amse
école d'économie d'aix-marseille
aix-marseille school of economics

Working Papers / Documents de travail

Explicit solutions for the asymptotically-optimal bandwidth in cross validation

Karim M. Abadir
Michel Lubrano

WP 2023 - Nr 36

Explicit solutions for the asymptotically-optimal bandwidth in cross validation *

Karim M. Abadir † Michel Lubrano‡

December 18, 2023

Abstract

We show that least squares cross-validation (CV) methods share a common structure which has an explicit asymptotic solution, when the chosen kernel is asymptotically separable in bandwidth and data. For density estimation with a multivariate Student t(ν) kernel, the CV criterion becomes asymptotically equivalent to a polynomial of only three terms. Our bandwidth formulae are simple and non-iterative (leading to very fast computations), their integrated squared-error dominates traditional CV implementations, they alleviate the notorious sample variability of CV, and overcome its breakdown in the case of repeated observations. We illustrate with univariate and bivariate applications, of density estimation and nonparametric regressions, to a large dataset of Michigan State University academic wages and experience.

Keywords: Bandwidth Choice; Cross Validation; Explicit Analytical Solution; Nonparametric Density Estimation, Academic Wages.
JEL Classification: C14, J31.

*We are grateful for comments received at the Bernoulli (ISI) Conference on Advances in Semiparametric Methods and Applications in Lisbon, International Society for Non-Parametric Statistics (ISNPS) Conference in Avignon, London-OxBridge Meeting, Netherlands Econometric Study Group, York Conference in honour of Mike Wickens, and seminars at Oxford and Greqam. We would also like to thank Jeff Racine for useful discussions and suggestions, and Omiros Papaspiliopoulos for suggesting the multivariate extension of our earlier univariate results. Support from the ESRC grants RES000230176 and RES062230790 is gratefully acknowledged.

The project leading to this publication has received funding from the French government under the "France 2030" investment plan managed by the French National Research Agency (reference :ANR-17-EURE-0020) and from Excellence Initiative of Aix-Marseille University - A*MIDEX. Complementary financing was provided by the "National Natural Science Foundation of China" (No. 71764008).

†American University in Cairo and Imperial College London, AUC Avenue, P.O. Box 74, New Cairo 11835, Egypt; London SW7 2AZ, UK. email:k.m.abadir@imperial.ac.uk, corresponding author

‡Aix-Marseille Univ., CNRS, AMSE, Marseille, France, 5 Bd Maurice Bourdet, Marseille F-13001, email:michel.lubrano@univ-amu.fr and School of Economics, Jiangxi University of Finance and Economics, Nanchang, China.

1 Introduction

Let $\{x_i\}_{i=1}^n$ be an i.i.d. sequence of the unvariate r.v. x, drawn from a density f that is a continuous function. The kernel density estimator introduced by Rosenblatt (1956) is

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

where h is the bandwidth and K is the kernel, and we can use the scaled kernels $K_h(u-x) = h^{-1}K(h^{-1}(u-x_i))$ to rewrite $\widehat{f}(u) = n^{-1}\sum_{i=1}^n K_h(u-x_i)$. The asymptotic expectation and variance of this estimator can be calculated, under the usual regularity conditions, leading to the asymptotic mean integrated squared error (AMISE)

$$\text{AMISE} = \frac{h^4}{4}k_{21}^2 I_2 + \frac{1}{nh}k_{02}, \tag{1}$$

where

$$k_{ij} = \int_{-\infty}^{\infty} t^i K(t)^j \,\mathrm{d}t, \qquad I_j = \int_{-\infty}^{\infty} f^{(j)}(u)^2 \,\mathrm{d}u,$$

the superscript $^{(j)}$ denoting the j-th derivative of the function. Minimizing the AMISE leads to

$$h_0 = k_{02}^{1/5}(nk_{21}^2 I_2)^{-1/5} \tag{2}$$

and to the Epanechnikov kernel $K_h(t) = 1_{|t|<h\sqrt{5}}3(h^2 - t^2/5)/\left(4\sqrt{5}h^3\right)$, the indicator function $1_{\mathcal{K}}$ returning 1 if condition $\mathcal{K}$ is satisfied and 0 otherwise. The multivariate generalization of the above results is in Subsection 4.2. These solutions are deterministic, but contain the unknown I_2.

It is widely recognized that a variety of kernels have good asymptotic efficiencies compared to the Epanechnikov kernel, whereas the choice of the bandwidth is crucial. For example, using the Gaussian instead of the Epanechnikov, the 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 $n^{-4/5}$. Moreover, this asymptotic optimality of the Epanechnikov kernel need not hold in finite samples and when the optimal h_0 is replaced by an estimate.

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). Instead, 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

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

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 over which the optimization is conducted. However, the last one contains both the unknown density and h. 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 ISE with its optimal h (unknown in practice) and the ISE 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.

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 I_2 in (2), and it leads to a minimum of the same AMISE objective function. However, Scott (2015, p. 179) noted that "BCV performed poorly for several difficult densities without a very large dataset."

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 et al. (1992) and its extension in Jones et al. (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 h_0, a rate that was established by Hall and Marron (1991) as $n^{1/2}$. SCV was extensively studied for multivariate density estimation in Duong (2004).

The CV method was applied to contexts other than density estimation. It is the main method for determining h in kernel regression models as illustrated in Muller (1987) and Li and Racine (2006, pp. 66–72). (The Nadaraya-Watson nonparametric regression formula is an estimate of the conditional expectation obtainable from joint densities.) Robinson and Moyeed (1989) have investigated the efficiency of various CV methods for spline smoothing regression with the objective to get a better trade-off between fit and smoothness. Other applications cover the determination of bandwidths in the estimation of spectra (such as in Velasco (2000)), the widespread Newey and West (1987) method that requires the estimation of spectra at the origin, as well as the more recent one by Robinson (2005).

None of the three CV methods introduced above give an explicit solution for their optimal h. We shall show that there is a common structure to all these CV methods, and we will use this to provide an explicit solution for their bandwidths. Furthermore, we conjecture that this structure extends to other CV problems where the objective functions can be written, locally to the optimum, as polynomial approximations in terms of h and h^{-1} upon choosing kernels from the class of "separable" kernels that we will define in the next section. The solutions we obtain are explicit (hence much quicker, by a factor of 20 in the univariate case), are more ISE-efficient than existing solutions, and solve two of the recognized problems of CV methods: their excess variability and their failure in the case of repeated observations.

2 Method for explicit solution of bandwidths

CV criteria necessitate the calculation of the $\int_{-\infty}^{\infty} \widehat{f}(u)^2 \mathrm{d}u$ seen in (3), which can be problematic if done numerically. The calculation involves a convolution that we solve explicitly here as a first step of our approach. The second step is to optimize the resulting criterion, and an explicit solution is allowed by a class of kernels that we introduce. These explicit analytical formulae will provide the speed, ISE efficiency and stability, and robustness to ties discussed earlier.

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 nonnegative integers. Define $D_h = K_h - K_0$, where K_0 is the Dirac delta function. SCV and its variants introduce an additional kernel L with bandwidth g, now requiring

$$D_h * D_h * L_g * L_g, \tag{5}$$

where L_g is the scaled version of kernel L such that $L_g(t) = g^{-1}L(g^{-1}t)$, the SCV-optimal g taking the form $\widehat{g} \sim Cn^p/\widehat{h}^2$ with C constant as $n \to \infty$ and p a constant to be detailed in Section 4. The notation $a_n \sim b_n$ means that $\lim_{n\to\infty} a_n/b_n = 1$, while $\widehat{h}$ and $\widehat{g}$ denote bandwidths that solve the optimization of a CV method. They are stochastic (unlike h_0), hence the hat.

There are two components to the solution. The first one is straightforward once we recall that the choice of a Gaussian kernel function ϕ has little effect on asymptotic efficiency while allowing simple explicit solutions, in which case we take $K = L = \phi$ to work out (4) and (5). To do so will require the Hermite polynomials

$$He_m(t) = \frac{(-1)^m \phi^{(m)}(t)}{\phi(t)} = t^m \sum_{j=0}^{1+\lfloor m/2 \rfloor} \frac{(-m)_{2j}}{j!(-2t^2)^j}, \tag{6}$$

where $m \in \mathbb{Z}_{0,+}$, $\lfloor m/2 \rfloor$ denotes the integer part of $m/2$, and $(-m)_{2j} = \prod_{i=1}^{2j}(-m+i-1)$ is Pochhammer's symbol; see Abadir (1999) for more details on He_m polynomials and their relation to the other type of Hermite polynomials denoted by H_m. See also Aldershof et al. (1995) for uses of these polynomials.

Lemma 1 *For $K = L = \phi$, (4) and (5) become, respectively,*

$$(K^{(q)} * K^{(r)})(a) = (-1)^{q+r} K_{\sqrt{2}}(a) He_{q+r}(a/\sqrt{2})/\sqrt{2^{q+r}}, \tag{7}$$

$$(D_h * D_h * L_g * L_g)(a) = K_{\sqrt{(2h^2+2g^2)}}(a) - 2K_{\sqrt{(h^2+2g^2)}}(a) + K_{g\sqrt{2}}(a), \tag{8}$$

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

The second component of the solution is to find a way to achieve asymptotic separability (in h and t) for a scaled kernel $K_h(t)$. This will allow a factorization of first-order conditions for h.

Definition 1 *A scaled kernel $K_h(t)$ is said to be asymptotically separable in h and t if its expansion around $h = 0$*

$$K_h(t) = h^{p_2} \sum\nolimits_{j \geq m} (h^{p_1})^j \psi_j(t) \qquad (0 < p_1 < \infty,\ |p_2| < \infty)$$

has a finite $m \in \mathbb{Z}$. This is a Laurent series, which generalizes Taylor series to allow for negative values of m and p_2.

This condition of a finite $m \in \mathbb{Z}$ does not hold for ϕ, but it holds for another kernel that can be made arbitrarily close to ϕ and that can be used instead of ϕ now that the convolutions have been worked out. Consider a Student t(ν) kernel, $K(t) = c_\nu/(1+t^2/\nu)^{(\nu+1)/2}$ with

$$c_\nu = \frac{\Gamma(\frac{\nu+1}{2})}{\sqrt{(\pi\nu)}\Gamma(\frac{\nu}{2})}, \quad k_{21} = \frac{\nu}{\nu-2}, \quad k_{02} = \frac{\Gamma\left(\frac{\nu}{2}+\frac{1}{2}\right)\Gamma\left(\frac{\nu}{2}+\frac{1}{4}\right)\Gamma\left(\frac{\nu}{2}+\frac{3}{4}\right)\sqrt{2}}{\nu^{\frac{3}{2}}\Gamma\left(\frac{\nu}{2}\right)^3\sqrt{\pi}}$$

from Lemma 2 of the Supplementary Material. The Gaussian is the limiting t(∞) case, but $\nu = 30$ makes the two virtually indistinguishable in practice. The scaled version of t(ν) is

$$K_h(t) = \frac{c_\nu}{h(1+t^2/(\nu h^2))^{(\nu+1)/2}} = \frac{c_\nu}{(h^2+t^2/\nu)^{(\nu+1)/2}} h^\nu. \tag{9}$$

As $\widehat{h} = O_p(n^{-1/5}) \xrightarrow{p} 0$, (9) becomes asymptotically separable in t and h since $K_h(t) = c_\nu \left(t^2/\nu\right)^{-(\nu+1)/2} h^\nu \left(1 + O(h^2)\right)$ as $h \to 0$ with $t \neq 0$ and ν finite, as implied by the binomial expansion. This asymptotic separability for small h does not hold in the Gaussian $\nu = \infty$ case, but it nevertheless

holds for any fixed large ν. This will allow subsequent derivations to give an explicit asymptotic formula for CV's solutions $\widehat{h}$. The only available expansion for the Gaussian kernel is $\exp(-t^2/(2h^2)) = 1 - t^2/(2h^2) + \ldots$, which has $m = -\infty$ in Definition 1 above, thus failing the required separability criterion on m. 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.

Separability applies to many other kernels, including the AMISE-optimal Epanechnikov

$$K_h(t) = 1_{|t|<h\sqrt{5}} \frac{3}{4\sqrt{5}} \left(\frac{1}{h} - \frac{t^2}{5h^3} \right).$$

It is not only asymptotically separable as $h \to 0$, but also exactly separable: no series expansion of a function is needed to separate h and t in $h^{-1} - h^{-3}t^2/5$. However, it is not regular because the support depends on h, but the assumption of continuity of the variate will get around the regularity issue. The case of the Epanechnikov kernel is treated in the Supplementary Material.

3 Univariate setup and illustration of simplified solution

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 as $\widehat{f}_{-j}(u) = (n-1)^{-1} \sum_{i \neq j} K_h(u - x_i)$. The last integral in the ISE in (3) is an expectation which can be estimated by $\widehat{f}_{n-1}(\boldsymbol{x}; h) = n^{-1} \sum_{j=1}^n \widehat{f}_{-j}(x_j) = n^{-1}(n-1)^{-1} \sum_{j=1}^n \sum_{i \neq j} K_h(z_{ij})$, where $\boldsymbol{x} = (x_1, \ldots, x_n)^\top$ and $z_{ij} = x_j - x_i$. UCV minimizes, with respect to h, the sum $S = S_1 + S_2 + S_3$, where

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

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

$$S_2 + S_3 \xrightarrow{p} -S_1 < 0 \tag{10}$$

by the consistency of $\widehat{f}$.

Using Lemma 1, we can work out $S_2 = n^{-1} K_{h\sqrt{2}}(0) + 2n^{-2} \sum_{j=1}^n \sum_{i>j} K_{h\sqrt{2}}(z_{ij})$, 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 $n/(n-1) = 1 + O(1/n)$,

$$S_2 + S_3 = \frac{K_{h\sqrt{2}}(0)}{n} + \frac{2 + O(1/n)}{n^2} \sum_{j=1}^{n} \sum_{i>j} \left[K_{h\sqrt{2}}(z_{ij}) - 2K_h(z_{ij}) \right], \quad (11)$$

where the first fraction is deterministic and of order $1/(nh)$. We now apply the second idea of the previous section, separable kernels, in order to tackle the optimization.

3.2 Limiting solution for simplified UCV

From (9), $K_{h\sqrt{2}}(0) = c_\nu/(h\sqrt{2})$. Applying (10) to (11), and since the UCV-optimal h is $\widehat{h} = O_p(n^{-1/5})$, it follows that the first term of (11) drops out asymptotically and the second term has a strictly negative and finite probability limit. This term that we drop (in this subsection only) is often called "diagonal" ($i = j$) or "nonstochastic". In this subsection, we will therefore minimize

$$R = 2\sum_{j=1}^{n} \sum_{i>j} K_{h\sqrt{2}}(z_{ij}) - 4\sum_{j=1}^{n} \sum_{i>j} K_h(z_{ij}), \quad (12)$$

where $R/n^2 \xrightarrow{p} -S_1 < 0$. The objective function (12) with a $\mathrm{t}(\nu)$ kernel becomes

$$R = 2c_\nu h^\nu \sum_{j=1}^{n} \sum_{i>j} \left[2^{\nu/2} \left(2h^2 + z_{ij}^2/\nu \right)^{-(\nu+1)/2} - 2\left(h^2 + z_{ij}^2/\nu \right)^{-(\nu+1)/2} \right]. \quad (13)$$

A substitution inside this double sum leads to the same UCV-optimal asymptotic solution:

Proposition 1 *For Student* $\mathrm{t}(\nu)$ *kernels and* $q \in \mathbb{R}_{0,+}$*, define the function*

$$y_n(q; \widehat{h}) = \sum_{j=1}^{n} \sum_{i>j} (\widehat{h}^2 + z_{ij}^2/\nu)^{-q-(\nu+1)/2}. \quad (14)$$

If a plug-in bandwidth, denoted by $\widehat{h}_{\mathrm{p}}$ *and satisfying* $\widehat{h}_{\mathrm{p}} = O_p(n^{-1/5})$*, is used in* $y_n(q; \widehat{h}_{\mathrm{p}})$ *only, then we get consistency of* $\widehat{f}$ *at the same rate achieved by the UCV bandwidth.*

Exploiting the asymptotic invariance of the $y_n(q; \cdot)$ function, we can rewrite the solution of optimizing R (see the first-order condition in the proof of Proposition 1) as

$$\widehat{h} = \left(\frac{\nu \left[2^{\nu/2} y_n(0; \widehat{h}_{\mathrm{p}}\sqrt{2}) - 2y_n(0; \widehat{h}_{\mathrm{p}}) \right]}{2(\nu+1) \left[2^{\nu/2} y_n(1; \widehat{h}_{\mathrm{p}}\sqrt{2}) - y_n(1; \widehat{h}_{\mathrm{p}}) \right]} \right)^{1/2}, \quad (15)$$

where the RHS makes use of plug-ins $\widehat{h}_{\mathrm{p}}$ satisfying $\widehat{h}_{\mathrm{p}} = O_p(n^{-1/5})$. By the formulation of R in (13) and the asymptotic invariance of $y_n(q;\cdot)$, we can verify that $\widehat{h}$ of (15) is of the same order as $\sqrt{(h^{-\nu}/h^{-\nu-2})}$, i.e., same order as h to be optimized. For UCV, this is $O_p(n^{-1/5})$.

Our method of solution can therefore be viewed as combining plug-in and CV approaches to get an explicit closed-form solution for the CV optimization problem. As our Proposition 1 shows, this entails no loss of asymptotic efficiency, and this will be seen to hold very well also for finite samples in the simulations of the Supplementary Material. Furthermore, as we will see with other more sophisticated CV methods below, our approach will enable us to achieve good performance that is theoretically attainable but has been elusive in practice so far because of the need to estimate unknown constants.

We now derive a plug-in to use as $\widehat{h}_{\mathrm{p}}$. We could substitute the rule of thumb $\widehat{h} = 1.06\widehat{\sigma}n^{-1/5}$ of Silverman (1986) mentioned before (3), 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 Supplementary Material, giving for $\nu > 2$

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

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.

By $R/n^2 \xrightarrow{p} -S_1 < 0$, the numerator and denominator in (15) 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 $y_n(q;\widehat{h})$ of (14), 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. In particular, the Epanechnikov kernel which is both spherical and separable leads to similar derivations whose results are in the Supplementary Material.

The combination between plug-in and CV approaches has been used also in Mammen et al. (2011). They introduce a bandwidth based on the weighted average of a plug-in method and a fully iterated CV, using Epanechnikov, quartic, and Gaussian kernels. The empirical intuition is that plug-in methods oversmooth while cross-validation ones undersmooth, and their argument for considering their combination is the important observation that practical implementation is crucial in achieving the theoretical potential of a method. However, they show that their asymptotic best weighted-average solution does not perform as well as hoped in small samples, both in term of average ISE and variability. In the Supplementary Material, our simulations show that both our Student plug-in $\widehat{h}_{\mathrm{S}}$ and our

CV solution manage to beat usual methods available in standard packages both in term of ISE and of variability. Our Proposition 1 assessed the *non-linear* combination of plug-in and CV, where the asymptotic optimality of our combination is now proved for UCV (albeit at UCV's suboptimal rate of convergence) and will similarly be established for SCV below (at the best rate of $n^{-1/2}$).

Other attempts have been made to improve the slow convergence rate of cross-validation methods. Using a kernel made of the linear combination of two Gaussian kernels, Savchuk et al. (2010) manage to reach the improved speed of $n^{-1/4}$. Their kernel is robust to rounding (ties in the data), but this implies a constrained choice for the two parameters necessary to calibrate their kernel. Our Student kernel can also be seen as a mixture (a Student density is obtained as an infinite mixture of Gaussians by a χ^2 mixing density), but with only one parameter ν to determine. Our kernel is also usable for SCV with its optimal $n^{-1/2}$ rate of convergence, as we shall see. In addition, the applications in Section 5 will show that our method is robust to rounding.

3.3 SCV criterion

Having analyzed UCV, we now introduce SCV. Jones et al. (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 ($\delta = 0$ otherwise). The result is the SCV objective function

$$S_{\mathrm{s}} = \frac{k_{02}}{nh} + \frac{\delta}{n} \left(D_h * D_h * L_g * L_g\right)(0) + \frac{1}{n^2} \sum_{j=1}^{n} \sum_{i \neq j} \left(D_h * D_h * L_g * L_g\right)(z_{ij}), \tag{17}$$

where 0 and z_{ij} are the arguments of the respective convolutions. They show that the asymptotically-optimal p in $g \sim Cn^p/h^2$ is $\widehat{p} = -23/45$ if $\delta = 1$ or $\widehat{p} = -44/85$ if $\delta = 0$, 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 $n^{-1/2}$ rate for the relative distance between the values of h minimizing MISE and S_{s}, while it is the slightly slower 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 $\widehat{g}_{\mathrm{s}} = O_p(n^{-1/10})$ dominates $\widehat{h}_{\mathrm{s}} = O_p(n^{-1/5})$. Nevertheless, the argument used for $\widehat{h}$ in connection with the Student kernel in Section 2 applies to $\widehat{g}_{\mathrm{s}}$ as well.

Although the $n^{-1/2}$ 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 $\widehat{h}_{\mathrm{s}}$. Kim et al. (1994) 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.

4 General solution for UCV and SCV

In the first subsection, we introduce the multivariate setup and give the resulting bandwidths that optimize the UCV and SCV criteria. They require a generalization of the plug-ins seen earlier to the multivariate case, which is done in the second subsection.

4.1 Solution of UCV and SCV bandwidths for multivariate kernels

Let the bandwidth matrix be $\boldsymbol{H} = h^2\boldsymbol{I}$, where $\boldsymbol{I}$ is the identity matrix. We do not tackle directly the case of $\boldsymbol{H}$ positive definite in full generality, which would require an additional $\frac{1}{2}d\,(d+1) - 1$ bandwidths to be derived. However, we will do so indirectly: we recommend orthogonalizing and normalizing the data first, then estimating the bandwidth as in this section, and finally reversing the orthonormalization. We did this in the applications of Section 5, and we will discuss both there and here below the generalization that it implies for $\boldsymbol{H}$. In this section, $\boldsymbol{x}$ now refers to the $d\times 1$ variate, rather than its $n \times d$ sample matrix which is $\boldsymbol{X} = (\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n)^\top$, and we define $\boldsymbol{z}_{ij} = \boldsymbol{x}_j - \boldsymbol{x}_i$ (whose elements are denoted by $z_{ij,m}$). The scaled kernel defined as $K_h(\boldsymbol{t}) = h^{-d}K(h^{-1}\boldsymbol{t})$ is used to write $\widehat{f}(\boldsymbol{u}) = n^{-1}\sum_{i=1}^{n} K_h(\boldsymbol{u} - \boldsymbol{x}_i)$.

The procedure for orthonormalization is as follows. Since the sample variance matrix $\boldsymbol{S}$ is positive definite, $\boldsymbol{S} = \boldsymbol{Q}\boldsymbol{\Lambda}\boldsymbol{Q}^\top$ and the square root of the matrix is the symmetric $\boldsymbol{S}^{1/2} = \boldsymbol{Q}\boldsymbol{\Lambda}^{1/2}\boldsymbol{Q}^\top$, with $\boldsymbol{\Lambda}$ the diagonal matrix of positive eigenvalues of $\boldsymbol{S}$, the columns of $\boldsymbol{Q}$ contain the orthonormal eigenvectors of $\boldsymbol{S}$, and $\boldsymbol{Q}\boldsymbol{Q}^\top = \boldsymbol{I}$. The orthonormalization is then $\boldsymbol{y} = \boldsymbol{S}^{-1/2}\boldsymbol{x}$ which has $\widehat{\mathrm{var}}(\boldsymbol{y}) = \boldsymbol{S}^{-1/2}\widehat{\mathrm{var}}(\boldsymbol{x})\boldsymbol{S}^{-1/2} = \boldsymbol{I}$, where $\boldsymbol{S}^{-1/2} = \boldsymbol{Q}\boldsymbol{\Lambda}^{-1/2}\boldsymbol{Q}^\top$; see Abadir and Magnus (2005) for matrix functions. (In general, the components of $\boldsymbol{y}$ are uncorrelated but mutually dependent.) Under general conditions, the sample variance is a consistent estimator of $\mathrm{var}(\boldsymbol{x})$ when it exists. Our paper is about asymptotically-optimal bandwidth formulae. These can no doubt be refined, but further support for our approach can be seen in the convergence results cited in the multivariate section of the Supplementary Material where we also have bandwidth formulae for the case of product kernels, in addition to the ones below which are for multivariate kernels.

The scaled multivariate t(ν) kernel is

$$K_h(\boldsymbol{t}) = c_{\nu,d}\,|\boldsymbol{H}|^{-1/2}\left(1+\frac{1}{\nu}\boldsymbol{t}^\top\boldsymbol{H}^{-1}\boldsymbol{t}\right)^{-(\nu+d)/2} = c_{\nu,d}h^\nu\left(h^2+\frac{1}{\nu}\sum_{m=1}^{d}t_m^2\right)^{-(\nu+d)/2},$$

where $c_{\nu,d} = (\pi\nu)^{-d/2}\,\Gamma\left(\frac{\nu+d}{2}\right)/\Gamma\left(\frac{\nu}{2}\right)$ generalizes the univariate $c_\nu = c_{\nu,1}$. In the case of a spherical multivariate kernel, such as here, the quadratic form in $\boldsymbol{t}$ shows that our procedure (orthonormalizing the data first) could be alternatively interpreted as having $\boldsymbol{H}$ proportional to the sample's variance matrix $\boldsymbol{S}$, since $\boldsymbol{y}^\top\boldsymbol{y} = \boldsymbol{x}^\top\boldsymbol{S}^{-1}\boldsymbol{x}$ in terms of the original data $\boldsymbol{x}$. This equivalence will not hold for the product kernels in the Supplementary Material, hence the general setup (of orthonormalizing the data first then using $\boldsymbol{H} = h^2\boldsymbol{I}$) introduced in this section.

Theorem 2 *Let $\widehat{h}$ denote the solution of a CV-optimal bandwidth, then we use $\widehat{h}_{\rm a}$ to denote our asymptotic solution satisfying $\lim_{n\to\infty}\widehat{h}_{\rm a}/\widehat{h} = 1$ and $\widehat{h}_{\rm aa}$ the leading term of its asymptotic expansion. Take plug-ins $\widehat{h}_{\rm p}, \widehat{g}_{\rm p}$ satisfying $\widehat{h}_{\rm p} = O(n^{-1/(4+d)})$ and $\widehat{g}_{\rm p} = O(n^{-1/(6+d)})$.*
(a) For UCV, with $y_n(q;h) = \sum_{j=1}^{n}\sum_{i>j}(h^2+\frac{1}{\nu}\boldsymbol{z}_{ij}^\top\boldsymbol{z}_{ij})^{-q-(\nu+d)/2}$, letting $\alpha_1 = 2^{-1-d/2}dn$,

$$\alpha_2 = \nu\left[2^{\nu/2}y_n(0;\widehat{h}_{\rm p}\sqrt{2}) - 2y_n(0;\widehat{h}_{\rm p})\right],\quad \alpha_3 = -2\,(\nu+d)\left[2^{\nu/2}y_n(1;\widehat{h}_{\rm p}\sqrt{2}) - y_n(1;\widehat{h}_{\rm p})\right],$$

we have

$$\widehat{h}_{\rm a} = \left(\frac{\alpha_1}{\alpha_2+\alpha_3\widehat{h}_{\rm p}^2}\right)^{1/(\nu+d)} \qquad \textit{and} \qquad \widehat{h}_{\rm aa} = (-\alpha_2/\alpha_3)^{1/2}. \tag{18}$$

(b) For SCV, with $y_n(q;h,g) = \sum_{j=1}^{n}\sum_{i>j}(h^2+2g^2+\frac{1}{\nu}\boldsymbol{z}_{ij}^\top\boldsymbol{z}_{ij})^{-q-(\nu+d)/2}$, letting

$$k_{02,d} = \left(\frac{\nu}{2\nu+d}\right)^{d/2}\frac{\left((\pi\nu)^{-d/2}\,\Gamma\left(\frac{\nu+d}{2}\right)/\Gamma\left(\frac{\nu}{2}\right)\right)^2}{\left(\pi\,(2\nu+d)\right)^{-d/2}\,\Gamma\,(\nu+d)\,/\Gamma\left(\nu+\frac{d}{2}\right)}, \tag{19}$$

$$\alpha_1 = \frac{k_{02,d}dn}{4c_{\nu,d}} + \frac{\delta dn}{2}\left((2+2n^{1/5})^{-1-d/2} - (1+2n^{1/5})^{-1-d/2}\right),$$
$$\alpha_2 = \nu\left[(2+2n^{1/5})^{(\nu-2)/2}y_n(0;\widehat{h}_{\rm p}\sqrt{2},\widehat{g}_{\rm p}) - (1+2n^{1/5})^{(\nu-2)/2}y_n(0;\widehat{h}_{\rm p},\widehat{g}_{\rm p})\right],$$
$$\alpha_3 = -\,(\nu+d)\left[(2+2n^{1/5})^{\nu/2}y_n(1;\widehat{h}_{\rm p}\sqrt{2},\widehat{g}_{\rm p}) - (1+2n^{1/5})^{\nu/2}y_n(1;\widehat{h}_{\rm p},\widehat{g}_{\rm p})\right],$$

we have

$$\widehat{h}_{\rm a} = \left(\frac{\alpha_1}{\alpha_2+\alpha_3\widehat{h}_{\rm p}^2}\right)^{1/(\nu+d)} \textit{and}\ \widehat{h}_{\rm aa} = \left(\frac{y_n(0;\widehat{h}_{\rm p}\sqrt{2},\widehat{g}_{\rm p}) - y_n(0;\widehat{h}_{\rm p},\widehat{g}_{\rm p})}{\left(1+\frac{d}{\nu}\right)\left[y_n(1;\widehat{h}_{\rm p}\sqrt{2},\widehat{g}_{\rm p}) - y_n(1;\widehat{h}_{\rm p},\widehat{g}_{\rm p})\right]} - 2\widehat{g}_{\rm aa}^2\right)^{1/2} \tag{20}$$

with

$$\widehat{g}_{\text{aa}} = \left(\frac{y_n(0; \widehat{h}_{\text{p}}, \widehat{g}_{\text{p}}) - y_n(0; 0, \widehat{g}_{\text{p}})}{2\left(1 + \frac{d}{\nu}\right)\left[y_n(1; \widehat{h}_{\text{p}}, \widehat{g}_{\text{p}}) - y_n(1; 0, \widehat{g}_{\text{p}})\right]} \right)^{1/2}. \tag{21}$$

The solutions $\widehat{h}_{\text{a}}$ require $\alpha_2 + \alpha_3 \widehat{h}_{\text{p}}^2 > 0$, which is guaranteed in large samples but might fail in small samples. If so, then the simpler asymptotic approximations $\widehat{h}_{\text{aa}}$ should be used instead. As for the plug-ins, in the univariate case we can use $\widehat{h}_{\text{p}}$ of (16), with $\nu > 2$, and the simple

$$\widehat{g}_{\text{p}} = \frac{\widehat{h}_{\text{p}}}{n^{-1/5}} n^{-1/10} = \widehat{h}_{\text{p}} n^{1/10} \tag{22}$$

from the discussion following (17); but the multivariate case requires the next subsection.

4.2 Multivariate plug-ins

Multivariate plug-in $\widehat{h}_{\text{p}}$. We consider the multivariate version of $h_0 = k_{02}^{1/5}(nk_{21}^2 I_2)^{-1/5}$ of (2) and recalculate its components to get $\widehat{h}_{\text{p}}$ in the case of a multivariate Student t(ν). Silverman's rule for variates with unit variance matrix is

$$\left(\frac{4}{(2+d)n} \right)^{1/(4+d)} \tag{23}$$

which is approximated by Scott (2015) as $n^{-1/(4+d)}$ since the constant ratio is always between 0.92 and 1.06 with $\lim_{d\to\infty} (4/(2+d))^{1/(4+d)} = 1$.

The multivariate AMISE generalizing (1) is

$$\text{AMISE} = \frac{h^4}{4} k_{21}^2 I_2 + \frac{1}{nh^d} k_{02,d} \tag{24}$$

leading to the generalization of (2) as

$$h_0 = \left(\frac{k_{02,d}\, d}{nk_{21}^2 I_2} \right)^{1/(4+d)}, \tag{25}$$

where $k_{21} = \nu/(\nu-2)$ as before, $k_{02,d}$ is in (19), and $I_2 = \int_{\mathbb{R}^d} (\sum_{j=1}^d \partial^2 f(\boldsymbol{u})/\partial u_j^2)^2 \,\mathrm{d}\boldsymbol{u}$ now; e.g., see Hardle and Muller (2000). It remains for us to work out I_2 for a multivariate Student t(ν) plug-in density, that is, our generalized Silverman's rule now multivariate. From Lemma 2(iv),

$$\begin{aligned} I_2 &= \frac{d\,(2+d)}{2^{\nu+d+1}\pi^{(d-1)/2} b^{d+4} \nu^{2+d/2}} \frac{\Gamma\left(\frac{\nu+d}{2}+2\right)\Gamma\left(\nu+\frac{d}{2}+2\right)}{\left(\Gamma\left(\frac{\nu}{2}\right)\right)^2 \Gamma\left(\frac{\nu+d+5}{2}\right)} \\ &\sim \frac{d\,(2+d)}{2^{d+2}\pi^{d/2} b^{d+4}} \left(1 + \frac{(d+4)\,(d+2)}{4\nu}\right)\left(1 + \frac{d\,(d+4)}{16\nu}\right)\left(1 - \frac{(d+4)\,(3d+12)}{16\nu}\right) \end{aligned} \tag{26}$$

giving

$$\lim_{\nu\to\infty} I_2 = \frac{d\,(2+d)}{2^{d+2}\pi^{d/2}\sigma^{d+4}}, \qquad \lim_{\nu\to\infty} h_0 = \left(\frac{4}{(2+d)\,n}\right)^{1/(4+d)} \sigma$$

which is Silverman's multivariate rule (23) when the scalar variance matrix is set to unity. Our extension of his rule for general ν follows from now having all the ingredients for (25) as

$$\widehat{h}_{\mathrm{S}} = \left(\frac{k_{02,d}d}{n\,(1-2/\nu)^2\,\widehat{I}_2}\right)^{1/(4+d)}, \tag{27}$$

where we use $\widehat{b} = \sqrt{(1-2/\nu)}$ in $\widehat{I}_2$ (with unit variance).

Multivariate plug-in $\widehat{g}_{\mathrm{p}}$. Theorem 3 of Duong and Hazelton (2005) implies a $\widehat{g}_{\mathrm{p}}$ for SCV, which is denoted there by g_1. Its formula is quite elaborate and requires combinations of sixth-order partial derivatives to be evaluated, but Duong (2007) gives a numerical way to compute these, which yields a $\widehat{g}_{\mathrm{p}}$ that we can use here. Alternatively, a rough approximation can be obtained by comparing their theorem's g_1 with (23) to get the relation

$$\frac{\widehat{g}_{\mathrm{p}}}{\widehat{h}_{\mathrm{p}}} \sim \frac{n^{-1/(6+d)}}{n^{-1/(4+d)}} = n^{\frac{2}{(6+d)(4+d)}},$$

as we had for the univariate case of (22), and we get

$$\widehat{g}_{\mathrm{p}} = \widehat{h}_{\mathrm{p}} n^{\frac{2}{(6+d)(4+d)}}. \tag{28}$$

However, $d = 1$ here would give $\widehat{g}_{\mathrm{p}} = \widehat{h}_{\mathrm{p}} n^{2/35}$, an overestimate (by $n^{3/70}$) of $\widehat{g}_{\mathrm{p}}$ compared to $\widehat{g}_{\mathrm{p}} = \widehat{h}_{\mathrm{p}} n^{1/10}$ of (22). Since the notation for orders of magnitude is an inequality relation, we adopt the larger order $\widehat{g}_{\mathrm{p}} = O(n^{-1/(6+d)})$ used in the optimality derivations of Duong and Hazelton (2005). For $d = 1$ in typical samples like 100 to 1000, the difference is 22% to 34%. Such differences do not have a large impact in practice, as will be seen in the next section, but much larger samples could require the calculations of Duong (2007) instead of the rough (28).

5 Academic wages at Michigan State University

We provide an empirical application on the distribution of academic wages and experience in the Michigan State University large database for 2012. An additional practical advantage of our explicit formulae is to avoid the troubles faced by existing CV approaches when there are some ties in the data, in this case some equal salaries and/or experience.

The database contains 6,402 entries (after deleting 22 lines which corresponded to a null wage). Deleting duplicate names, as the same person can be appointed by several departments, we were left with $n = 5,050$ distinct individuals earning 4,070 different salaries. The minimum yearly wage is \$3,600 (due to part time). The first quartile is \$52,070. The mean wage is \$90,380. The 0.995 quantile is \$298,832. The maximum yearly wage is \$952,400 and corresponds to a fixed-term contract on an endowed Chair of the chemistry department. We want to make inference on the wage distribution and then on the bivariate relation between wages and experience.

Table 1: Bandwidths for the wage dataset of Michigan State University

		Student kernel			Gaussian kernel		
ν	$\widehat{h}_{\mathrm{S}}$	UCV $\widehat{h}_{\mathrm{a}}$	SCV $\widehat{h}_{\mathrm{a}}$	SCV $\widehat{h}_{\mathrm{aa}}$	Silverman	UCV	SCV
4	5.513	2.84	5.828	5.353	8.95	1.17	4.15
6	7.182	4.204	7.485	6.970			

This very asymmetric wage distribution has a Kolmogorov-Smirnov measured complexity equal to $d_n(\boldsymbol{x}) = 0.120$, as per the implementation details in the Supplementary Material, which would suggest choosing between $\nu = 4$ and $\nu = 6$. In Table 1, we present our various implied choices for a bandwidth and the alternative answers of the literature. They illustrate the usual breakdown of standard UCV in the presence of repeated observations. One of the assumptions needed for using a cross-validation method is that the observations are draws from a continuous random variable. Otherwise, the presence of a point mass piling up is detected by least squares cross-validation, which then chooses a small bandwidth to deal with these point masses. Wage datasets typically contain point mass piling up as several individuals (those with the same qualification and experience) tend to have similar wages. The value obtained for standard UCV corresponds to the lower bound of the grid search of the Brent algorithm of `bw.ucv` in `R`. At the other extreme, the Silverman rule `bw.nrd` in `R` gives the highest value. (The plug in was obtained as `bw.nrd(x)` in `R`. The unmodified traditional formula $\widehat{h} = 1.06\widehat{\sigma}n^{-1/5}$ produces an even larger value of 10.41.) Both represent unreliable window sizes, the effect of which is depicted on the second panel of Figure 1, over-smoothing and under-smoothing.

None of our formulae suffer these drawbacks. Our UCV's $\widehat{h}_{\mathrm{a}}$ helps to identify small details of the wage distribution, while our two integral-free SCV $\widehat{h}_{\mathrm{a}}$ and $\widehat{h}_{\mathrm{aa}}$ (and also our generalized rule of Silverman $\widehat{h}_{\mathrm{S}}$) give a smoother density. We use $\nu = 6$ in the left panel of Figure 1 and see the following features. The wage density presents several bumps that are well identified when using SCV $\widehat{h}_{\mathrm{aa}}$ which we find here to be the best method, also because of the recommendation to use it for asymmetric densities (see our simulations in Table 8 of the Supplementary Material). The two main

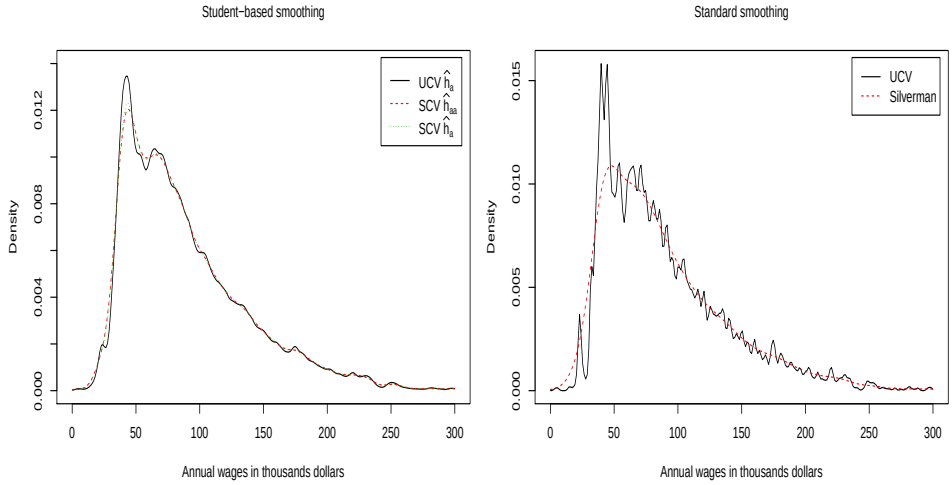

Figure 1: Wage density estimation for Michigan State University in 2012

modes of the distribution are well identified with all our methods. Our UCV's $\widehat{h}_{\mathrm{a}}$, even if it provides slightly more variability than our other bandwidths, helps to identify the first very small mode of the distribution which corresponds to 49 Teachers with a *fixed-term* contract and who are all paid \$22,870 a year. The second and main mode is at \$43,374. It corresponds mainly to Research Associates with a *fixed-term* contract. The third mode is at \$64,868. Around this mode, most wages correspond either to Specialists or to Assistant Professors with a labour contract which is either *Not Tenured/Continuing System* or *Tenure System Probationary*. