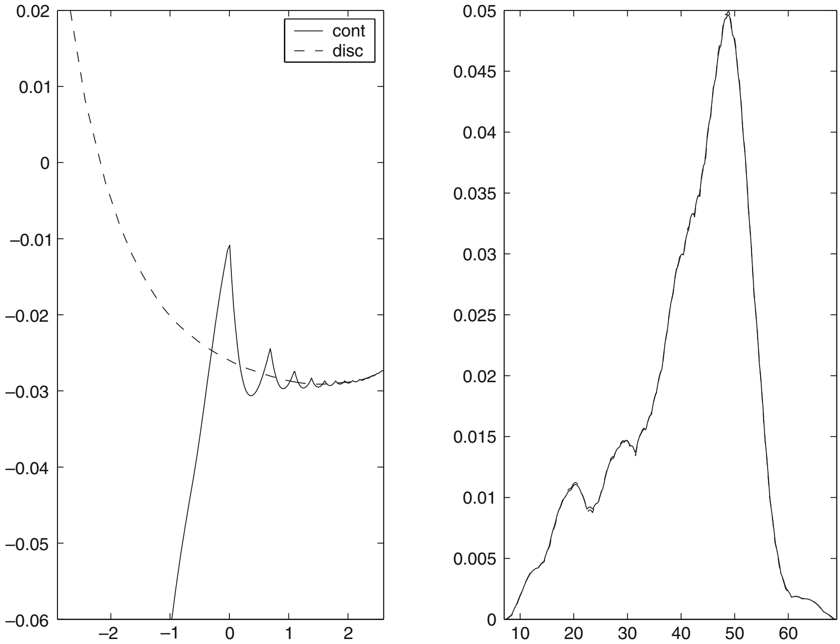

Fig. 3 The cross-validation score functions for the Stanford Heart Transplant data; plots are in $\log_e$ scale and are on the left-hand side. The estimates of the density for the Stanford Heart Transplant data based on the original discrete sample (*solid line*) and based on the contaminated sample (*dashed line*) are on the right hand side; for both the cross-validation bandwidth $\bar{h}_{50,\mathrm{cv}}$ was used

Acknowledgements The first author is grateful to the University of Birmingham and ORS award scheme for sponsoring her PhD at the School of Mathematics and Statistics in the University of Birmingham.

Appendix

Before proving Theorem 1, we state the following lemma which is a standard result and hence its proof is omitted here.

Lemma 1 *Under the assumptions* A1, A3 *and* A4 *and if* $\gamma > 0$,

$$\sup_{h\in I_n}\left|\frac{\tilde{M}_Y(h) - \mathrm{ISE}_{Y_n}(h)}{\mathrm{MISE}_X}\right| \to 0, \tag{14}$$

almost surely as $n \to \infty$, *where* $\tilde{M}_Y(h) = M_Y(h) + \int f_Y^2(x)\mathrm{d}x$.

Proof of Theorem 1 By Borel-Cantelli lemma, to prove (6), it suffices to show that

$$P\left[\sup_h\left|\frac{\mathrm{ISE}_X(h) - \mathrm{ISE}_{Y_n}(h)}{\mathrm{MISE}_X(h)}\right| > \varepsilon\right] < C\cdot n^{-\eta}, \tag{15}$$

where $\eta > 1$ and C denotes a generic constant. Note the the left-hand side of this inequality is less than $\sup_h E|(\mathrm{ISE}_X(h) - \mathrm{ISE}_{Y_n}(h))/\mathrm{MISE}_X(h)|^{2k} \cdot \varepsilon^{-2k}$ for any positive integer k. Cauchy–Schwartz inequality implies that

$$E\left|\mathrm{ISE}_X(h) - \mathrm{ISE}_{Y_n}(h)\right|^{2k} \leqslant \left[C \cdot \left(\left[\int \left(f_X - f_{Y_n}\right)^2\right]^{2k} + E\left[\int \left(\hat{f}_X - \hat{f}_{Y_n}\right)^2\right]^{2k}\right)\right.$$
$$\left.\times \left(E\left(\int \left(f_X - \hat{f}_X\right)^2\right)^{2k} + E\left(\int \left(f_{Y_n} - \hat{f}_{Y_n}\right)^2\right)^{2k}\right)\right]^{\frac{1}{2}}$$
$$\equiv C\,(\mathrm{B11} + \mathrm{B12})\,(\mathrm{B21} + \mathrm{B22})\,,$$

For the term B11, Taylor expansion for f_X gives

$$\int \left(f_{Y_n} - f_X\right)^2 = \frac{n^{-4\gamma}}{4} \int \left(\int u^2 f_X''(z - \theta_u u n^{-\gamma}) f_U(u) \mathrm{d}u\right)^2 \mathrm{d}z \leqslant C \cdot n^{-4\gamma}.$$

