

EXPLICIT SOLUTIONS FOR THE ASYMPTOTICALLY-OPTIMAL BANDWIDTH IN CROSS VALIDATION

Karim Abadir, Michel Lubrano

▶ **To cite this version:**

Karim Abadir, Michel Lubrano. EXPLICIT SOLUTIONS FOR THE ASYMPTOTICALLY-OPTIMAL BANDWIDTH IN CROSS VALIDATION. 2010. halshs-00472750

HAL Id: halshs-00472750

https://shs.hal.science/halshs-00472750

Preprint submitted on 13 Apr 2010

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

GREQAM

Groupement de Recherche en Economie
Quantitative d'Aix-Marseille - UMR-CNRS 6579
Ecole des Hautes Etudes en Sciences Sociales
Universités d'Aix-Marseille II et III

Document de Travail n°2010-06

EXPLICIT SOLUTIONS FOR THE ASYMPTOTICALLY-OPTIMAL BANDWIDTH IN CROSS VALIDATION

Karim ABADIR
Michel LUBRANO

February 2010

DT-GREQAM

Explicit solutions for the asymptotically-optimal bandwidth in cross validation

Karim M. Abadir and Michel Lubrano*

Imperial College London and GREQAM-CNRS

February 2010

*Karim M. Abadir, Imperial College, London SW7 2AZ, UK (email: k.m.abadir@imperial.ac.uk); Michel M. Lubrano, Greqam-Cnrs, Centre de la Vieille Charité, Marseille F-13002, France (email:michel.lubrano@univmed.fr). We are grateful for comments received at the Bernoulli (ISI) Conference on Advances in Semiparametric Methods and Applications (Lisbon), London-OxBridge Meeting, York Conference in honour of Mike Wickens, and seminars at Oxford and Greqam. Support from the ESRC grants RES000230176 and RES062230790 is gratefully acknowledged. This paper was written on the occasion of three visits made by Karim Abadir in Marseille at the invitation of the Université de la Méditérannée.

Abstract

Least squares cross-validation (CV) methods are often used for automated bandwidth selection. We show that they share a common structure which has an explicit asymptotic solution. Using the framework of density estimation, we consider unbiased, biased, and smoothed CV methods. We show that, with a Student t(ν) kernel which includes the Gaussian as a special case, the CV criterion becomes asymptotically equivalent to a simple polynomial. This leads to optimal-bandwidth solutions that dominate the usual CV methods, definitely in terms of simplicity and speed of calculation, but also often in terms of integrated squared error because of the robustness of our asymptotic solution. We present simulations to illustrate these features and to give practical guidance on the choice of ν.

KEY WORDS: bandwidth choice; cross validation; nonparametric density estimation; analytical solution.

1 Introduction

Let $\{x_i\}_{i=1}^n$ be an i.i.d. sequence with unknown common density f that is a continuous function. The kernel density estimator introduced by Rosenblatt (1956) is given by

$$\widehat{f}(u) := \frac{1}{nh} \sum_{i=1}^{n} K\left(\frac{u - x_i}{h}\right),$$

where h is the bandwidth and K is the kernel. We will assume that the kernel is nonnegative, in which case the scaled kernels $K_h(u - x) := h^{-1}K(h^{-1}(u - x_i))$ are proper p.d.f.s and

$$\widehat{f}(u) = \frac{1}{n} \sum_{i=1}^{n} K_h(u - x_i) \tag{1}$$

is the sample mean of these. It is widely recognized that a variety of kernels (including the Gaussian) have good asymptotic efficiencies compared to the optimal one, the Epanechnikov kernel, whereas the choice of the bandwidth is crucial. For example, using the Gaussian instead of the Epanechnikov, the asymptotic mean integrated squared error (AMISE) is multiplied by a factor of $(6\sqrt{\pi/125})^{-4/5} \approx 1.04$, implying a *relative* loss of only 4% and an absolute loss that vanishes at the rate of $n^{-4/5}$. Subject to some regularity conditions, optimizing the AMISE gives

$$\widehat{h} = \left(\frac{k_{02}}{k_{21}^2 I_2}\right)^{1/5} n^{-1/5}, \tag{2}$$

where $k_{ij} := \int_{-\infty}^{\infty} t^i K(t)^j \, \mathrm{d}t$ and $I_j := \int_{-\infty}^{\infty} f^{(j)}(u)^2 \, \mathrm{d}u$, with $f^{(j)}(\cdot)$ denoting the j-th derivative of $f(\cdot)$. Plug-in methods substitute estimates for the remaining unknown quantity I_2 of (2) by using a nonparametric estimate, as in Hall and Marron (1987) or Jones and Sheather (1991); but they can go as far as replacing f in I_2 by a Gaussian density, a method commonly referred to as the rule of Silverman (1986).

Rudemo (1982) and Bowman (1984) introduced the least squares cross-validation (CV) method to determine the bandwidth that minimizes the integrated squared

error (ISE) asymptotically. The formula for the ISE is

$$\begin{aligned} \text{ISE} &:= \int_{-\infty}^{\infty} (f(u) - \widehat{f}(u))^2 \, \mathrm{d}u \\ &= \int_{-\infty}^{\infty} f(u)^2 \, \mathrm{d}u + \int_{-\infty}^{\infty} \widehat{f}(u)^2 \, \mathrm{d}u - 2 \int_{-\infty}^{\infty} \widehat{f}(u) f(u) \, \mathrm{d}u, \end{aligned} \tag{3}$$

where all three components are assumed finite with probability 1. The first integral in (3) does not affect the procedure and can be omitted from the optimization. The second integral is in terms of the data (known) and the h to be optimized. However, the last one contains the unknown density. CV overcomes this problem by considering an alternative criterion that has the same expectation as the ISE and is based on a resampling scheme. The validity of this method relies on a strong result by Stone (1984) which shows that the ISEs with optimal h (unknown in practice) and with h obtained by CV coincide asymptotically. But the speed of convergence is rather slow. The method is said to suffer from a great deal of sample variability, and it is costly to compute for large samples. Silverman (1982) proposed to use the fast Fourier transform as an approximation for reducing computational cost, while Härdle and Scott (1992) recommended binning techniques.

This CV criterion is an unbiased estimator of the mean integrated squared error (MISE), and we shall refer to it as unbiased CV (UCV) to stress this. The biased CV (BCV) criterion proposed by Scott and Terrell (1987) is a biased estimator of the MISE, but it reduces the sample variability of the UCV criterion. It was derived as a method of estimating the unknown integral in the denominator of (2), and it minimizes the same AMISE objective function.

The BCV of Scott and Terrell (1987) was followed by a number of alternative BCVs; including the modified CV of Stute (1992), the smoothed CV (SCV) of Hall, Marron and Park (1992) and its extension in Jones, Marron and Park (1991). The latter is particularly interesting because it derives the functional form of an additional bandwidth that helps CV achieve the fastest rate of convergence relative to $\widehat{h}$, a rate that was established by Hall and Marron (1991) as $\sqrt{n}$.

None of these methods give an explicit solution for the optimal h. Furthermore, there is a common structure to all these CV methods, not just in density estimation but also in nonparametric regression; e.g. see Li and Racine (2006). In fact, it is a structure that is also shared by other problems, such as the determination of bandwidths in the estimation of spectra; inter alia, see Velasco (2000), the widespread Newey and West (1987) method that requires the estimation of spectra at the origin, and the more recent one by Robinson (2005).

In Section 2, we introduce a method to obtain explicit solutions for asymptotically optimal bandwidths in problems sharing this common structure. In Sections 3–5, we apply it to solving for the optimal bandwidths in UCV, BCV, and the SCV version of Jones, Marron and Park (1991), respectively. In Section 6, we present simulations to illustrate the finite-sample robustness of the results to various densities and to give guidance on choices that need to be made in practice when implementing our method of solution. We confirm that our simple explicit solutions (one for each of UCV, BCV, SCV) for the asymptotic bandwidth are very efficient (in terms of ISE) and robust, solving CV's notorious sampling variability problem as well as giving huge numerical efficiency gains. Section 7 concludes. An appendix collects some derivations that are needed in the text.

2 Method for explicit solution of bandwidths

Let $*$ denote the convolution symbol. UCV, BCV, and their variants require the calculation of

$$K^{(q)} * K^{(r)}, \tag{4}$$

where $q, r \in \mathbb{Z}_{0,+}$, the nonegative integers. Define

$$D_h := K_h - K_0, \tag{5}$$

where K_0 is the Dirac delta function. SCV and its variants introduce an additional kernel L with bandwidth g, requiring the calculation of

$$D_h * D_h * L_g * L_g, \tag{6}$$

where L_g is the scaled version of kernel L such that $L_g(t) := g^{-1}L(g^{-1}t)$, the optimal g taking the form

$$\widehat{g} \sim Cn^p/\widehat{h}^2 \tag{7}$$

with C constant as $n \to \infty$ and p a constant to be detailed in Section 5. The notation $a_n \sim b_n$ means that $\lim_{n\to\infty} a_n/b_n = 1$.

There are two components to the solution. The first one is straightforward once we recall that the choice of a Gaussian kernel function has little effect on asymptotic efficiency while allowing simple explicit solutions, in which case we take $K = L = \phi$ to work out (4) and (6). To do so will require the Hermite polynomials

$$He_m(t) := \frac{(-1)^m \phi^{(m)}(t)}{\phi(t)} = t^m \sum_{j=0}^{1+\lfloor m/2 \rfloor} \frac{(-m)_{2j}}{j!(-2t^2)^j}, \tag{8}$$

where $m \in \mathbb{Z}_{0,+}$, $\lfloor m/2 \rfloor$ denotes the integer part of $m/2$, and

$$(-m)_{2j} := \prod_{i=1}^{2j} (-m + i - 1);$$

see Abadir (1999) for more details on He_m polynomials and their relation to the other type of Hermite polynomials denoted by H_m.

Lemma 1 *For $K = L = \phi$,(4) and (6) become, respectively,*

$$\left(K^{(q)} * K^{(r)}\right)(a) = (-1)^{q+r} \frac{K_{\sqrt{2}}(a)\, He_{q+r}\left(a/\sqrt{2}\right)}{\sqrt{2^{q+r}}} \tag{9}$$

and

$$(D_h * D_h * L_g * L_g)(a) = K_{\sqrt{2h^2+2g^2}}(a) - 2K_{\sqrt{h^2+2g^2}}(a) + K_{g\sqrt{2}}(a), \tag{10}$$

where a is the argument of the convolution and $K_b(t) := b^{-1}K(b^{-1}t)$.

Proof. See the appendix.

The second component of the solution is to find a way to achieve the asymptotic separability (in h and t) of a kernel $K(t/h)$. This allows a factorization of first-order conditions for h. This does not hold for ϕ, but it holds for another p.d.f. that can be made arbitrarily close to ϕ and that can be used instead of ϕ after the convolutions have been worked out as in the previous lemma.

Consider a Student t(ν) kernel, $K(t) = c_\nu/(1 + t^2/\nu)^{(\nu+1)/2}$, where

$$c_\nu := \Gamma\left(\frac{\nu+1}{2}\right) / \left(\sqrt{\pi\nu}\Gamma\left(\frac{\nu}{2}\right)\right). \tag{11}$$

The Gaussian kernel is the limiting t(∞) case, but $\nu = 30$ makes the two virtually indistinguishable for all practical purposes. The scaled version of this Student t(ν) kernel is

$$K_h(t) = \frac{c_\nu}{h(1 + t^2/(\nu h^2))^{(\nu+1)/2}} = \frac{c_\nu}{(h^2 + t^2/\nu)^{(\nu+1)/2}} h^\nu. \tag{12}$$

As $\widehat{h} = O_p(n^{-1/5}) \xrightarrow{p} 0$, (12) becomes asymptotically separable in t and h:

$$K_h(t) = c_\nu \left(t^2/\nu\right)^{-(\nu+1)/2} h^\nu \left(1 + O(h^2)\right)$$

as $h \to 0$ and ν is finite. This is implied by the binomial expansion, as

$$(h^2 + a)^{-b} = a^{-b} + O(h^2), \quad (a, b \text{ finite and } h \to 0), \tag{13}$$

which we will need again later. This quasi-separability for small h does not hold in the limiting $\nu = \infty$ Gaussian case, but it nevertheless holds for any fixed large ν that makes t(ν) indistinguishable from the Gaussian. This allows the subsequent derivations to give an explicit formula for the asymptotically optimal $\widehat{h}$. The only available expansion for the Gaussian kernel relies on $\exp(-t^2/(2h^2)) = 1 - t^2/(2h^2) + \ldots$, which is not valid for $h \to 0$. To use the terminology of complex analysis, $h = 0$ is an "essential singularity" of the function. The binomial expansion of the Student t(ν) kernel does not suffer this drawback, even for any arbitrarily large but finite ν.

We are now in a position to apply these results to optimal bandwidth selection in CV problems.

3 Application 1: UCV

This section contains three parts to clarify the ideas in this first application. This level of detail will not be given for applications 2 (BCV) and 3 (SCV).

First, we rewrite the UCV criterion by using Lemma 1. Second, we analyze the criterion to shed light on the asymptotic behaviour of its components, and this results in some straightforward approximations for the optimal $\widehat{h}$. This solution allows us to determine the required orders of magnitude and understand how the method works. Third, this analysis leads to an asymptotic representation of the first-order conditions for $\widehat{h}$ as simple polynomials for which we give solutions that are numerically-efficient, of the order of 24 times faster than CV methods. Our solutions are also often more accurate in terms of minimizing the ISE, as will be seen in Section 6.

3.1 UCV criterion

The first step of the UCV procedure is to delete one observation at a time, say x_j $(j = 1, \ldots, n)$, then calculate the usual kernel estimator based on the remaining $n-1$ data points

$$\widehat{f}_{-j}(u) := \frac{1}{n-1} \sum_{i \neq j} K_h(u - x_i), \quad j = 1, \ldots, n. \tag{14}$$

The last integral in the ISE in (3) is an expectation which can be estimated by using the sample mean of (14)

$$\overline{\widehat{f}}_{n-1}(\boldsymbol{x}; h) := \frac{1}{n} \sum_{j=1}^{n} \widehat{f}_{-j}(x_j) = \frac{1}{n(n-1)} \sum_{j=1}^{n} \sum_{i \neq j} K_h(z_{ij}), \tag{15}$$

where $\boldsymbol{x} := (x_1, \ldots, x_n)'$, denoting a transpose by a prime, and

$$z_{ij} := x_j - x_i. \tag{16}$$

A matrix with typical element z_{ij} would be skew-symmetric.

UCV minimizes with respect to h the sum $S := S_1 + S_2 + S_3$, where

$$S_1 := \int_{-\infty}^{\infty} f(u)^2 \, \mathrm{d}u, \quad S_2 := \int_{-\infty}^{\infty} \widehat{f}(u)^2 \, \mathrm{d}u, \quad S_3 := -2\overline{\widehat{f}}_{n-1}(\boldsymbol{x}; h).$$

This procedure is justified by the fact that $\mathrm{E}(S) = \mathrm{E}(\mathrm{ISE})$, the latter being the definition of the MISE. Since $S_1 > 0$ and does not depend on n, it does not tend to 0 as $n \to \infty$ and we need

$$S_2 + S_3 \xrightarrow{p} -S_1 < 0 \tag{17}$$

for consistency of $\widehat{f}$.

Using Lemma 1, we can work out

$$S_2 = \frac{1}{n} K_{h\sqrt{2}}(0) + \frac{2}{n^2} \sum_{j=1}^{n} \sum_{i>j} K_{h\sqrt{2}}(z_{ij}),$$

where we separated out the term having $i = j$ and used the fact that K is an even function of z_{ij} to rewrite the range of the inner summation ($\sum_{i \neq j} = 2 \sum_{i>j}$). Using $n/(n-1) = 1 + O(1/n)$ gives

$$S_2 + S_3 = \frac{K_{h\sqrt{2}}(0)}{n} + \frac{2 + O(1/n)}{n^2} \sum_{j=1}^{n} \sum_{i>j} \left[K_{h\sqrt{2}}(z_{ij}) - 2K_h(z_{ij}) \right], \tag{18}$$

where the first fraction is deterministic and of order $1/(nh)$. We now apply the second idea of the previous section, the $\mathrm{t}(\nu)$ kernel, in order to tackle the optimization.

3.2 Asymptotic approximation

From the scaled Student $\mathrm{t}(\nu)$ kernel in (12), $K_{h\sqrt{2}}(0) = c_\nu/(h\sqrt{2})$. Applying (17) to (18), and the fact that the UCV-optimal h is $\widehat{h} = O_p(n^{-1/5})$, it follows that the first term of (18) drops out asymptotically and the second term has a strictly negative and finite probability limit. In this subsection, we will therefore minimize

$$R := 2 \sum_{j=1}^{n} \sum_{i>j} K_{h\sqrt{2}}(z_{ij}) - 4 \sum_{j=1}^{n} \sum_{i>j} K_h(z_{ij}), \tag{19}$$

where $R/n^2 \xrightarrow{p} -S_1 < 0$ at the optimum and so each of the two terms in (19) is of order n^2 or *larger* (but with cancelling leading terms). We now exploit this remark.

The objective function (19) with a $\mathrm{t}(\nu)$ kernel becomes

$$R = 2c_\nu h^\nu \sum_{j=1}^{n} \sum_{i>j} \left[2^{\nu/2} \left(2h^2 + z_{ij}^2/\nu \right)^{-(\nu+1)/2} - 2 \left(h^2 + z_{ij}^2/\nu \right)^{-(\nu+1)/2} \right]. \tag{20}$$

Differentiating (20) with respect to h, we get the first-order condition

$$\nu \sum_{j=1}^{n} \sum_{i>j} \left[2^{\nu/2} \left(2\widehat{h}^2 + z_{ij}^2/\nu \right)^{-(\nu+1)/2} - 2 \left(\widehat{h}^2 + z_{ij}^2/\nu \right)^{-(\nu+1)/2} \right] \tag{21}$$

$$= \quad 2\left(\nu+1\right)\widehat{h}^2 \sum_{j=1}^{n} \sum_{i>j} \left[2^{\nu/2} \left(2\widehat{h}^2 + z_{ij}^2/\nu \right)^{-(\nu+3)/2} - \left(\widehat{h}^2 + z_{ij}^2/\nu \right)^{-(\nu+3)/2} \right].$$

Applying $R = O_p(n^2)$ and $\widehat{h} = O_p(n^{-1/5})$ to (20) and recalling the remark following (19), the function

$$y_n(q, \widehat{h}) := \sum_{j=1}^{n} \sum_{i>j} \left(\widehat{h}^2 + z_{ij}^2/\nu \right)^{-(q+1)/2} \tag{22}$$

has order $n^{2+\nu/5}$ or larger for $q = \nu$. Therefore, the leading term of the expansion of (22) is the one from which $\widehat{h}$ is absent and it is the one obtained by using the binomial expansion (13). In other words, any small values of the same order as $\widehat{h}$ will do, asymptotically, for (22), and we will explore now two such possibilities for a plug-in that we will denote generically by $\widehat{h}_{\mathrm{p}}$.

First, we could substitute Silverman's (1986) rule of thumb $\widehat{h} = 1.06\widehat{\sigma}n^{-1/5}$ mentioned after (2), with $\widehat{\sigma}^2$ denoting the sample variance of $\{x_i\}_{i=1}^n$. A more elaborate version would use again (2) but with f replaced by a Student density instead of the Gaussian. The ingredients for this are in Lemma 2 of the appendix, and they give

$$\widehat{h}_{\mathrm{S}} := \left(\frac{4\left(1-2/\nu\right)^{9/2}\left(\nu-3/16\right)^2\left(\nu+17/8\right)\left(\nu+5/2\right)\left(\nu+7/2\right)}{3\left(\nu-1/4\right)\left(\nu+1\right)^2\left(\nu+3\right)^2} \right)^{1/5} \widehat{\sigma} n^{-1/5} \tag{23}$$

with $\lim_{\nu\to\infty} \widehat{h}_{\mathrm{S}} / \left(\widehat{\sigma} n^{-1/5}\right) = (4/3)^{1/5} \approx 1.06$ implying Silverman's rule as a special case. Second, we could generalize the popular method of Jones and Sheather (1991), using a Student (rather than Gaussian) density and kernel, resulting in an estimate of I_2 given by

$$\widehat{I}_2 := \frac{1}{n^2\widehat{\lambda}^5} \sum_{i,j} K^{(4)}\left(\frac{z_{ij}}{\widehat{\lambda}} \right) \tag{24}$$

$$= \frac{\left(4\nu-1\right)\left(\nu+1\right)\left(\nu+3\right)}{4\sqrt{2\pi}n^2\widehat{\lambda}^5\nu^5} \sum_{i,j} \frac{\left(\nu+2\right)\left(\nu+4\right)z_{ij}^4/\widehat{\lambda}^4 - 6\nu\left(\nu+4\right)z_{ij}^2/\widehat{\lambda}^2 + 3\nu^2}{\left(1 + z_{ij}^2/\left(\widehat{\lambda}^2\nu\right)\right)^{(\nu+9)/2}}$$

with $\lambda := \left(2K^{(4)}(0)/\left(nI_3k_{21}\right)\right)^{1/7}$ estimated by

$$\widehat{\lambda} := \left(\frac{\sqrt{2}\left(\nu-2\right)^{9/2}\left(2\nu+7\right)\left(2\nu+9\right)\left(2\nu+11\right)\left(8\nu+25\right)}{5\nu^{7/2}\left(\nu+1\right)\left(\nu+3\right)\left(\nu+5\right)^2\left(4\nu-1\right)}\right)^{1/7}\widehat{\sigma}n^{-1/7} \tag{25}$$

leading to

$$\widehat{h}_{\mathrm{JS}} := \left(\frac{\left(\nu-2\right)^2\left(16\nu-3\right)^2\left(4\nu-1\right)}{\sqrt{\pi}2^{11}\nu^5\widehat{I}_2}\right)^{1/5} n^{-1/5} \tag{26}$$

with the ingredients derived in Lemma 3 of the appendix.

Exploiting the asymptotic invariance of the $y_n(\cdot,\cdot)$ function, we can rewrite the solution (21) as

$$\widehat{h} = \sqrt{\frac{\nu\left[2^{\nu/2}y_n(\nu,\widehat{h}_{\mathrm{p}}\sqrt{2}) - 2y_n(\nu,\widehat{h}_{\mathrm{p}})\right]}{2\left(\nu+1\right)\left[2^{\nu/2}y_n(\nu+2,\widehat{h}_{\mathrm{p}}\sqrt{2}) - y_n(\nu+2,\widehat{h}_{\mathrm{p}})\right]}}, \tag{27}$$

where the RHS makes use of a plug-in $\widehat{h}_{\mathrm{p}}$, be it $\widehat{h}_{\mathrm{S}}$ or $\widehat{h}_{\mathrm{JS}}$, giving an explicit asymptotic solution for $\widehat{h}$. Note that $R/n^2 \xrightarrow{p} -S_1 < 0$ implies that the numerator and denominator should both be negative at the optimum, thus restricting the allowable solutions for h. Note also that $z_{ij}^2/\nu = \left(x_j - x_i\right)^2/\nu$, appearing in $y_n(\nu,\widehat{h})$ of (22), is a measure of distance between the data points. It is quadratic because of the adoption of a spherical p.d.f. as a kernel, and this applies more generally to other spherical kernels.

3.3 Exact solution

Omitting only the term denoted by $O(1/n)$ in (18), but not the first deterministic term which is now $c_\nu/(nh\sqrt{2})$, similar derivations lead to the first-order condition

$$\begin{aligned}\frac{n}{2\sqrt{2}} &= \nu\widehat{h}_{\mathrm{u}}^{\nu+1}\left[2^{\nu/2}y_n(\nu,\widehat{h}_{\mathrm{u}}\sqrt{2}) - 2y_n(\nu,\widehat{h}_{\mathrm{u}})\right] \\ &\quad -2\left(\nu+1\right)\widehat{h}_{\mathrm{u}}^{\nu+3}\left[2^{\nu/2}y_n(\nu+2,\widehat{h}_{\mathrm{u}}\sqrt{2}) - y_n(\nu+2,\widehat{h}_{\mathrm{u}})\right],\end{aligned} \tag{28}$$

where $\widehat{h}_{\mathrm{u}}$ is the UCV solution. As before, the content of the square brackets can be accurately approximated by using $\widehat{h}_{\mathrm{p}}$. This makes (28) an equation of the form

$\alpha_1 = \alpha_2 \widehat{h}^{\nu+1} + \alpha_3 \widehat{h}^{\nu+3}$, where

$$\begin{aligned} \alpha_1 &= \frac{n}{2\sqrt{2}}, \quad \alpha_2 = \nu \left[2^{\nu/2} y_n(\nu, \widehat{h}_{\mathrm{p}}\sqrt{2}) - 2y_n(\nu, \widehat{h}_{\mathrm{p}}) \right], \\ \alpha_3 &= -2(\nu+1) \left[2^{\nu/2} y_n(\nu+2, \widehat{h}_{\mathrm{p}}\sqrt{2}) - y_n(\nu+2, \widehat{h}_{\mathrm{p}}) \right], \end{aligned}$$

which is easy to solve numerically. An alternative form of writing $\alpha_1 = \alpha_2 \widehat{h}^{\nu+1} + \alpha_3 \widehat{h}^{\nu+3}$ is

$$\widehat{h} = \left(\frac{\alpha_1}{\alpha_2 + \alpha_3 \widehat{h}^2} \right)^{1/(\nu+1)}, \tag{29}$$

which can be approximated for $\nu > 2$ by using $\widehat{h}_{\mathrm{p}}^2$ of (23) or (26) on the RHS, giving an explicit asymptotic formula for $\widehat{h}$ which we will call $\widehat{h}_{\mathrm{a}}$:

$$\boxed{\widehat{h}_{\mathrm{a}} := \left(\frac{\alpha_1}{\alpha_2 + \alpha_3 \widehat{h}_{\mathrm{p}}^2} \right)^{1/(\nu+1)}}. \tag{30}$$

One should note however that this solution exists if and only if $\alpha_2 + \alpha_3 \widehat{h}_{\mathrm{p}}^2 > 0$, a condition which is guaranteed in large samples, but might fail in small samples. In this case, the simpler asymptotic approximation (27), reexpressed as

$$\widehat{h}_{\mathrm{aa}} := \sqrt{-\alpha_2/\alpha_3}, \tag{31}$$

should be used. Note that iterating (29), instead of using $\widehat{h}_{\mathrm{p}}^2$ in (30), would give the exact UCV solution except for the inconsequential approximation of $1/(n-1)$ by $1/n$ in the objective function (18).

4 Application 2: BCV

Scott and Terrell (1987) optimize the AMISE and eventually arrive at their BCV objective function (their equation (3.17)). In our notation,

$$S_{\mathrm{b}} := \frac{k_{02}}{nh} + \frac{k_{21}^2}{4n^2 h} \sum_{j=1}^{n} \sum_{i \neq j} \left(\int_{-\infty}^{\infty} K^{(2)}(u) \, K^{(2)}(u + z_{ij}/h) \, \mathrm{d}u \right),$$

where $k_{02}/(nh)$ is a good estimator of the integrated variance in the MISE, while the second part is the modified estimator of integrated squared bias which achieves the stability of the BCV criterion relative to UCV. Using Lemma 1 and

$$He_4(b) = b^4 - 6\ b^2 + 3$$

which is calculated from the formula for Hermite polynomials in (8), we get

$$S_{\rm b} = \frac{k_{02}}{nh} + \frac{k_{21}^2}{8n^2}\sum_{j=1}^{n}\sum_{i>j}\left(\frac{z_{ij}^4}{4h^4} - \frac{3z_{ij}^2}{h^2} + 3\right)K_{h\sqrt{2}}(z_{ij}), \tag{32}$$

where K is an even function of z_{ij}, hence the range of the inner summation.

As before, using the Student t(ν) kernel (12) with $h\sqrt{2}$ instead of h, as required for (32), we get

$$S_{\rm b} = \frac{k_{02}}{nh} + \frac{c_\nu k_{21}^2}{8\sqrt{2}n^2}\sum_{j=1}^{n}\sum_{i>j}\left(\frac{z_{ij}^4}{4}h^{\nu-4} - 3z_{ij}^2h^{\nu-2} + 3h^\nu\right)(h^2 + z_{ij}^2/(2\nu))^{-(\nu+1)/2} \tag{33}$$

and the exact first-order solution for $\nu > 4$ is

$$\begin{aligned}
&\frac{8\sqrt{2}k_{02}n}{c_\nu k_{21}^2} \qquad\qquad (34)\\
=\ &\widehat{h}_{\rm b}^{\nu-3}\sum_{j=1}^{n}\sum_{i>j}\left(\left(\frac{\nu}{4}-1\right)z_{ij}^4 - 3(\nu-2)z_{ij}^2\widehat{h}_{\rm b}^2 + 3\nu\widehat{h}_{\rm b}^4\right)\left(\widehat{h}_{\rm b}^2 + z_{ij}^2/(2\nu)\right)^{-(\nu+1)/2}\\
&-(\nu+1)\widehat{h}_{\rm b}^{\nu-1}\sum_{j=1}^{n}\sum_{i>j}\left(\frac{z_{ij}^4}{4} - 3z_{ij}^2\widehat{h}_{\rm b}^2 + 3\widehat{h}_{\rm b}^4\right)\left(\widehat{h}_{\rm b}^2 + z_{ij}^2/(2\nu)\right)^{-(\nu+3)/2},
\end{aligned}$$

where $\widehat{h}_{\rm b}$ is the BCV solution. The same arguments in the previous subsection about $\widehat{h} = O_p(n^{-1/5})$ indicate that this is essentially an equation of the form $\beta_1 = \beta_2\widehat{h}^{\nu-3} + \beta_3\widehat{h}^{\nu-1}$, which leads to

$$\boxed{\widehat{h}_{\rm a} := \left(\frac{\beta_1}{\beta_2 + \beta_3\widehat{h}_{\rm p}^2}\right)^{1/(\nu-3)}}. \tag{35}$$

We can make the same remark as before concerning the positivity of $\beta_2 + \beta_3\widehat{h}_{\rm p}^2$, but this time we have a supplementary restriction on the value of ν which should be

greater than 4. In addition, like (31) was a simplification of (30), here we have the simplifying asymptotic approximation

$$\widehat{h}_{\mathrm{aa}} := \sqrt{-\beta_2/\beta_3}. \tag{36}$$

We use it instead of (35) whenever $\beta_2 + \beta_3 \widehat{h}_{\mathrm{p}}^2 < 0$.

5 Application 3: SCV

Jones, Marron and Park (1991) estimate the integrated squared bias $\int (K_h * f - f)^2$ (or equivalently $\int (D_h * f)^2$) by smoothing this particular appearance of f, effectively a plug-in that uses a second kernel L and bandwidth g. They also combine this with the option of using the idea of Jones and Sheather (1991), in which case they set an indicator function $\delta = 1$ below (and $\delta = 0$ otherwise). The result is the SCV objective function

$$S_{\mathrm{s}} := \frac{k_{02}}{nh} + \frac{\delta}{n} \left(D_h * D_h * L_g * L_g\right)(0) + \frac{1}{n^2} \sum_{j=1}^{n} \sum_{i \neq j} \left(D_h * D_h * L_g * L_g\right)(z_{ij}),$$

where 0 and z_{ij} are the arguments of the respective convolutions. They show that the asymptotically-optimal p in $g \sim C n^p / h^2$ is given by

$$\widehat{p} = \begin{cases} -23/45 & (\delta = 1) \\ -44/85 & (\delta = 0), \end{cases} \tag{37}$$

but the constant C depends on the unknown f again. They experiment with a couple of plug-in methods to estimate C, but they do not work well and they will not be necessary in the case of our method where we optimize with respect to both h and g.

The case of $\delta = 1$ achieves the best $1/\sqrt{n}$ rate for the relative distance between the values of h minimizing MISE and S_{s}, while it is the slightly worse rate of $n^{-8/17}$ that is obtained if $\delta = 0$. Note that $\widehat{g}_{\mathrm{s}}$ dominates $\widehat{h}_{\mathrm{s}}$, where these are the optimizers of S_{s}; e.g., if we take $\widehat{p}$ to be $-\frac{1}{2}$ henceforth, then $\widehat{g}_{\mathrm{s}} = O_p\left(n^{-1/10}\right)$ dominates $\widehat{h}_{\mathrm{s}} = O_p\left(n^{-1/5}\right)$.

Nevertheless, the argument used for $\widehat{h}$ in connection with the Student kernel in Section 2 applies to $\widehat{g}_{\rm s}$ as well.

Although the $1/\sqrt{n}$ rate is achieved by SCV, the best possible multiplicative constant established in Fan and Marron (1992) is not quite reached by the limiting variance of the normalized $\widehat{h}_{\rm s}$. Kim, Park and Marron (1994) show how to modify the method to achieve this lower bound, but their results show that samples as big as $n = 1,000$ are not big enough to reach these asymptotics and they say (p.120) that their method is "mostly of theoretical interest". We therefore do not include their extension.

Using Lemma 1 and the symmetry of the Student t(ν) kernels (we use the same ν for K and L), we can work out the criterion explicitly as

$$\begin{aligned} S_{\rm s} &= \frac{k_{02}}{nh} + \frac{\delta c_\nu}{n\sqrt{2}}\left(\frac{1}{\sqrt{h^2+g^2}} - \frac{2^{3/2}}{\sqrt{h^2+2g^2}} + \frac{1}{g}\right) \\ &\quad + \frac{2c_\nu}{n^2}\left[\left(2h^2+2g^2\right)^{\nu/2} y_n(\nu, h\sqrt{2}, g) - 2\left(h^2+2g^2\right)^{\nu/2} y_n(\nu,h,g) + 2^{\nu/2} g^\nu y_n(\nu,0,g)\right], \end{aligned}$$

where

$$y_n(q,h,g) := \sum_{j=1}^{n}\sum_{i>j}\left(h^2+2g^2+z_{ij}^2/\nu\right)^{-(q+1)/2}. \tag{38}$$

Since $\partial y_n(q,h\sqrt{2},g)/\partial h = (h/g)\,\partial y_n(q,h\sqrt{2},g)/\partial g$ and

$$\frac{\partial y_n(q,h,g)}{\partial g} = \frac{2g}{h}\frac{\partial y_n(q,h,g)}{\partial h} = -2\,(q+1)\,g y_n(q+2,h,g),$$

defining

$$\begin{aligned} y_n^\dagger(q,h,g) &:= \left(h^2+2g^2\right)^{(q-2)/2} y_n(q,h,g) \\ &= \left(h^2+2g^2\right)^{(q-2)/2} \sum_{j=1}^{n}\sum_{i>j}\left(h^2+2g^2+z_{ij}^2/\nu\right)^{-(q+1)/2} \end{aligned} \tag{39}$$

allows us to write the exact first-order conditions for g and h, respectively, as

$$\begin{aligned} &\frac{\delta n}{2^{5/2}}\left(\frac{1}{(\widehat{h}_{\rm s}^2+\widehat{g}_{\rm s}^2)^{3/2}} - \frac{2^{5/2}}{(\widehat{h}_{\rm s}^2+2\widehat{g}_{\rm s}^2)^{3/2}} + \frac{1}{\widehat{g}_{\rm s}^3}\right) \\ = \;& \nu\left[y_n^\dagger(\nu,\widehat{h}_{\rm s}\sqrt{2},\widehat{g}_{\rm s}) - 2y_n^\dagger(\nu,\widehat{h}_{\rm s},\widehat{g}_{\rm s}) + y_n^\dagger(\nu,0,\widehat{g}_{\rm s})\right] \\ & -(\nu+1)\left[y_n^\dagger(\nu+2,\widehat{h}_{\rm s}\sqrt{2},\widehat{g}_{\rm s}) - 2y_n^\dagger(\nu+2,\widehat{h}_{\rm s},\widehat{g}_{\rm s}) + y_n^\dagger(\nu+2,0,\widehat{g}_{\rm s})\right] \end{aligned} \tag{40}$$

and

$$\begin{aligned} & \frac{k_{02}n}{4c_\nu \widehat{h}_{\rm s}^3} + \frac{\delta n}{2^{5/2}} \left(\frac{1}{(\widehat{h}_{\rm s}^2 + \widehat{g}_{\rm s}^2)^{3/2}} - \frac{2^{3/2}}{(\widehat{h}_{\rm s}^2 + 2\widehat{g}_{\rm s}^2)^{3/2}} \right) \qquad (41) \\ = \; & \nu \left[y_n^\dagger(\nu, \widehat{h}_{\rm s}\sqrt{2}, \widehat{g}_{\rm s}) - y_n^\dagger(\nu, \widehat{h}_{\rm s}, \widehat{g}_{\rm s}) \right] - (\nu+1) \left[y_n^\dagger(\nu+2, \widehat{h}_{\rm s}\sqrt{2}, \widehat{g}_{\rm s}) - y_n^\dagger(\nu+2, \widehat{h}_{\rm s}, \widehat{g}_{\rm s}) \right], \end{aligned}$$

where we notice that the terms on the RHS of (41) have already been calculated in (40). Also, (41) can be used to simplify (40) by subtraction as

$$\begin{aligned} & \frac{k_{02}n}{4c_\nu \widehat{h}_{\rm s}^3} + \frac{\delta n}{2^{5/2}} \left(\frac{2^{3/2}}{(\widehat{h}_{\rm s}^2 + 2\widehat{g}_{\rm s}^2)^{3/2}} - \frac{1}{\widehat{g}_{\rm s}^3} \right) \qquad (42) \\ = \; & \nu \left[y_n^\dagger(\nu, \widehat{h}_{\rm s}, \widehat{g}_{\rm s}) - y_n^\dagger(\nu, 0, \widehat{g}_{\rm s}) \right] - (\nu+1) \left[y_n^\dagger(\nu+2, \widehat{h}_{\rm s}, \widehat{g}_{\rm s}) - y_n^\dagger(\nu+2, 0, \widehat{g}_{\rm s}) \right]. \end{aligned}$$

We shall consider solutions of (41) and (42).

As before, the asymptotic invariance of the $y_n(\cdot,\cdot,\cdot)$ function allows us to replace its arguments $\widehat{h}_{\rm s}, \widehat{g}_{\rm s}$ by $\widehat{h}_{\rm p}, \widehat{g}_{\rm p}$, where $\widehat{h}_{\rm p}$ is defined in either of (23) or (26), and $\widehat{g}_{\rm p}$ is defined as

$$\widehat{g}_{\rm p} := \frac{\widehat{h}_{\rm p}}{n^{-1/5}} n^{-1/10} = \widehat{h}_{\rm p} n^{1/10}, \qquad (43)$$

hence replacing $y_n^\dagger(q, a\widehat{h}_{\rm s}, \widehat{g}_{\rm s})$ in (41) and (42) by $(a^2\widehat{h}_{\rm s}^2 + 2\widehat{g}_{\rm s}^2)^{(q-2)/2} y_n(q, a\widehat{h}_{\rm p}, \widehat{g}_{\rm p})$ for all q and a, leading to polynomial-type first-order conditions, as we shall see by the end of this paragraph. Furthermore, an asymptotic approximation for $\widehat{g}_{\rm s}$ can be obtained from (42) by dropping the LHS terms as in the previous two sections, and we get

$$\widehat{g}_{\rm aa} := \sqrt{\frac{y_n(\nu, \widehat{h}_{\rm p}, \widehat{g}_{\rm p}) - y_n(\nu, 0, \widehat{g}_{\rm p})}{2\,(1+1/\nu) \left[y_n(\nu+2, \widehat{h}_{\rm p}, \widehat{g}_{\rm p}) - y_n(\nu+2, 0, \widehat{g}_{\rm p}) \right]}}, \qquad (44)$$

where we have used twice on the RHS $(2 + \widehat{h}_{\rm s}^2/\widehat{g}_{\rm s}^2)^{q/2} = (2 + O_p\left(n^{-1/5}\right))^{q/2} \sim 2^{q/2}$, a large-$n$ asymptotic expansion that is more accurate for small q (i.e. small ν). The corresponding asymptotic approximation for $\widehat{h}_{\rm s}$ is obtained from (41) and by using the binomial expansion

$$\left(2\widehat{h}_{\rm s}^2 + 2\widehat{g}_{\rm s}^2\right)^q = \left(\widehat{h}_{\rm s}^2 + 2\widehat{g}_{\rm s}^2\right)^q \left(1 + \frac{\widehat{h}_{\rm s}^2}{\widehat{h}_{\rm s}^2 + 2\widehat{g}_{\rm s}^2}\right)^q \sim \left(\widehat{h}_{\rm s}^2 + 2\widehat{g}_{\rm s}^2\right)^q, \qquad (45)$$

yielding

$$\widehat{h}_{\rm aa} := \sqrt{\frac{y_n(\nu, \widehat{h}_{\rm p}\sqrt{2}, \widehat{g}_{\rm p}) - y_n(\nu, \widehat{h}_{\rm p}, \widehat{g}_{\rm p})}{(1+1/\nu)\left[y_n(\nu+2, \widehat{h}_{\rm p}\sqrt{2}, \widehat{g}_{\rm p}) - y_n(\nu+2, \widehat{h}_{\rm p}, \widehat{g}_{\rm p})\right]} - 2\widehat{g}_{\rm aa}^2}. \tag{46}$$

Like in the previous two sections, an asymptotic solution that keeps the LHS of (41) can be obtained by using (43) to write $\widehat{g}_{\rm p}^2/\widehat{h}_{\rm p}^2 = n^{1/5}$ and

$$\begin{aligned} &\frac{k_{02}n}{4c_\nu} + \frac{\delta n}{2^{5/2}}\left(\frac{1}{(1+n^{1/5})^{3/2}} - \frac{2^{3/2}}{(1+2n^{1/5})^{3/2}}\right) \\ = \;& \widehat{h}^{\nu+1}\nu\left[(2+2n^{1/5})^{(\nu-2)/2} y_n(\nu, \widehat{h}_{\rm p}\sqrt{2}, \widehat{g}_{\rm p}) - (1+2n^{1/5})^{(\nu-2)/2} y_n(\nu, \widehat{h}_{\rm p}, \widehat{g}_{\rm p})\right] \\ & -\widehat{h}^{\nu+3}(\nu+1)\left[(2+2n^{1/5})^{\nu/2} y_n(\nu+2, \widehat{h}_{\rm p}\sqrt{2}, \widehat{g}_{\rm p}) - (1+2n^{1/5})^{\nu/2} y_n(\nu+2, \widehat{h}_{\rm p}, \widehat{g}_{\rm p})\right], \end{aligned} \tag{47}$$

which is a polynomial of the form $\gamma_1 = \gamma_2\widehat{h}^{\nu+1} + \gamma_3\widehat{h}^{\nu+3}$ yielding as before

$$\boxed{\widehat{h}_{\rm a} := \left(\frac{\gamma_1}{\gamma_2 + \gamma_3\widehat{h}_{\rm p}^2}\right)^{1/(\nu+1)}} \tag{48}$$

if we use the plug-in $\widehat{h}_{\rm p}$ on the RHS. However, unlike in the previous two sections, it is not the case that $\widehat{h}_{\rm aa}$ of (46) equals $\sqrt{-\gamma_2/\gamma_3}$, because of the presence of terms like $(a^2 + 2n^{1/5})^q$ that are due to g.

Unlike in the previous two sections where we did not have g, we now have the following additional result. Having an asymptotic solution $\widehat{g}_{\rm aa}$ and $\widehat{h}_{\rm aa}$ allows us to estimate the constant C (that depends on the unknown density) in $\widehat{g} \sim C/(\widehat{h}^2\sqrt{n})$ as

$$\widehat{C} := \widehat{g}_{\rm aa}\widehat{h}_{\rm aa}^2\sqrt{n}, \tag{49}$$

which obviates the need for a plug-in rule for C as in Jones, Marron and Park (1991) who find that such rules do not work well for C. Furthermore, we can now replace $\widehat{g}_{\rm s}^2$ by $\widehat{C}^2/(\widehat{h}_{\rm s}^4 n)$ in the first-order condition (41) to solve for only one unknown, $\widehat{h}_{\rm s}$, from

$$\begin{aligned} &\frac{k_{02}n\widehat{h}_{\rm s}^{2\nu-3}}{4c_\nu} + \frac{\delta n\widehat{h}_{\rm s}^{2\nu+6}}{2}\left(\frac{1}{(2\widehat{h}_{\rm s}^6 + b_n)^{3/2}} - \frac{1}{(\widehat{h}_{\rm s}^6 + b_n)^{3/2}}\right) \\ = \;& \widehat{h}_{\rm s}^4\nu\left[(2\widehat{h}_{\rm s}^6 + b_n)^{(\nu-2)/2} y_n(\nu, \widehat{h}_{\rm p}\sqrt{2}, \widehat{g}_{\rm p}) - (\widehat{h}_{\rm s}^6 + b_n)^{(\nu-2)/2} y_n(\nu, \widehat{h}_{\rm p}, \widehat{g}_{\rm p})\right] \\ & -(\nu+1)\left[(2\widehat{h}_{\rm s}^6 + b_n)^{\nu/2} y_n(\nu+2, \widehat{h}_{\rm p}\sqrt{2}, \widehat{g}_{\rm p}) - (\widehat{h}_{\rm s}^6 + b_n)^{\nu/2} y_n(\nu+2, \widehat{h}_{\rm p}, \widehat{g}_{\rm p})\right], \end{aligned} \tag{50}$$

where $b_n := 2\widehat{C}^2/n$ and $y_n(q, a\widehat{h}_{\rm p}, \widehat{g}_{\rm p})$ is calculated only once (for $a = 1, \sqrt{2}$ and $q = \nu, \nu + 2$) for *all* iterations over $\widehat{h}_{\rm s}$.

6 Simulations

Among the different possibilities suggested in the literature,[1] we selected five generating densities: Gaussian, Student with three degrees of freedom, two mixtures of normals and the lognormal. We chose t(3) to avoid the Cauchy (i.e. t(1)) and t(2) because their variance does not exist, thus precluding the use of plug-in rules as a starting point. We reported in Table 1 the expression of these densities together with

Table 1: Generating processes and a measure of its intrinsic complexity.

Density	Expression	$B(f)$
Gaussian	N(0,1)	1.30
Bimodal Mixture	0.5N(-1,4/9)+0.5N(1,4/9)	1.87
Student	t(3)	2.58
Skewed Bimodal Mixture	0.75N(0,1)+0.25N(3/2,1/9)	3.39
Lognormal	exp(N(0,1))	7.17

the measure of complexity proposed in Fan and Marron (1992).[2] We have chosen two

[1] For example, see Scott and Terrell (1987), Park and Marron (1990), or Marron and Wand (1992). Scott and Terrell (1987) use four different densities as a benchmark: Gaussian, Cauchy, lognormal, and the mixture of Gaussians 0.75 N(0,1) + 0.25 N(3/2,1/9). Their sample sizes are 400 or larger. Park and Marron (1990) make use of the Gaussian and a variety of mixtures of Gaussians with two sample sizes, $n = 100$ and $n = 400$. Marron and Wand (1992) use fourteen different mixture of normals to investigate the degree of approximation committed when using the AMISE instead of the MISE.

[2] Note the alternative measure of complexity proposed in Wand and Devroye (1993), which is based on an L_1 measure. Both the L_2 measure of Fan and Marron (1992) and the L_1 measure of Wand and Devroye (1993) are scale and location independent. The former is based on a nonparametric counterpart of the famous Cramér-Rao lower bound.

sample sizes $n = 150$ and $n = 450$. For each case, we compute $\widehat{h}_{\text{ise}}$ as the minimizer of the true ISE in (3), given that the true density is known in the simulations.[3] We then compare the efficiency of each method i by reporting the mean of the Monte Carlo ratio $\text{ISE}(\widehat{h}_i)/\text{ISE}(\widehat{h}_{\text{ise}})$. The ISEs are computed on a fixed grid (not sample dependent) of 67 points using Simpson's rule on the following intervals: [-5,5] for the Gaussian density and the two mixtures of Gaussians, [-8,8] for the Student t(3), and $[\text{e}^{-5}, \text{e}^{2}]$ for the lognormal. These grids were chosen so as to cover most of the probability of the theoretical density.[4] We generate 2500 replications for each experiment. In order to reduce the variance of the Monte Carlo experiments, first we impose the same starting seed for each density, and second we take the smaller sample $n = 150$ as a sub-sample of the larger sample $n = 450$.

The coming subsections tackle the following issues. First, we compare (23) and (26) to see which is preferable as the initial plug-in $\widehat{h}_{\text{p}}$ for our method. The second to fourth subsections then study the performance of our UCV, BCV, and SCV formulae, respectively.

6.1 Choosing our initial plug-in $\widehat{h}_{\text{p}}$

With (23) and (26), we discussed two possible initial values for our proposed formulae. Both come from usual plug-in methods generalized for the use of a Student t(ν) kernel. Table 2 shows that these starting values do a fairly good job in term of efficiency. We first note that the efficiency of the plug-in of Jones and Sheather (1991) based on a Gaussian assumption can always be improved either by using the simple rule of Silverman in the empirically rare case of the Gaussian process (which is true by

[3] The estimate $\widehat{h}_{\text{ise}}$ is searched over an initial grid of 9 values covering the interval $[\widehat{h}_{\text{S}}/10, 2\widehat{h}_{\text{S}}]$, where $\widehat{h}_{\text{S}}$ is given in (23). The initial range is automatically enlarged if the optimum is on the boundary. The initial grid is then iteratively split until the required precision is obtained.

[4] The lower bound of e^{-5} was employed to deal with the usual boundary problems of the lognormal. When comparing ISEs for different window sizes, all ISEs are affected by comparable truncation errors.

design for this rule), or by using a Student kernel (instead of a Gaussian) with either 10 degrees of freedom in the case of moderately difficult processes or with 3 degrees of freedom in the case of the more difficult lognormal.

How can we explain the changes in the results as ν varies? The Student kernel entails a loss of relative efficiency measured by the ratio

$$\left(\int K_{\mathrm{t}(\nu)}(t)^2 \, \mathrm{d}t / \int K_{\mathrm{E}}(t)^2 \, \mathrm{d}t \right)^{4/5},$$

where K_{E} is the Epanechnikov and $K_{\mathrm{t}(\nu)}$ the Student $\mathrm{t}(\nu)$ kernel used. This ratio is 1.06 for $\nu = 30$, 1.08 for $\nu = 10$, and goes up to 1.37 for $\nu = 3$. As complexity increases, the loss of efficiency is more than compensated by a better care of the influence of the observations that are outside the immediate neighbourhood of the point where the density is fitted. This is a kind of robustification, the first of two such features that we will note in the simulations. The value $\nu = 10$ seems to be a good compromise between efficiency and robustification for most of the moderately complicated situations.

We note that the relative efficiency of $\widehat{h}_{\mathrm{JS}}$ will be difficult to beat because, on average, $\widehat{h}_{\mathrm{ise}}$ is only 20% more efficient than $\widehat{h}_{\mathrm{JS}}$. This is simply a question about the real efficiency of any cross validation method, compared to a plug-in method. The answer depends on the complexity of the generating process. We must finally note that $\widehat{h}_{\mathrm{JS}}$ is relatively costly to compute, while the cost of $\widehat{h}_{\mathrm{S}}$ is comparatively negligible; see Table 4 below.

6.2 UCV

We have discussed how to choose the value of ν as an inverse function of the complexity of the density to estimate in the case of a plug-in rule only. Here, we will reconsider it in the context of UCV. We will also have to answer four questions. How well does our asymptotic $\widehat{h}_{\mathrm{a}}$ of (30) do? Is it better to solve the more elaborate exact first-order conditions (28) for $\widehat{h}_{\mathrm{u}}$, using an iterative method? Or should we maximize the

Table 2: Plug-in with a Student kernel

Density	Bandwidth	Kernel's ν	$n = 150$	$n = 450$
Gaussian	$\widehat{h}_{\mathrm{S}}$	3	2.68	2.17
		10	1.47	1.27
		30	**1.38**	**1.21**
	$\widehat{h}_{\mathrm{JS}}$	3	3.21	2.39
		10	1.56	1.30
		30	1.48	1.26
Bimodal Mixture	$\widehat{h}_{\mathrm{S}}$	3	1.47	1.32
		10	1.14	1.19
		30	1.29	1.38
	$\widehat{h}_{\mathrm{JS}}$	3	1.79	1.49
		10	**1.12**	**1.11**
		30	1.16	1.14
Student t(3)	$\widehat{h}_{\mathrm{S}}$	3	1.64	1.38
		10	1.68	1.68
		30	2.03	2.08
	$\widehat{h}_{\mathrm{JS}}$	3	2.00	1.58
		10	**1.30**	**1.21**
		30	1.32	1.23
Skewed Mixture	$\widehat{h}_{\mathrm{S}}$	3	1.26	1.16
		10	1.24	1.39
		30	1.42	1.66
	$\widehat{h}_{\mathrm{JS}}$	3	1.53	1.32
		10	**1.13**	**1.15**
		30	1.24	1.25
Lognormal	$\widehat{h}_{\mathrm{S}}$	3	2.05	2.68
		10	5.35	7.99
		30	6.20	9.40
	$\widehat{h}_{\mathrm{JS}}$	3	**1.18**	**1.20**
		10	2.44	2.73
		30	2.88	3.27

Each line reports the expectation of the ratio $\mathrm{ISE}(\widehat{h}_i)/\mathrm{ISE}(\widehat{h}_{\mathrm{ise}})$, where $\widehat{h}_{\mathrm{ise}}$ is the optimal bandwidth for a Student kernel with ν degrees of freedom.

objective function (18) for $\widehat{h}_{\mathrm{u}}$, using a grid search in case this function has several local minima in small samples? Is there any benefit from using the more expensive $\widehat{h}_{\mathrm{JS}}$ as our initial plug-in $\widehat{h}_{\mathrm{p}}$?

For the Gaussian process, Table 3 shows that $\widehat{h}_{\mathrm{a}}$ does not improve efficiency over the simple starting value $\widehat{h}_{\mathrm{S}}$, which we expected since $\widehat{h}_{\mathrm{S}}$ is designed for this case. In fact, taking $\nu = 30$ gives essentially a Gaussian kernel, and the efficiency of $\widehat{h}_{\mathrm{a}}$ is almost indistinguishable from that of $\widehat{h}_{\mathrm{S}}$ in Table 2. Table 3 shows that iterating to get the exact $\widehat{h}_{\mathrm{u}}$ is of no use and even leads to a worse situation here. For the bimodal mixture, $\widehat{h}_{\mathrm{a}}$ improves efficiency over the simple starting value $\widehat{h}_{\mathrm{S}}$ and reaches the efficiency obtained with $\widehat{h}_{\mathrm{JS}}$ for $\nu = 10$. As $\widehat{h}_{\mathrm{a}}$ is much cheaper to compute than $\widehat{h}_{\mathrm{JS}}$, it thus constitutes an interesting alternative. Iterating is again of no use here. A similar situation appears for the Student and the skewed mixture. If we decide to iterate, it is better in all cases to choose $\nu = 3$. Iterating for any given sample realization can leads us away from the optimum, and it is better to use the more "robust" low ν in this case, as discussed in the previous subsection. This brings us to the second "robustification" comment. Starting with Table 3, we will notice that $\widehat{h}_{\mathrm{a}}$ often performs better than the exact solution of first-order equations. This gives us an idea of how robust $\widehat{h}_{\mathrm{a}}$ is to undesirable sample variations that CV methods are known for, especially UCV. Note that all the methods seen so far in the table produce quickly convergent results as the sample size grows.

The lognormal case is, in a sense, rather special. It is well known that in the case of strong asymmetry, it is not ideal to find a single window size to estimate a density on the whole of its support, when using a symmetric kernel. The preferred solutions would be to use asymmetric kernels as in Abadir and Lawford (2004) or to transform the data as underlined by Wand, Marron and Ruppert (1991). However, $\widehat{h}_{\mathrm{a}}$ with $\nu = 3$ still does very well. Additionally, here we examine the more expensive $\widehat{h}_{\mathrm{JS}}$ as a plug-in for our method, and we find that it is does better than using $\widehat{h}_{\mathrm{S}}$ as our plug-in for the RHS of $\widehat{h}_{\mathrm{a}}$. Actually, $\widehat{h}_{\mathrm{a}}$ with $\widehat{h}_{\mathrm{JS}}$ as a plug-in does better than the

Table 3: Unbiased cross-validation with a Student kernel

Density	Bandwidth	Kernel's ν	$n = 150$	$n = 450$
Gaussian	$\widehat{h}_{\mathrm{a}}$	3	2.27	1.85
		10	1.48	1.28
		30	**1.39**	**1.22**
	$\widehat{h}_{\mathrm{u}}$	3	1.62	1.43
		10	1.95	1.59
		30	2.10	1.66
Bimodal Mixture	$\widehat{h}_{\mathrm{a}}$	3	1.36	1.26
		10	**1.14**	**1.13**
		30	1.25	1.30
	$\widehat{h}_{\mathrm{u}}$	3	1.32	1.23
		10	1.44	1.30
		30	1.48	1.33
Student	$\widehat{h}_{\mathrm{a}}$	3	1.61	1.46
		10	**1.41**	**1.31**
		30	1.81	1.78
	$\widehat{h}_{\mathrm{u}}$	3	1.59	1.43
		10	1.76	1.51
		30	1.83	1.54
Skewed Mixture	$\widehat{h}_{\mathrm{a}}$	3	1.24	1.20
		10	**1.19**	**1.21**
		30	1.37	1.52
	$\widehat{h}_{\mathrm{u}}$	3	1.29	1.20
		10	1.38	1.26
		30	1.41	1.28
Lognormal	$\widehat{h}_{\mathrm{a}}$	3	1.17	1.13
		10	3.49	4.42
		30	5.34	7.67
	$\widehat{h}_{\mathrm{a}}$ (with $\widehat{h}_{\mathrm{p}} = \widehat{h}_{\mathrm{JS}}$)	3	**1.13**	**1.09**
		10	1.79	1.79
		30	2.54	2.76
	$\widehat{h}_{\mathrm{u}}$	3	1.31	1.22
		10	1.36	1.24
		30	1.38	1.24

Bold numbers indicate the best method for each density. Starting values are $\widehat{h}_{\mathrm{S}}$ if not stated otherwise.

latter on its own; compare to Table 3. The asymptotic $\widehat{h}_{\mathrm{a}}$ improves a lot the efficiency of $\widehat{h}_{\mathrm{S}}$ and also that of $\widehat{h}_{\mathrm{JS}}$. The efficiency results for this simple $\widehat{h}_{\mathrm{a}}$ are strikingly good (very close to minimum-ISE) for a generating density that is as troublesome as the lognormal. Again, by a careful choice of ν, it is not useful to iterate.

UCV methods have a tendency to produce smaller window sizes than plug-in methods. It is thus useful to compare the graphs of the empirical distributions of the various estimates of h in order to have a more precise idea about their location and dispersion. The bimodal mixture of Gaussians and the lognormal represent opposite levels of difficulty of estimation, and so offer graphs with interesting interpretations.

Figure 1: Window size dispersion for unbiased cross validation:

Bimodal Gaussian mixture density with a Student kernel ($\nu = 10$).

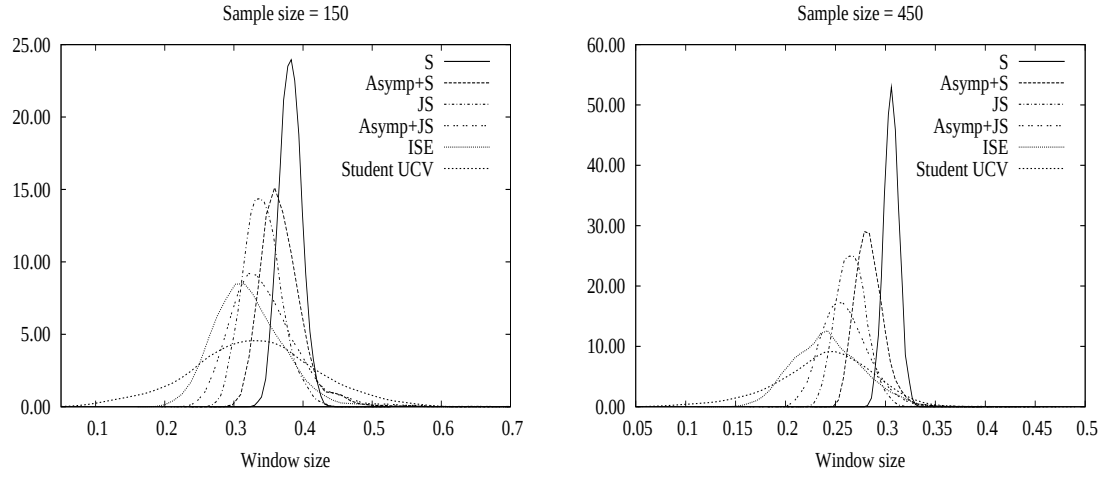


Figure 1 presents the empirical density of the various estimates of h for the bimodal mixture of Gaussians using a Student kernel with $\nu = 10$. The distribution of $\widehat{h}_{\mathrm{u}}$ is rather dispersed for a small sample size, but seems to converge to the distribution of $\widehat{h}_{\mathrm{ise}}$ as the sample size increases. The distribution of $\widehat{h}_{\mathrm{a}}$ is less dispersed than that of $\widehat{h}_{\mathrm{ise}}$, but is always contained in a reasonable probability interval of the latter, a robustness feature that we mentioned earlier in this subsection. In contrast, the distribution of $\widehat{h}_{\mathrm{S}}$ is more and more concentrated and goes to the borderline of a

reasonable probability interval of the distribution of $\widehat{h}_{\text{ise}}$ as the sample size grows.

Figure 2: Window size dispersion for unbiased cross validation:

Lognormal density with a Student kernel ($\nu = 3$).

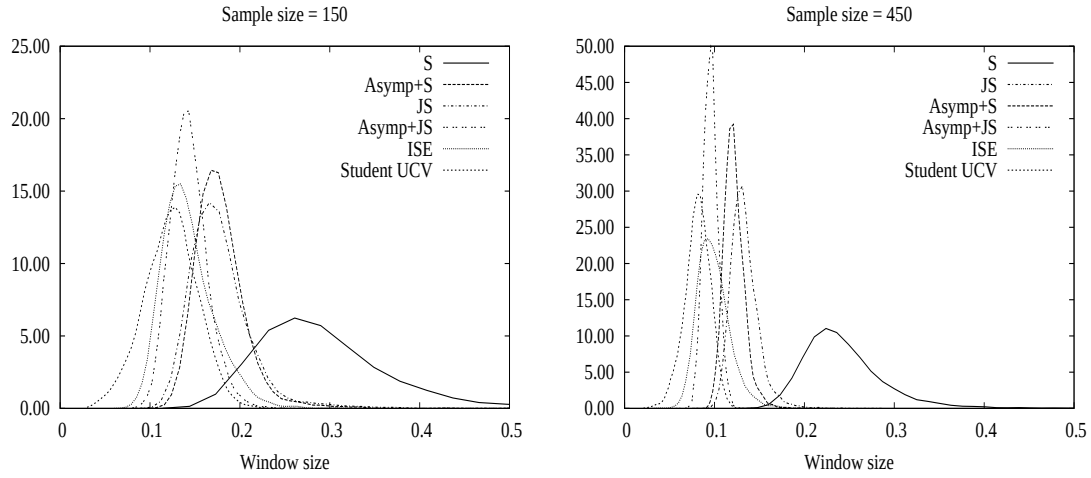


Figure 2 presents a rather different picture for estimating the lognormal density using a Student kernel with $\nu = 3$. The distribution of $\widehat{h}_{\text{S}}$ is very dispersed and lies far away from the distribution of the other bandwidth estimates. This seems to be the case because $\widehat{\sigma}^2$ is not a good measure of dispersion when contrasting the two lognormal tails.[5] However, when $\widehat{h}_{\text{S}}$ is used as a plug-in for $\widehat{h}_{\text{a}}$, the distribution of $\widehat{h}_{\text{a}}$ is nearly identical to that of $\widehat{h}_{\text{JS}}$. When the sample size grows, the distribution of $\widehat{h}_{\text{a}}$ gets closer to that of $\widehat{h}_{\text{ise}}$. The exact solution $\widehat{h}_{\text{u}}$, obtained by iterating the first-order condition of UCV, produces window estimates that are slightly smaller than the ISE-optimal ones. Everything behaves smoothly in this graph because $\nu = 3$. For $\nu = 30$, the kernel is basically Gaussian and we have three clearly identified groups in the windows we considered: $\widehat{h}_{\text{S}}$ and $\widehat{h}_{\text{a}}$ (with $\widehat{h}_{\text{S}}$ as a starting value), then $\widehat{h}_{\text{JS}}$ and the corresponding $\widehat{h}_{\text{a}}$, finally $\widehat{h}_{\text{ise}}$ and $\widehat{h}_{\text{u}}$. With $\nu = 30$, all the non-iterated methods

[5]Silverman (1986) suggests using a more robust measure of dispersion given by $\min(\widehat{\sigma}, (q_{0.75} - q_{0.25})/1.34)$, where q_α is the α quantile. This could improve greatly the performance for this particular case, but gave mixed results for the other generating processes.

are disqualified. We clearly see here the beneficial impact of using a Student kernel with ν smaller than 30.

Table 4: Relative execution time

Method	$n = 150$	$n = 450$
$\widehat{h}_{\mathrm{a}}$ (with $\widehat{h}_{\mathrm{p}} = \widehat{h}_{\mathrm{S}}$)	1	1
$\widehat{h}_{\mathrm{JS}}$	2	2
$\widehat{h}_{\mathrm{u}}$	16	16
integral-free UCV	24	22
ISE-optimal	21	6

Computer time is normalized in terms of the execution time of $\widehat{h}_{\mathrm{a}}$.

Let us finally turn to computational efficiency. Computations were done on a Pentium 4 running at 3 GHz. The $\widehat{h}_{\mathrm{a}}$ based on (30) involves no iteration, but a loop of size $n \times (n-1)/2$, while $\widehat{h}_{\mathrm{JS}}$ involves a loop of size n^2. The UCV $\widehat{h}_{\mathrm{u}}$ based on the first-order condition (28) involves iterations and a loop of size $n \times (n-1)/2$. The optimization of our integral-free UCV objective function (18) and of the ISE-optimal $\widehat{h}_{\mathrm{ise}}$ are based on an adaptative grid search. The former involves the basic calculation of a loop of size $n \times (n-1)/2$, while the latter has two loops of size $n \times m$ with m denoting the size of the basic grid over which the function is evaluated.

Table 4 displays execution times relative to that of $\widehat{h}_{\mathrm{a}}$. This ratio seems to be independent of the sample size, except for the ISE-optimal calculation. The asymptotic $\widehat{h}_{\mathrm{a}}$ of (30) can be up to 24 times quicker than the integral-free UCV, while $\widehat{h}_{\mathrm{u}}$ can be 1.5 times quicker than a direct optimzation of the integral-free UCV using an adaptative grid search. We can conclude that $\widehat{h}_{\mathrm{a}}$ is numerically very efficient compared to the integral-free UCV, while $\widehat{h}_{\mathrm{u}}$ brings some additional efficiency compared to the latter. The gains in absolute (not relative) comuptational times can be considerable when dealing with large datasets such as the ones that arise in finance, where typical

methods are too slow (they take hours) to yield a useful answer.

6.3 BCV

BCV was proposed as an alternative to UCV, in order to provide efficiency gains and to deal with the sample variability of UCV solutions (a gain that has already been achieved by our asymptotic $\widehat{h}_{\mathrm{a}}$ of (30) for UCV). The aim of this second Monte Carlo experiment is to investigate the reality of these potential gains for our explicit exact solutions of BCV bandwidths and their asymptotic versions, as well as comparing the results with our new UCV results.

Before presenting results in Table 5, we must underline the computational difficulties for getting the BCV estimate. It was not possible to solve the first-order (34), as the program ends too many times with a non-convergence. Instead, we were looking for the minimum of the integral-free BCV objective function S_{b} in (33) by a grid search. But in most cases, (33) reaches a minimum as $h \to \infty$. We had to look for a local minimum in the range $[\widehat{h}_{\mathrm{S}}/10, 1.25\widehat{h}_{\mathrm{S}}]$. Whenever there was no local minimum, we selected the upper value of the interval. Scott and Terrell (1987) mention related difficulties in their optimizations when samples are small; see their Sections 5 and 6. The samples they consider for BCV are typically very large for what is commonly available in economics.

Let us now compare the results of Table 3 and of Table 5, case by case. For the Gaussian density, BCV improves clearly over the UCV, as found by Scott and Terrell (1987). For all the other cases, there is no major improvement except in 2 cases of mixtures and small n where the BCV's $\widehat{h}_{\mathrm{a}}$ of (35) does well. In general, most of the time we had to use the asymptotic approximation $\widehat{h}_{\mathrm{aa}}$ of (36) instead of $\widehat{h}_{\mathrm{a}}$ of (35). In the cases we analyze, the BCV approach is infrequently beneficial. Here again, we find that $\nu = 10$ does better than $\nu = 30$ (approximately Gaussian kernel) most of the time, hence supporting the idea of considering a Student kernel. There is no case $\nu = 3$ in the table because it is required that $\nu > 4$ for the BCV derivations; see

Table 5: Biased cross-validation with a Student kernel

Density	Bandwidth	Kernel's ν	$n = 150$	$n = 450$
Gaussian	$\widehat{h}_{\mathrm{a}}$	10	1.69	1.41
		30	*1.33*	*1.20*
	$\widehat{h}_{\mathrm{b}}$	10	**1.23**	**1.15**
		30	*1.29*	*1.22*
Bimodal Mixture	$\widehat{h}_{\mathrm{a}}$	10	**1.07**	**1.12**
		30	1.36	1.42
	$\widehat{h}_{\mathrm{b}}$	10	1.53	1.72
		30	1.85	1.82
Student	$\widehat{h}_{\mathrm{a}}$	10	1.53	1.44
		30	1.94	1.95
	$\widehat{h}_{\mathrm{b}}$	10	2.10	2.67
		30	*1.49*	1.64
Skewed Mixture	$\widehat{h}_{\mathrm{a}}$	10	**1.13**	1.25
		30	1.45	1.72
	$\widehat{h}_{\mathrm{b}}$	10	1.63	2.01
		30	1.93	2.40
Lognormal	$\widehat{h}_{\mathrm{a}}$	10	5.25	7.34
		30	6.09	8.60
	$\widehat{h}_{\mathrm{a}}$ (with $\widehat{h}_{\mathrm{p}} = \widehat{h}_{\mathrm{JS}}$)	10	2.51	2.63
		30	2.84	3.06
	$\widehat{h}_{\mathrm{b}}$	10	4.85	7.75
		30	2.05	1.66

Starting values are $\widehat{h}_{\mathrm{S}}$ if not stated otherwise. Figures in italics indicate an improvement over the corresponding case in Table 3. Figures in bold indicate the best solution for the process.

the condition before (34). We shall henceforth only compare the cases $\nu = 10$ and $\nu = 30$.

6.4 SCV

How do our SCV formulae perform, and how do they compare to UCV and BCV? In Table 6, $\widehat{h}_{\text{aa}}$ represents the asymptotic approximation (46) which requires $\widehat{g}_{\text{aa}}$ of (44), while $\widehat{h}_{\text{a}}$ is the asymptotic (48). For $\widehat{h}_{\text{s}}$, we have two versions, both iterated solutions of first-order conditions: the first arises from (47), while the second is obtained from (50).

For the Gaussian case, SCV does better than UCV, except if we compare $\widehat{h}_{\text{aa}}$ here to $\widehat{h}_{\text{a}}$ there. There are no substantial differences among the various methods. For the other processes, the gain is not evident. In general, the simple asymptotic approximation $\widehat{h}_{\text{aa}}$ provides a good solution with $\nu = 10$. The second $\widehat{h}_{\text{s}}$ is better than the first one, except for the special case of the lognormal where the ranking is reversed. A choice of $\nu = 10$ dominates $\nu = 30$ here too. Overall, SCV tends to do better than UCV in small samples, except in the Student and lognormal cases. This result is in accordance with the simulation results of Jones, Marron and Park (1991) where smoothing the MISE give better convergence results when not far from the Gaussian. But when far from it, the UCV gave better results despite its variability.

7 Conclusion

In this paper, we introduce a general method for solving explicitly the optimization of CV-type problems. We use this approach for the optimization of UCV, BCV, and SCV criteria in density estimation. We obtain an explicit first-order condition for the bandwidth that optimizes each of these criteria. We then obtain an explicit asymptotic formula for the optimal bandwidth in each of the three cases. The asymptotic formula is displayed in (30), (35), and (48). It requires no iteration, is simple and

Table 6: Smoothed cross-validation with a Student kernel

Density	Bandwidth	Kernel's ν	$n = 150$	$n = 450$
Gaussian	$\widehat{h}_{\mathrm{aa}}$	10	1.64	1.42
		30	1.40	1.23
	$\widehat{h}_{\mathrm{a}}$	10	*1.31*	*1.20*
		30	*1.33*	*1.19*
	$\widehat{h}_{\mathrm{s}}$ of (47)	10	*1.62*	*1.32*
		30	*1.38*	*1.20*
	$\widehat{h}_{\mathrm{s}}$ of (50)	10	*1.59*	*1.27*
		30	*1.43*	*1.19*
Bimodal Mixture	$\widehat{h}_{\mathrm{aa}}$	10	*1.10*	*1.13*
		30	1.26	1.33
	$\widehat{h}_{\mathrm{a}}$	10	1.35	1.39
		30	1.38	1.46
	$\widehat{h}_{\mathrm{s}}$ of (47)	10	2.80	2.57
		30	2.42	2.03
	$\widehat{h}_{\mathrm{s}}$ of (50)	10	*1.11*	*1.23*
		30	*1.23*	1.43
Student	$\widehat{h}_{\mathrm{aa}}$	10	1.50	1.43
		30	1.93	1.95
	$\widehat{h}_{\mathrm{a}}$	10	1.74	1.66
		30	2.04	2.04
	$\widehat{h}_{\mathrm{s}}$ of (47)	10	1.94	1.55
		30	1.84	1.78
	$\widehat{h}_{\mathrm{s}}$ of (50)	10	1.86	1.79
		30	1.86	2.14
Skewed Mixture	$\widehat{h}_{\mathrm{aa}}$	10	*1.15*	1.24
		30	1.38	1.58
	$\widehat{h}_{\mathrm{a}}$	10	1.45	1.63
		30	1.50	1.75
	$\widehat{h}_{\mathrm{s}}$ of (47)	10	2.49	2.67
		30	2.20	2.23
	$\widehat{h}_{\mathrm{s}}$ of (50)	10	*1.17*	1.42
		30	*1.35*	1.73
Lognormal	$\widehat{h}_{\mathrm{aa}}$	10	4.63	6.77
		30	5.98	8.97
	$\widehat{h}_{\mathrm{a}}$	10	5.03	6.92
		30	6.05	8.91
	$\widehat{h}_{\mathrm{s}}$ of (47)	10	3.14	2.63
		30	4.18	3.74
	$\widehat{h}_{\mathrm{s}}$ of (50)	10	5.70	8.25
		30	5.85	9.49

Starting values are $\widehat{h}_{\mathrm{S}}$. Figures in italics indicate an improvement over the corresponding case ($\widehat{h}_{\mathrm{aa}}$ and $\widehat{h}_{\mathrm{a}}$ here considered together) in Table 3. Figures in bold indicate the best solution for the process when including UCV and BCV in the comparison.

very fast to calculate, is ISE-efficient, and is very robust. The latter two features (efficiency and robustness) of our explicit asymptotic solution are a compensation for CV's notorious sampling variability which has preoccupied many in this field and has led to many modifications of CV-type criteria in an attempt to stabilize it. Our results apply to non-i.i.d. setups as well, with a minor modification of the index of some sums as, for example, shown by Hart and Vieu (1990) and Hall, Lahiri and Truong (1995).

Appendix

Proof of Lemma 1. By definition,

$$\left(K^{(q)} * K^{(r)}\right)(a) = \int_{-\infty}^{\infty} K^{(q)}(t)\, K^{(r)}(a-t)\,\mathrm{d}t;$$

and we drop the argument a henceforth from the LHS for convenience. Using $K = \phi$,

$$K^{(q)} * K^{(r)} = \int_{-\infty}^{\infty} \phi^{(q)}(t)\,\phi^{(r)}(a-t)\,\mathrm{d}t = \mathrm{D}_{w_1}^{q}\mathrm{D}_{w_2}^{r}\int_{-\infty}^{\infty}\phi(w_1+t)\,\phi(w_2+a-t)\,\mathrm{d}t,$$

where D_w^q is shorthand for the q-th derivative with respect to w, evaluated at $w = 0$. Using the convolution of two Gaussians,

$$\begin{aligned} K^{(q)} * K^{(r)} &= \frac{1}{\sqrt{2}}\mathrm{D}_{w_1}^{q}\mathrm{D}_{w_2}^{r}\phi\left(\frac{w_1+w_2+a}{\sqrt{2}}\right) = \frac{\phi^{(q+r)}\left(a/\sqrt{2}\right)}{\sqrt{2}} \\ &= (-1)^{q+r}\,\frac{\phi\left(a/\sqrt{2}\right) He_{q+r}\left(a/\sqrt{2}\right)}{2^{(q+r+1)/2}} \end{aligned}$$

by the definition of Hermite polynomials.

To work out $D_h * D_h * L_g * L_g$, we start with

$$L_g * L_g = \frac{1}{g^2}\int_{-\infty}^{\infty}\phi\left(\frac{t}{g}\right)\phi\left(\frac{a-t}{g}\right)\mathrm{d}t = \frac{1}{g}\int_{-\infty}^{\infty}\phi(u)\,\phi\left(\frac{a}{g}-u\right)\mathrm{d}u$$

by a change of variable. Applying the result of the previous convolution and using $He_0 \equiv 1$,

$$L_g * L_g = \frac{\phi\left(a/\left(g\sqrt{2}\right)\right)}{g\sqrt{2}} = L_{g\sqrt{2}} = K_{g\sqrt{2}}$$

Next,

$$D_h * D_h = K_h * K_h - 2K_h + K_0 = K_{h\sqrt{2}} - 2K_h + K_0,$$

hence

$$\begin{aligned} D_h * D_h * L_g * L_g &= \left(K_{h\sqrt{2}} - 2K_h + K_0\right) * K_{g\sqrt{2}} \\ &= K_{h\sqrt{2}} * K_{g\sqrt{2}} - 2K_h * K_{g\sqrt{2}} + K_{g\sqrt{2}}. \end{aligned}$$

The remaining convolutions can be worked out by means of

$$K_b * K_c = \frac{1}{bc}\int_{-\infty}^{\infty} \phi\left(\frac{t}{b}\right)\phi\left(\frac{a-t}{c}\right)\mathrm{d}t = \frac{1}{\sqrt{b^2+c^2}}\phi\left(\frac{a}{\sqrt{b^2+c^2}}\right) = K_{\sqrt{b^2+c^2}}$$

to give the required result.

Lemma 2 *Let $\nu > 2$.*

(i) For a Student $\mathrm{t}(\nu)$ *kernel,* $k_{21} := \int_{-\infty}^{\infty} t^2 K(t)\,\mathrm{d}t = \nu/(\nu-2)$ *and*

$$k_{02} := \int_{-\infty}^{\infty} K(t)^2\,\mathrm{d}t = \frac{\sqrt{2}\Gamma\left(\frac{\nu}{2}+\frac{1}{2}\right)\Gamma\left(\frac{\nu}{2}+\frac{1}{4}\right)\Gamma\left(\frac{\nu}{2}+\frac{3}{4}\right)}{\sqrt{\pi}\nu^{\frac{3}{2}}\Gamma\left(\frac{\nu}{2}\right)^3} \sim \frac{\left(1-\frac{3}{16\nu}\right)^2\left(1-\frac{1}{4\nu}\right)}{2\sqrt{\pi}},$$

here $k_{02} \sim a(\nu)$ meaning that the function $a(\nu)$ is made up of the leading terms of the asymptotic expansion of k_{02} for large ν.

(ii) For a scaled Student $\mathrm{t}(\nu)$ *density with variance σ^2,*

$$\begin{aligned} I_2 := \int_{-\infty}^{\infty} f^{(2)}(u)^2\,\mathrm{d}u &= \frac{3\nu(\nu+1)^2(\nu+3)^2 c_\nu^2}{\sigma^5(\nu-2)^{5/2}(2\nu+9)^{1/2}(2\nu+7)(2\nu+5)c_{2\nu+9}} \\ &\sim \frac{3(\nu+1)^2(\nu+3)^2(4\nu-1)^2}{\sigma^5 4\sqrt{\pi\nu}(\nu-2)^{5/2}(2\nu+7)(2\nu+5)(8\nu+17)} \end{aligned}$$

where $c_\nu := \Gamma\left(\frac{\nu+1}{2}\right) / \left(\sqrt{\pi\nu}\Gamma\left(\frac{\nu}{2}\right)\right)$.

Proof. (i) For k_{21}, the result is simply the usual variance of a $\mathrm{t}(\nu)$. For k_{02}, the integrating constant $c_{2\nu+1}$ of the $\mathrm{t}(2\nu+1)$ density implies that

$$\begin{aligned} k_{02} &= \int_{-\infty}^{\infty} \frac{c_\nu^2}{(1+t^2/\nu)^{\nu+1}}\,\mathrm{d}t = \sqrt{\frac{\nu}{2\nu+1}}\frac{\left(\Gamma\left(\frac{\nu+1}{2}\right)/\left(\sqrt{\pi\nu}\Gamma\left(\frac{\nu}{2}\right)\right)\right)^2}{\Gamma(\nu+1)/\left(\sqrt{\pi(2\nu+1)}\Gamma\left(\nu+\frac{1}{2}\right)\right)} \\ &= \frac{\sqrt{2}\Gamma\left(\frac{\nu}{2}+\frac{1}{2}\right)\Gamma\left(\frac{\nu}{2}+\frac{1}{4}\right)\Gamma\left(\frac{\nu}{2}+\frac{3}{4}\right)}{\sqrt{\pi}\nu^{\frac{3}{2}}\Gamma\left(\frac{\nu}{2}\right)^3} \sim \frac{1}{2\sqrt{\pi}}\left(1-\frac{3}{16\nu}\right)^2\left(1-\frac{1}{4\nu}\right), \end{aligned}$$

where the last equality follows from Legendre's duplication formula

$$\Gamma(\eta) = \frac{2^{\eta-1}}{\sqrt{\pi}}\Gamma\left(\frac{\eta}{2}\right)\Gamma\left(\frac{\eta+1}{2}\right),$$

and the asymptotic equivalence from the general approximation for the ratio of two gamma functions

$$\begin{aligned}\frac{\Gamma(a+\nu/2)}{\Gamma(b+\nu/2)} &= \left(\frac{\nu}{2}\right)^{a-b}\left(1+\frac{(a-b)(a+b-1)}{\nu}+O\left(\frac{1}{\nu^2}\right)\right)\\ &\sim \left(\frac{\nu}{2}\right)^{a-b}\left(1+\frac{(a-b)(a+b-1)}{\nu}\right).\end{aligned}$$

(ii) The Student $\mathrm{t}(\nu)$ density with variance σ^2 is

$$f(u) = \frac{c_\nu}{\sigma\sqrt{1-2/\nu}\left(1+u^2/\left(\nu\sigma^2\left(1-2/\nu\right)\right)\right)^{(\nu+1)/2}},$$

hence

$$f^{(2)}(u)^2 = \frac{\left(1+1/\nu\right)^2 c_\nu^2}{\sigma^6\left(1-2/\nu\right)^3}\frac{\left(1-\left(\nu+2\right)u^2/\left(\nu\sigma^2\left(1-2/\nu\right)\right)\right)^2}{\left(1+u^2/\left(\nu\sigma^2\left(1-2/\nu\right)\right)\right)^{\nu+5}}.$$

By the change of variable $t = u\sqrt{2\nu+9}/\sqrt{\nu\sigma^2\left(1-2/\nu\right)}$,

$$I_2 = \int_{-\infty}^{\infty} f^{(2)}(u)^2\,\mathrm{d}u = \frac{\left(1+1/\nu\right)^2 c_\nu^2}{\sigma^5\left(1-2/\nu\right)^{5/2}\left(2+9/\nu\right)^{1/2} c_{2\nu+9}}\int_{-\infty}^{\infty}\frac{c_{2\nu+9}\left(1-\frac{\nu+2}{2\nu+9}t^2\right)^2}{\left(1+t^2/\left(2\nu+9\right)\right)^{\nu+5}}\,\mathrm{d}t.$$

From the Student $\mathrm{t}(2\nu+9)$ density,

$$\begin{aligned}I_2 &= \frac{\left(1+1/\nu\right)^2 c_\nu^2}{\sigma^5\left(1-2/\nu\right)^{5/2}\left(2+9/\nu\right)^{1/2} c_{2\nu+9}}\left(1-2\frac{\nu+2}{2\nu+7}+\left(\frac{\nu+2}{2\nu+9}\right)^2\frac{3\left(2\nu+9\right)^2}{\left(2\nu+7\right)\left(2\nu+5\right)}\right)\\ &= \frac{3\nu\left(\nu+1\right)^2\left(\nu+3\right)^2 c_\nu^2}{\sigma^5\left(\nu-2\right)^{5/2}\left(2\nu+9\right)^{1/2}\left(2\nu+7\right)\left(2\nu+5\right) c_{2\nu+9}}.\end{aligned}$$

Using

$$c_\nu = \frac{\Gamma\left(\frac{\nu+1}{2}\right)}{\sqrt{\pi\nu}\Gamma\left(\frac{\nu}{2}\right)} \sim \frac{1-\frac{1}{4\nu}}{\sqrt{2\pi}} \qquad \text{and} \qquad c_{2\nu+9} = \frac{\Gamma\left(\nu+5\right)}{\sqrt{\pi\left(2\nu+9\right)}\Gamma\left(\nu+\frac{9}{2}\right)} \sim \frac{1+\frac{17}{8\nu}}{\sqrt{\pi\left(2+\frac{9}{\nu}\right)}} \tag{51}$$

gives the required asymptotic result.

Lemma 3 *(i) For a Student* $\mathrm{t}(\nu)$ *kernel,*

$$\begin{aligned} K^{(4)}(t) &= \frac{c_\nu (\nu+1)(\nu+3)\left((\nu+2)(\nu+4)t^4 - 6\nu(\nu+4)t^2 + 3\nu^2\right)}{\nu^4 \left(1+t^2/\nu\right)^{(\nu+9)/2}} \\ &\sim \frac{(4\nu-1)(\nu+1)(\nu+3)\left((\nu+2)(\nu+4)t^4 - 6\nu(\nu+4)t^2 + 3\nu^2\right)}{4\sqrt{2\pi}\nu^5 \left(1+t^2/\nu\right)^{(\nu+9)/2}}. \end{aligned}$$

(ii) For a scaled Student $\mathrm{t}(\nu)$ *density with* $\nu > 2$ *and variance* σ^2*,*

$$\begin{aligned} I_3 := \int_{-\infty}^{\infty} f^{(3)}(u)^2 \,\mathrm{d}u &= \frac{15\nu(\nu+1)^2(\nu+3)^2(\nu+5)^2 c_\nu^2}{\sigma^7(\nu-2)^{7/2}(2\nu+13)^{1/2}(2\nu+7)(2\nu+9)(2\nu+11)c_{2\nu+13}} \\ &\sim \frac{15(\nu+1)^2(\nu+3)^2(\nu+5)^2(4\nu-1)^2}{\sigma^7 4\sqrt{\pi\nu}(\nu-2)^{7/2}(2\nu+7)(2\nu+9)(2\nu+11)(8\nu+25)}, \end{aligned}$$

where $c_\nu := \Gamma\left(\frac{\nu+1}{2}\right) / \left(\sqrt{\pi\nu}\Gamma\left(\frac{\nu}{2}\right)\right)$.

Proof. (i) This follows directly from $K(t) = c_\nu/(1+t^2/\nu)^{(\nu+1)/2}$ and (51).

(ii) From the Student $\mathrm{t}(\nu)$ density with variance σ^2 (see Lemma 2(ii)),

$$f^{(3)}(u)^2 = \frac{9\nu(\nu+1)^2(\nu+3)^2 c_\nu^2}{\sigma^{10}(\nu-2)^5} u^2 \frac{\left(1-(\nu+2)u^2/\left(3\sigma^2(\nu-2)\right)\right)^2}{\left(1+u^2/\left(\sigma^2(\nu-2)\right)\right)^{\nu+7}}.$$

By the change of variable $t = u/\sqrt{\sigma^2(\nu-2)}$,

$$I_3 = \int_{-\infty}^{\infty} f^{(3)}(u)^2 \,\mathrm{d}u = \frac{9\nu(\nu+1)^2(\nu+3)^2 c_\nu^2}{\sigma^7(\nu-2)^{7/2}} \int_{-\infty}^{\infty} t^2 \frac{\left(1-(\nu+2)t^2/3\right)^2}{(1+t^2)^{\nu+7}} \,\mathrm{d}t.$$

From the Student $\mathrm{t}(2\nu+13)$ density,

$$\begin{aligned} I_3 &= \frac{9\nu(\nu+1)^2(\nu+3)^2 c_\nu^2}{\sigma^7(\nu-2)^{7/2} c_{2\nu+13}\sqrt{2\nu+13}} \\ &\quad \times \left(\frac{1}{2\nu+11} - \frac{2(\nu+2)}{(2\nu+9)(2\nu+11)} + \frac{5(\nu+2)^2}{3(2\nu+7)(2\nu+9)(2\nu+11)} \right) \\ &= \frac{15\nu(\nu+1)^2(\nu+3)^2(\nu+5)^2 c_\nu^2}{\sigma^7(\nu-2)^{7/2}(2\nu+13)^{1/2}(2\nu+7)(2\nu+9)(2\nu+11)c_{2\nu+13}}. \end{aligned}$$

Using (51) for c_ν and

$$c_{2\nu+13} = \frac{\Gamma(\nu+7)}{\sqrt{\pi(2\nu+13)}\Gamma\left(\nu+\frac{13}{2}\right)} \sim \frac{1+\frac{25}{8\nu}}{\sqrt{\pi\left(2+\frac{13}{\nu}\right)}}$$

gives the required asymptotic result.

References

Abadir, K. M. (1999), "An Introduction to Hypergeometric Functions for Economists," *Econometric Reviews*, 18, 287–330.

Abadir, K. M. and Lawford, S. (2004), "Optimal Asymmetric Kernels," *Economics Letters*, 83, 61–68.

Bowman, A. W. (1984), "An Alternative Method of Cross-validation for the Smoothing of Density Estimates," *Biometrika*, 71, 353–360.

Fan, J. and Marron, J. S. (1992), "Best Possible Constant for Bandwidth Selection," *Annals of Statistics*, 20, 2057–2070.

Hall, P., Lahiri, S. N. and Truong, Y. K. (1995), "On bandwidth choice for density estimation with dependent data," *Annals of Statistics*, 23, 2241–2263.

Hall, P. and Marron, J. S. (1987), "Estimation of Integrated Squared Density Derivatives," *Statistics and Probability Letters*, 6, 109–115.

Hall, P. and Marron, J. S. (1991), "Lower Bounds for Bandwidth Selection in Density Estimation," *Probability Theory and Related Fields*, 90, 149–173.

Hall, P., Marron, J. S. and Park, B. U. (1992), "Smoothed Cross-Validation," *Probability Theory and Related Fields*, 92, 1–20.

Hart, J. D. and Vieu, P. (1990), "Data-driven bandwidth choice for density estimation based on dependent data," *Annals of Statistics*, 18, 873–890.

Härdle, W. and Scott, D. (1992), "Smoothing by Weighted Averaging of Rounded Points," *Computational Statistics*, 7, 97–128.

Jones, M. C., Marron, J. S. and Park, B. U. (1991), "A Simple Root n Bandwidth Selector," *Annals of Statistics*, 19, 1919–1932.

Jones, M. C. and Sheather, S. J. (1991), "Using Nonstochastic Terms to Advantage in Kernel-based Estimation of Integrated Squared Density Estimates," *Statistics and Probability Letters*, 11, 511–514.

Kim, W. C. , Park, B. U. and Marron, J. S. (1994), "Asymptotically Best Bandwidth Selectors in Kernel Density Estimation," *Statistics and Probability Letters*, 19, 119–127.

Li, Q. and Racine, J. S. (2006), *Nonparametric Econometrics: Theory and Practice*, Princeton University Press, Princeton.

Marron, J. S. and Wand, M. P. (1992), "Exact Mean Integrated Squared Error," *Annals of Statistics*, 20, 712–736.

Newey, W. K. and West, K. D. (1987), "A Simple, Positive Semi-Definite, Heteroskedasticity and Autocorrelation Consistent Covariance Matrix," *Econometrica*, 55, 703–708.

Park, B. U. and Marron, J. S. (1990), "Comparison of Data Driven Bandwidth Selectors," *Journal of the American Statistical Association*, 85, 66–72.

Robinson, P. M. (2005), "Robust Covariance Matrix Estimation: HAC Estimates With Long Memory/Antipersistence Correction," *Econometric Theory*, 21, 171–180.

Rosenblatt, M. (1956), "Remarks on Some Nonparametric Estimates of a Density Function," *Annals of Mathmatical Statistics*, 27, 832–837.

Rudemo, M. (1982), "Empirical Choice of Histograms and Kernel Density Estimators," *Scandinavian Journal of Statistics*, 9, 65–78.

Scott, D. W. and Terrell, G. R. (1987), "Biased and Unbiased Cross-validation in Density Estimation," *Journal of the American Statistical Association*, 82, 1131–1146.

Silverman, B. W. (1982), "Kernel Density Estimation Using the Fast Fourier Transform," *Applied Statistics*, 31, 93–99.

Silverman, B. W. (1986), *Density Estimation for Statistics and Data Analysis*, Chapman and Hall, New-York.

Stone, C. (1984), "An Asymptotically Optimal Window Selection Rule for Kernel Density Estimates," *Annals of Statistics*, 12, 1285–1297.

Stute, W. (1992), "Modified Cross-Validation in Density Estimation," *Journal of Statistical Planning and Inference*, 30, 293–305.

Velasco, C. (2000), "Local Cross-Validation for Spectrum Bandwidth Choice," *Journal of Time Series Analysis*, 21, 329–361.

Wand, M. P. and Devroye, L. (1993), "How Easy is a Given Density to Estimate," *Computational Statistics and Data Analysis*, 16, 311–323.

Wand, M. P., Marron, J. S. and Ruppert, D. (1991), "Transformations in Density Estimation," *Journal of the American Statistical Association*, 86, 343–353.