To bound term B12, write first

$$\left(\int \left(\hat{f}_X - \hat{f}_{Y_n}\right)^2\right)^{2k} \leqslant \frac{C}{(nh)^{4k}} \left[\left(\sum_{i=1}^{n} W_i\right)^{2k} + \left(\sum_{i<j} V_{ij}\right)^{2k}\right], \tag{16}$$

where $W_i = h \int \left(K(t) - K\left(t - \frac{U_{i,n}}{h}\right)\right)^2 \mathrm{d}t$ and $V_{ij} = \int \left(K\left(\frac{z - X_i}{h}\right) - K\left(\frac{z - X_i - U_{i,n}}{h}\right)\right)$ $\left(K\left(\frac{z - X_j}{h}\right) - K\left(\frac{z - X_j - U_{j,n}}{h}\right)\right) \mathrm{d}z$. For any i.i.d. $Z_1, Z_2, \ldots, Z_n$,

$$E\left(\sum_{i=1}^{n} Z_i\right)^{2k} \leqslant C\left[E\left(\sum_{i=1}^{n} (Z_i - EZ_i)\right)^{2k} + n^{2k}(EZ_1)^{2k}\right] \tag{17}$$

and note that $EW_1 \leqslant Cn^{-1/5 + 2\alpha(1/5 - \gamma)}$ and

$$E(W_1)^{2k} = h^{2k} \int \left(\int \left(K(t) - K\left(t - \frac{u}{h}\right)\right)^2 \mathrm{d}t\right)^{2k} f_{U,n}(u)\ \mathrm{d}u \leqslant Cn^{2k(-\frac{1}{5} + 2\alpha(\frac{1}{5} - \gamma))}.$$

This together with (17) and Rosenthal's inequality (see e.g. Hall and Heyde, 1980, p. 23) ($\sum_{i=1}^{j} (W_i - EW_i)$ is a martingale with $\mathcal{F}_j = \sigma\{X_i, U_i, i = 1, \ldots, \min(j, n)\}$) imply that

$$E\left(\sum_{i=1}^{n} W_i\right)^{2k} \leqslant Cn^{8k/5 + 4k\alpha(1/5 - \gamma)}.$$

To bound the latter sum in (16), define $V_i = E(V_{i,j}|X_i, U_i)$, $\tilde{V}_{i,j} = V_{i,j} - V_i - V_j + EV_{i,j}$ and $\tilde{V}_i = V_i - EV_i$. Since

$$E\left[K\left(\frac{z-X_i}{h}\right) - K\left(\frac{z-X_i-U_{i,n}}{h}\right)\right] = \frac{h}{2} f_X''(z) \int u^2 f_{U_n}(u)\mathrm{d}u + o\left(hn^{-2\gamma}\right)$$
$$= C \cdot f_X''(z) n^{-1/5-2\gamma} + o\left(hn^{-2\gamma}\right), \qquad (18)$$

then $EV_{ij} = Cn^{-\frac{2}{5}-4\gamma} + o\left(n^{-\frac{2}{5}-4\gamma}\right)$ and $V_i = \int Cf_X''(z) n^{-\frac{1}{5}-2\gamma}\left[K\left(\frac{z-X_i}{h}\right) - K\left(\frac{z-Y_{i,n}}{h}\right)\right]\mathrm{d}z$. Also

$$E(V_{ij})^{2k} \leqslant Ch^{2k+1-4k\alpha} \int |u_1|^{2k\alpha} |u_2|^{2k\alpha} f_{U_n}(u_1) f_{U_n}(u_2)\mathrm{d}u_1\mathrm{d}u_2 \int f_X^2(x)\mathrm{d}x$$
$$\leqslant Cn^{-1/5-2k/5+4k\alpha(1/5-\gamma)}. \qquad (19)$$

Hence $E(V_i)^{2k} \leqslant Cn^{-\frac{4k}{5}-4k\gamma+2k\alpha(\frac{1}{5}-\gamma)}$, hence $E\tilde{V}_{ij}^{2k} \leqslant Cn^{-1/5-2k/5+4k\alpha(1/5-\gamma)}$ and by Rosenthal's inequality $E\left(\sum_{j=i+1}^{n} \tilde{V}_{i,j}\right)^{2k} \leqslant Cn^{2k/5+4k\alpha(1/5-\gamma)}$, which implies that

$$E\left(\sum_{1\leqslant i<j\leqslant n} \tilde{V}_{i,j}\right)^{2k} \leqslant Cn^{7k/5+4k\alpha(1/5-\gamma)}.$$

Since also

$$E\left(\sum_{i=1}^{n} \tilde{s}_i\right)^{2k} \leqslant Cn^{k/5-4k\gamma+2k\alpha(1/5-\gamma)},$$

thus

$$E\left(\sum_{i<j} V_{ij}\right)^{2k} \leqslant Cn^{7k/5}\left[n^{4k\alpha(1/5-\gamma)} + n^{9k/5-8k\gamma}\right]$$

and hence

$$E\left[\int \left(\hat{f}_X - \hat{f}_{Y_n}\right)^2\right]^{2k} \leqslant Cn^{-8k/5+4k\alpha(1/5-\gamma)}. \qquad (20)$$

For the term B21, note that

$$E\left[\int \left(f_X - \hat{f}_X\right)^2\right]^{2k} \leqslant C\left(\left[\int \left(f_X - E\hat{f}_X\right)^2\right]^{2k} + E\left[\int \left(\hat{f}_X - E\hat{f}_X\right)^2\right]^{2k}\right)$$

and

$$\int \left(f_X - E\hat{f}_X\right)^2 = \int \left(\frac{\sigma_K^2 h^2}{2} f_X'' + o(h^2)\right)^2 \leqslant C \cdot h^4 = C \cdot n^{-4/5}.$$

Note also that

$$E\left[\int \left(\hat{f}_X - E\hat{f}_X\right)^2\right]^{2k} \leqslant \frac{C}{n^{4k}h^{4k}} \left(E\left[\sum_{i=1}^{n} W_i\right]^{2k} + E\left[\sum_{i<j} V_{ij}\right]^{2k}\right).$$

where $W_i = h\int (K(t) - EK_h)^2 \mathrm{d}t$, $V_{ij} = \int [(K((z - X_i)/h) - EK_h)(K((z - X_j)h) - EK_h)]\mathrm{d}z$ and $EK_h = E\left[K\left((z - X_1)/h\right)\right]$. W_i's are non-random and $E\left[\sum_{i=1}^{n} W_i\right]^{2k} = C\cdot(nh)^{2k} = O\left(n^{8k/5}\right)$. The sum $\sum_{i<j} V_{ij}$ is treated in analogous way to the latter sum in (16). Using the fact that $EV_{ij} = 0$, $E\left|V_{ij}\right|^{2l} = O\left(h^{2l+1}\right)$ and $E\left(V_{ij}^2|\mathcal{F}_{j-1}\right) = O(h^3)\, f_X(X_i)$ gives

$$E\left[\int \left(\hat{f}_X - E\hat{f}_X\right)^2\right]^{2k} \leqslant \frac{1}{n^{16k/5}} \cdot O\left(n^{8k/5} + n^{7k/5}\right) = O\left(n^{-8k/5}\right).$$

Similarly for B22, $E[\int (f_{Y_n} - \hat{f}_{Y_n})^2]^{2k} \leqslant C \cdot n^{-\frac{8k}{5}}$ and therefore $E|\mathrm{ISE}_X(h) - \mathrm{ISE}_{Y_n}(h)|^{2k} \leqslant C\, n^{-2k\left(\frac{4}{5} - \frac{\alpha}{5} + \alpha\gamma\right)}$. Since $\mathrm{MISE}_X(h) = O(n^{-4/5})$ for $h = O(n^{-1/5})$, the probability in (15) is bounded by

$$E\left|\frac{\mathrm{ISE}_X(h) - \mathrm{ISE}_{Y_n}(h)}{\mathrm{MISE}_X(h)}\right|^{2k} \cdot \varepsilon^{-2k} \leqslant \frac{C}{\varepsilon^{2k}} \cdot n^{-2k\alpha\left(\gamma - \frac{1}{5}\right)} = \frac{C}{\varepsilon^{2k}} \cdot n^{-\eta}, \quad \eta > 1,$$

for a sufficiently large k and thus (15) holds which proves (6). □

The discretised case is proven similarly to the continuous case, the only significantly different points are:

- for the term B11, using expansion (5),

$$\int \left(f_X(z) - f_{\tilde{Y}_n}(z)\right)^2 \mathrm{d}z = \frac{D^2 n^{-2\gamma}}{12} \sum_{j=-\infty}^{+\infty} \int_{x_j}^{x_{j+1}} \left(f_X'(z)\right)^2 \mathrm{d}z + o(n^{-2\gamma}) \leqslant Cn^{-2\gamma};$$

- for the term B12, the moments EW_i and EW_i^{2k} have the same orders as for the continuous case, but, the convergence in (18) is slower. Using Taylor expansion for f_X,

$$\left| f_X(z) - \int_{x_j}^{x_{j+1}} \frac{f_X(x)}{Dn^{-\gamma}} \mathrm{d}x \right| \leqslant \frac{|f_X'(z)| \cdot Dn^{-\gamma}}{2} + \int_{x_j}^{x_{j+1}} \frac{(x-z)^2}{2Dn^{-\gamma}} f_X''(z + \theta_{xz}(x-z))\mathrm{d}x,$$

and

$$\left| E\left[K\left(\frac{z - X_i}{h}\right) - K\left(\frac{z - \tilde{X}_i - U_{i,n}}{h}\right)\right]\right| \leqslant \frac{Dn^{-\gamma}h}{2} |f_X'(z)| \sum_{j=-\infty}^{+\infty} \int_{\frac{z-x_{j+1}}{h}}^{\frac{z-x_j}{h}} K(t)\mathrm{d}t$$

$$+ h^2 |f_X'(z)| \sum_{j=-\infty}^{+\infty} \int_{\frac{z-x_{j+1}}{h}}^{\frac{z-x_j}{h}} tK(t)\mathrm{d}t + o(n^{-1/5-\gamma})$$

$$\leqslant Cn^{-1/5-\gamma};$$

- thus

$$B12 \leqslant C\left(n^{-8k/5-4k\alpha(\gamma-1/5)} + n^{-4k\gamma} + n^{-k-2k\gamma-2k\alpha(\gamma-1/5)}\right),$$

which proves the result for $\gamma > 2/5$.

Proof of Theorem 3 We apply here the methods developed in Hall and Marron (1987) and use the Lemmas 2–7 stated further in this section. The MISE_X' admits the expansion

$$0 = \mathrm{MISE}_X'(h_{X,\mathrm{opt}}) = \mathrm{MISE}_X'(\hat{h}_{Y,\mathrm{cv}}) + (h_{X,\mathrm{opt}} - \hat{h}_{Y,\mathrm{cv}})\mathrm{MISE}_X''(h^*),$$

where h^* lies between $\hat{h}_{Y,\mathrm{cv}}$ and $h_{X,\mathrm{opt}}$. Denote

$$\begin{aligned} D_1(h) &= \mathrm{MISE}_X(h) - \mathrm{ISE}_X(h), \\ D_2(h) &= \mathrm{ISE}_X(h) - M_X(h), \\ D_3(h) &= M_X(h) - M_Y(h) = \frac{1}{n^2 h} \sum_{i,j} \left(K^*\left(\frac{X_i - X_j}{h}\right) - K^*\left(\frac{Y_i - Y_j}{h}\right)\right). \end{aligned}$$

Then $\mathrm{MISE}_X'(\hat{h}_{Y,\mathrm{cv}}) = D_1'(\hat{h}_{Y,\mathrm{cv}}) + D_2'(\hat{h}_{Y,\mathrm{cv}}) + D_3'(\hat{h}_{Y,\mathrm{cv}})$. Since $M_Y'(\hat{h}_{Y,\mathrm{cv}}) = 0$

$$(\hat{h}_{Y,\mathrm{cv}} - h_{X,\mathrm{opt}})\mathrm{MISE}_X''(h^*) = D_1'(\hat{h}_{Y,\mathrm{cv}}) + D_2'(\hat{h}_{Y,\mathrm{cv}}) + D_3'(\hat{h}_{Y,\mathrm{cv}}). \quad (21)$$

Lemmas 5, 6 and Lemma 3.2 in Hall and Marron (1987) together with Eq. (21) imply

$$(\hat{h}_{Y,\mathrm{cv}} - h_{X,\mathrm{opt}})\mathrm{MISE}''_X(h^*) = D'_1(h_{X,\mathrm{opt}}) + D'_2(h_{X,\mathrm{opt}}) + D'_3(h_{X,\mathrm{opt}}) + o(n^{-7/10}). \tag{22}$$

From the proof of Theorem 2.1 in Hall and Marron (1987),

$$(\hat{h}_{X,\mathrm{opt}} - h_{X,\mathrm{opt}})\mathrm{MISE}''_X(h^*) = D'_1(h_{X,\mathrm{opt}}) + o(n^{-7/10}),$$

hence

$$(\hat{h}_{Y,\mathrm{cv}} - \hat{h}_{X,\mathrm{opt}})\mathrm{MISE}''_X(h^*) = D'_2(h_{X,\mathrm{opt}}) + D'_3(h_{X,\mathrm{opt}}) + o(n^{-7/10}). \tag{23}$$

Finally, the fact that $\mathrm{MISE}''_X(h^*) = Cn^{-2/5}$ and Lemma 7 give

$$n^{3/10}(\hat{h}_{Y,\mathrm{cv}} - \hat{h}_{X,\mathrm{opt}}) \to N(0, \sigma^2_{\mathrm{cv}}),$$

in distribution as $n \to \infty$. □

Notation

Denote

$$\begin{aligned} V_{ij}(h) &= K^*\left(\frac{X_i - X_j}{h}\right) - K^*\left(\frac{Y_i - Y_j}{h}\right), \\ W_{ij}(h) &= \frac{X_i - X_j}{h} K^{*'}\left(\frac{X_i - X_j}{h}\right) - \frac{Y_i - Y_j}{h} K^{*'}\left(\frac{Y_i - Y_j}{h}\right), \end{aligned}$$

and $V_i(h) = E[V_{ij}(h)|X_i, U_i]$, $W_i(h) = E[W_{ij}(h)|X_i, U_i]$ and define $\tilde{V}_{ij}(h) = V_{ij}(h) - V_i(h) - V_j(h) + EV_{ij}$, $\tilde{W}_{ij}(h) = W_{ij}(h) - W_i(h) - W_j(h) + EW_{ij}$. Then

$$\begin{aligned} \frac{1}{2} D'_3(h) &= \frac{1}{n^2h^2}\sum_{i<j} \tilde{V}_{ij}(h) + \frac{1}{n^2h^2}\sum_{i<j} \tilde{W}_{ij}(h) \\ &\quad + \frac{1}{nh^2}\sum_{i=1}^{n} (V_i(h) - E(V_i(h)) + W_i(h) - E(W_i(h))) \\ &\quad + \frac{1}{h^2} E\left(V_i(h) + W_i(h)\right) + R \equiv I_1 + I_2 + I_3 + I_4 + R, \end{aligned} \tag{24}$$

where $R = n^{-1}I_3 + n^{-1}I_4$ is a negligible part. Define also

$$I_{31} = \frac{1}{nh^2}\sum_{i=1}^{n}\left[W_i(h) - E(W_i(h))\right],$$
$$I_{32} = \frac{1}{nh^2}\sum_{i=1}^{n}\left[V_i(h) - E(V_i(h))\right],$$

then $I_3 = I_{31} + I_{32}$. In the discretised case Y_is change to $\tilde{Y}_i$ in the definitions of $V_{ij},\ W_{ij}$.

Lemma 2 *If* $\gamma > \gamma_0$,

$$\sup_{h\in I_n} E|n^{7/10}D_3'(h)|^{2k} \to 0;$$

if $\gamma = \gamma_0$,

$$\sup_{h\in I_n} E|n^{7/10}D_3'(h)|^{2k} \leqslant C(b_1, b_2),$$

where $I_n = [b_1 n^{-\frac{1}{5}}, b_2 n^{-\frac{1}{5}}]$.

Proof For the continuous case, the expected values of V_{ij}, W_{ij} admit

$$EV_{ij}{=}h\int K^{*}(z)\left[-\frac{\mu_2}{2n^{2\gamma}}\left(f_X''(y+zh)f_X(y){+}f_X(x)f_X''(y+zh)\right){+}O\left(n^{-4\gamma}\right)\right]\mathrm{d}z\ \mathrm{d}y,$$
$$EW_{ij}{=}h\int zK^{*'}(z)\left[-\frac{\mu_2}{2n^{2\gamma}}\left(f_X''(y{+}zh)f_X(y){+}f_X(x)f_X''(y+zh)\right)+O\left(n^{-4\gamma}\right)\right]\mathrm{d}z\ \mathrm{d}y.$$

Hence using Taylor expansion of f_X, $E\left(V_{ij} + W_{ij}\right) = O(n^{-4\gamma}h^3) + O(n^{-2\gamma}h^5)$. Note that

$$E\left[K^{*}\left(\frac{X - X_i}{h}\right)|X\right] - E\left[K^{*}\left(\frac{X-Y}{h}\right)\right] = h\left(\int f_X^2(x)\mathrm{d}x{-}f_X(X_i) + o(h^2)\right)$$

and

$$E\left[K^{**}\left(\frac{X - X_i}{h}\right)|X\right] - E\left[K^{**}\left(\frac{X-Y}{h}\right)\right]$$
$$= h\left(-\int f_X^2(x)dx + f_X(X_i) + o(h^2)\right),$$

where $K^{**}(z) = zK^{*'}(z)$. Therefore, $V_i + W_i = o(n^{-3/5})$ and by Rosenthal's inequality,

$$E\left[\frac{1}{nh^2}\sum_{i=1}^{n}(V_i(h) + W_i(h) \;-\; E(V_i(h) + W_i(h)))\right]^{2k} \leqslant o\left(n^{-7k/5}\right). \quad (25)$$

The moments of V_{ij} are bounded by

$$E\left(V_{ij}^{2k}\right) = h\int\left(\frac{u-u_1}{h}K^{*'}(z) - \frac{(u-u_1)^2}{2h^2}K^{*''}(z^*)\right)^{2k}$$
$$\times f_{XU_n}(x, x+zh, u, u_1)\mathrm{d}x\,\mathrm{d}z\,\mathrm{d}u\,\mathrm{d}u_1 \leqslant Ch^{1-2k}n^{-2k\gamma}. \quad (26)$$

Hence $E\left(\tilde{V}_{ij}^{2k}\right) \leqslant Cn^{(2k-1)/5-2k\gamma}$ and therefore by Rosenthal's inequality,

$$E\left(\frac{1}{n^2h^2}\sum_{i<j}\tilde{V}_{ij}(h)\right)^{2k} \leqslant Cn^{-k-2k\gamma}. \quad (27)$$

By analogous arguments, $E\left(\frac{1}{n^2h^2}\sum_{i<j}\tilde{W}_{ij}(h)\right)^{2k} \leqslant Cn^{-k-2k\gamma}$.

For the discretised case $E\left(V_{ij} + W_{ij}\right) = O(n^{-2\gamma}h^3) + O(n^{-\gamma}h^5)$ and the rest of the proof follows the same lines as in the continuous case.

Lemma 3 *For* $s, t \in [b_1 n^{-\frac{1}{5}}, b_2 n^{-\frac{1}{5}}]$,

$$E\left|n^{7/10}\left(D_3'(sn^{-1/5}) - D_3'(tn^{-1/5})\right)\right|^{2k} \leqslant C|s-t|^{2k}.$$

Proof In the continuous case, Taylor expansion for K^* gives

$$\frac{1}{sn^{-1/5}}K^*\left(\frac{X_i - X_j}{sn^{-1/5}}\right) - \frac{1}{tn^{-1/5}}K^*\left(\frac{X_i - X_j}{tn^{-1/5}}\right)$$
$$= n^{1/5}\left[\frac{t-s}{ts}K^*\left(\frac{X_i - X_j}{tn^{-1/5}}\right) + \frac{X_i - X_j}{sn^{-1/5}}\cdot\frac{t-s}{ts}K^{*'}\left(\frac{X_i - X_j}{tn^{-1/5}}\right)\right.$$
$$\left. + \frac{(X_i - X_j)^2}{2sn^{-2/5}}\cdot\frac{(t-s)^2}{t^2s^2}K^{*''}\left(\frac{X_i - X_j}{t^* n^{-1/5}}\right)\right]\cdot I\{|X_i - X_j| \leqslant |b_2|n^{-1/5}\},$$

and

$$K^*\left(\frac{X_i - X_j}{tn^{-1/5}}\right) - K^*\left(\frac{X_i - X_j + U_i - U_j}{tn^{-1/5}}\right)$$
$$= -\frac{U_i - U_j}{tn^{-1/5}}K^{*'}\left(\frac{X_i - X_j}{tn^{-1/5}}\right) + \frac{(U_i - U_j)^2}{2t^2n^{-2/5}}K^{*''}(z^*), \quad (28)$$

where t^* is between s and t and z^* is between $n^{1/5}(X_i-X_j)/t$ and $n^{1/5}(X_i-X_j+U_i-U_j)/t$. Therefore,

$$\left|\frac{\tilde{V}_{ij}(sn^{-1/5})}{sn^{-1/5}}-\frac{\tilde{V}_{ij}(tn^{-1/5})}{tn^{-1/5}}\right| \leqslant Cn^{1/5}|s-t|n^{1/5-\gamma}I\{|X_i-X_j|\leqslant tn^{-1/5}\}, \quad (29)$$

and

$$E\left|\frac{\tilde{V}_{ij}(sn^{-1/5})}{sn^{-1/5}}-\frac{\tilde{V}_{ij}(tn^{-1/5})}{tn^{-1/5}}\right|^{2k} \leqslant Cn^{(4k-1)/5-2k\gamma}|s-t|^{2k},$$

which combined with Rosenthal's inequality gives

$$E\left(\frac{1}{n^2h}\sum_{i<j}\frac{\tilde{V}_{ij}(sn^{-1/5})}{sn^{-1/5}}-\frac{\tilde{V}_{ij}(tn^{-1/5})}{tn^{-1/5}}\right)^{2k} \leqslant Cn^{-k-2k\gamma}|s-t|^{2k}.$$

In the same way, one obtains the bounds for $\tilde{W}_{ij}$,

$$E\left(\frac{1}{n^2h}\sum_{i<j}\frac{\tilde{W}_{ij}(sn^{-1/5})}{sn^{-1/5}}-\frac{\tilde{W}_{ij}(tn^{-1/5})}{tn^{-1/5}}\right)^{2k} \leqslant Cn^{-k-2k\gamma}|s-t|^{2k}.$$

To find bounds for moments of V_i, note that

$$\begin{aligned}
&\frac{1}{sn^{-\frac{1}{5}}}\int K^*\left(\frac{X_i-x}{sn^{-\frac{1}{5}}}\right)f_X(x)\mathrm{d}x-\frac{1}{tn^{-\frac{1}{5}}}\int K^*\left(\frac{X_i-x}{tn^{-\frac{1}{5}}}\right)f_X(x)\mathrm{d}x\\
&=\int K^*(z)\left[zn^{-\frac{1}{5}}(t-s)f_X'(X_i)+\frac{z^2n^{-\frac{2}{5}}}{2}(t^2-s^2)f_X''(X_i)+o(n^{-\frac{2}{5}})|s-t|\right]\mathrm{d}z\\
&=o\left(n^{-\frac{2}{5}}\right)|s-t|.
\end{aligned}$$

Thus by Rosenthal's inequality again,

$$E\left(\frac{1}{nh}\sum_{i=1}^{n}\frac{\tilde{V}_i(sn^{-1/5})}{sn^{-1/5}}-\frac{\tilde{V}_i(tn^{-1/5})}{tn^{-1/5}}\right)^{2k} \leqslant o\left(n^{-7k/5}\right),$$

where $\tilde{V}_i(h)=V_i(h)-EV_i(h)$. Since the arguments for W_i are same, we omit the details.

In the discretised case, we need to replace $(U_i-U_j)t^{-1}$ in (28) by $[(X_i-X_j)-(\tilde{Y}_i-\tilde{Y}_j)]t^{-1}$ and the inequality (29) remains true. □

Lemma 4 *There exists a positive number $\eta > 0$, such that*

$$\sup_{t\in[b_1,b_2]} |D_3'(tn^{-1/5})| = O_P(n^{-3/5-\eta}).$$

Proof Lemma 2 implies that $E|D_3'(n^{-1/5}t)|^{2k} \leqslant Cn^{-7k/5}$, therefore by the Chebyshev's inequality

$$P\left(|n^{3/5+\eta}D_3'(n^{-1/5}t)|^{2k} \geqslant \epsilon\right) \leqslant \frac{E|D_3'(n^{-1/5}t)|^{2k}}{\epsilon^{2k}n^{-6k/5-2k\eta}} \leqslant Cn^{-k/5+2k\eta} \to 0,$$

provided $\eta < 1/10$. □

Lemma 5 *There exists a positive number $\eta > 0$, such that*

$$\hat{h}_{Y,\mathrm{cv}} - h_{X,\mathrm{opt}} = O_P(n^{-1/5-\eta}).$$

Proof We assume that $\hat{h}_{Y,\mathrm{cv}} \in [b_1n^{-1/5}, b_2n^{-1/5}]$, hence by Lemma 4 and Lemma 3.2 in Hall and Marron (1987),

$$|D_1'(\hat{h}_{Y,\mathrm{cv}})| + |D_2'(\hat{h}_{Y,\mathrm{cv}})| + |D_3'(\hat{h}_{Y,\mathrm{cv}})| = O_P(n^{-3/5-\eta}). \tag{30}$$

Taylor expansion of MISE_X gives

$$\mathrm{MISE}_X'(h_{X,\mathrm{opt}})+(\hat{h}_{Y,\mathrm{cv}}-h_{X,\mathrm{opt}})\mathrm{MISE}_X''(h^*)=D_1'(\hat{h}_{Y,\mathrm{cv}})+D_2'(\hat{h}_{Y,\mathrm{cv}})+D_3'(\hat{h}_{Y,\mathrm{cv}})$$

where h^* lies between $\hat{h}_{Y,\mathrm{cv}}$ and $h_{X,\mathrm{opt}}$ and hence $h^* = O(n^{-1/5})$. Therefore, $\mathrm{MISE}_X''(h^*) = C\,n^{-2/5}$, which together with (30) imply that $\hat{h}_{Y,\mathrm{cv}} - h_{X,\mathrm{opt}} = O_P(n^{-1/5-\eta})$. □

Lemma 6 *For any $\eta > 0$,*

$$\sup_{|t-t_0|<n^{-1/5-\eta}} n^{7/10}|D_3'(t) - D_3'(t_0)| \to 0$$

in probability.

Proof This is proved in analogous way to the proof of Lemma 3.2 in Hall and Marron (1987) and details are omitted here. □

Lemma 7 *If $\gamma > \gamma_0$, then*

$$n^{7/10}D_3'(h_{X,\mathrm{opt}}) \to 0$$

in probability. If $\gamma = \gamma_0$, then

$$n^{7/10}\left(D_2'(h_{X,\mathrm{opt}}) + D_3'(h_{X,\mathrm{opt}})\right) \to N(0, \sigma^2_{Y,\mathrm{cv}}) \tag{31}$$

in distribution.

Proof If $\gamma > \gamma_0$, then Lemma 2 implies that $\mathrm{Var}(D_3'(h_{X,\mathrm{opt}})) = o(n^{-7/5})$ and therefore $n^{7/10}D_3'(h_{X,\mathrm{opt}}) \to 0$ in probability.

Assume now that $\gamma = \gamma_0$. In the continuous case, using the decomposition (24), denote

$$T_{11} = \frac{1}{n^2h^2}\sum_{i<j}\left(\tilde{V}_{ij}(h) + \tilde{W}_{ij}(h)\right),$$
$$T_{12} = \frac{1}{n^2h^2}\sum_{i<j}\left(\tilde{B}_{1,ij}(h) + \tilde{B}_{2,ij}(h)\right),$$
$$T_2 = \frac{1}{nh^2}\sum_{i=1}^{n}\left(\tilde{B}_{1,i}(h) + \tilde{B}_{2,i}(h)\right),$$

where $\tilde{B}_{k,ij} = B_{k,ij}-B_{k,i}-B_{k,j}+EB_{k,ij}$, $\tilde{B}_{k,i} = B_{k,i}-EB_{k,i}$ and $B_{k,i} = E[B_{k,ij}|X_i]$ with $B_{1,ij} = K\left(\frac{X_i-X_j}{h}\right)$ and $B_{2,ij} = \frac{X_i-X_j}{h}K'\left(\frac{X_i-X_j}{h}\right)$. Set $T_1 = T_{11} + T_{12}$. Then $D_2'+D_3' = T_1+T_2+I_3+I_4+R$. Note that (25) implies $\mathrm{Var}(I_3) = o(n^{-7/5})$, hence to prove (31), it suffices to show that $n^{7/10}\,(T_1 + T_2) \to N(0,\sigma^2_{Y,\mathrm{cv}})$. It is easy to see that T_1 and T_2 are uncorrelated. Write $T_1 = \sum_{i<j}A(i,j)$ and $T_2 = \sum_{i=1}^{n}a(i)$ and define $Q_i = \sum_{j=i+1}^{n}A(i,j) + a(i)$, then $EQ_i = 0$ and $T_1 + T_2 = \sum_{i=1}^{n}Q_i$ is a martingale with σ-field $\mathcal{F}_{n,i} = \sigma\{X_j, U_j, j = 1,\dots,\min(i,n)\}$. Note that the variables $A(i,j), A(i,k)$ are uncorrelated if i,j,k are all different and

$$\mathrm{Var}(A(1,2)) = \left[\frac{2\mu_{U,2}}{n^{4+2\gamma}h^5}\left(2R(K^{*'}) + R(K_1^*)\right) + \frac{1}{n^4h^3}R(K_1)\right]R(f_X) + o(n^{-\frac{1}{5}}),$$
$$\mathrm{Var}(a(1)) = \frac{h^2}{n^2}\sigma_K^4\left(\int\left(f_X''(x)\right)^2 f_X(x)\mathrm{d}x - \left(\int f_X''(x)f_X(x)\mathrm{d}x\right)^2\right). \tag{32}$$

Hence

$$s_n^2 \equiv \mathrm{Var}(T_1 + T_2) = \sum_{i=1}^{n}\mathrm{Var}(Q_i) = \frac{n(n-1)}{2}\mathrm{Var}(A(1,2)) + n\mathrm{Var}(a(1)). \tag{33}$$

Since also Q_is are independent from $\mathcal{F}_{n,i-1}$,

$$s_n^{-2}\sum_i E\left[Q_i^2|\mathcal{F}_{n,i-1}\right] \to 1, \tag{34}$$

in probability.

By Chebyshev's inequality and (32),

$$P(|Q_i| > s_n\epsilon) \leqslant E|Q_i|^2 s_n^{-2}\epsilon^{-2} = O\left(n^{-1}\right), \tag{35}$$

and by the Cauchy–Schwartz inequality,

$$E\left[Q_i^2 I\{|Q_i| > s_n\epsilon\}\right] \leqslant (EQ_i^4)^{1/2} P^{1/2}(|Q_i| > s_n\epsilon) \leqslant O(n^{-1/2} EQ_i^2). \tag{36}$$

Hence

$$s_n^{-2}\sum_{i=1}^{n} E\left[Q_i^2 I\{|Q_i| > s_n\epsilon\}|\mathcal{F}_{n,i-1}\right] \leqslant \frac{nEQ_i^2}{n\mathrm{Var}(Q_i)} \cdot O(n^{-1/2}) = O(n^{-1/2}) \to 0.$$

This and (34) imply, using Brown's theorem (Hall and Heyde, 1980, p. 58), that $T_1 + T_2 \to N(0, s_n^2)$ in distribution. Substituting $h = h_{X,\mathrm{opt}} = C_h n^{-1/5}$ in the expressions for variances in (32) finishes the proof of (31).

In the discretised case,

$$\begin{aligned}\mathrm{Var}(A(1,2)) = {} & \frac{1}{n^4h^3}\int\left(z - \sum_{j=-\infty}^{+\infty}\frac{\tilde{x}_j\left(I_{S_{j(x)}}(z) - I_{S_{1j}}(x)\right)}{h}\right)^2 \\ & \times \left(2K^{*'}(z) + zK^{*''}(z)\right)^2 f_X^2(x)\mathrm{d}x\,\mathrm{d}z \\ & + \frac{2\mu_{U,2}}{n^{4+2\gamma}h^5}\left(2R(K^{*'}) + R(K_1^*)\right)R(f_X) + \frac{1}{n^4h^3}R(K_1)R(f_X) + o(n^{-1/5})\end{aligned}$$

and the rest of the proof remains the same and is omitted here.

Proof of Proposition 1
Proof of (12) The proof is analogous to the proof of Theorem 3 with the function $D_3(h)$ replaced by

$$D_{3L}(h) = M_X(h) - \bar{M}_{Y,L}(h) = \frac{1}{n^2h}\sum_{i,j}\left[K^*\left(\frac{X_i - X_j}{h}\right) - \frac{1}{L}\sum_{l=1}^{L}K^*\left(\frac{Y_i^l - Y_j^l}{h}\right)\right],$$

where $Y_i^l = X_i + U_i^l$. We also need to redefine V_{ij} and W_{ij}. The terms $K^*((Y_i - Y_j)h^{-1})$ and $K_1^*((Y_i - Y_j)h^{-1})$ in the definition of V_{ij} and W_{ij} are replaced by the means $L^{-1}\sum_{l=1}^{L} K^*((Y_i^l - Y_j^l)h^{-1})$ and $L^{-1}\sum_{l=1}^{L} K_1^*((Y_i^l - Y_j^l)h^{-1})$, respectively.

In the proof of Lemma 3, the differences $U_i - U_j$ and $(U_i - U_j)^2$ in the right-hand side of (28) need to be replaced by the means $L^{-1}\sum_{l=1}^{L} U_i^l - U_j^l$ and $L^{-1}\sum_{l=1}^{L}(U_i^l - U_j^l)^2$. All the following inequalities remain the same in this proof.

In the second part of Lemma 7, the terms $A(i,j)$ are dependent on L and their variance is expanded to

$$\mathrm{Var}(A(1,2)) = \left[\frac{2\mu_{U,2}}{Ln^{4+2\gamma}h^5}\left(2R(K^{*'}) + R(K_1^*)\right) + \frac{1}{n^4h^3}R(K_1)\right]R(f_X) + o(n^{-\frac{1}{5}}),$$

which with $h = C_h n^{-\frac{1}{5}}$ implies that $n^{\frac{7}{10}} \left(D_2'(h_{X,\mathrm{opt}}) + D_{3L}'(h_{X,\mathrm{opt}})\right) \to N(0, \sigma^2_{L,\mathrm{cv}})$, in distribution.

Proof of (13) Equation (23) holds for each of the bandwidths $\hat{h}_{l,\mathrm{cv}}$ and thus

$$\left(\bar{h}_{\mathrm{cv},L} - \hat{h}_{X,\mathrm{opt}}\right) \mathrm{MISE}_X''(h^*) = D_2'(h_{X,\mathrm{opt}}) + D_{3L}'(h_{X,\mathrm{opt}}) + o(n^{-7/10}),$$

which is the same as the analogous expansion for the bandwidth $\hat{H}_{L,\mathrm{cv}}$, hence (13) holds.

References

Andrews, D. F., Herzberg, A. M. (1985). *Data.* Berlin Heidelberg New York: Springer.
Bowman, A. W. (1984). An alternative method of cross-validation for the smoothing of density estimates. *Biometrika 71*, 353–360.
Brinkman, N. D. (1981). Ethanol fuel – a single-cylinder engine study of efficiency and exhaust emissions. *SAE Transactions 90*, 1410–1424.
Chiu, S. T. (1991). The effect of discretization error on bandwidth selection for kernel density estimation. *Biometrika 78*, 436–441.
Hall, P. (1983). Large sample optimality of least squares cross-validation in density estimation, *Annals of Statistics 11*, 1156–1174.
Hall, P., Heyde, C. C. (1980). *Martingale Limit Theory and its Application*. New York–London: Academic Press
Hall, P., Marron, J. S. (1987). Extent to which least-squares cross-validation minimises integrated square error in nonparametric density estimation. *Probability Theory and Related Fields 74*, 567–581.
Jones, M. C., Marron, J. S., Sheather, S. J. (1996). A brief survey of bandwidth selection for density estimation. *Journal of the American Statistical Association 91,* 401–407.
Loader, C. R. (1999). Bandwidth selection: classical or plug-in? *Annals of Statistics 27,* 415–438.
Machado, J. A. F., Santos Silva, J. M. C. (2002). Quantiles for Counts. No. CWP22/02 in CeMMAP, Centre for Microdata Methods and Practice, Institute for Fiscal Studies, Depertment of Economics, UCL.
Marron, J. S., Wand, M. P. (1992). Exact mean integrated squared error. *Annals of Statistics 20,* 712–736.
Rosenblatt, M. (1956). Remarks on some nonparametric estimates of a density function. *Annals of Mathematical Statistics 27,* 832–837.
Ruppert, D., Carroll, R. J. Maca, J. D. (1999). Nonparametric regression in the presence of measurement error. *Biometrika 86,* 541–554.
Silverman, B. W. (1986). *Density Estimation for Statistics and Data Analysis*. In, *Monographs on Statistics and Applied Probability*. London: Chapman and Hall.
Stone, C. J. (1984). An asymptotically optimal window selection rule for kernel density estimates. *Annals of Statistics 12,* 1285–1297.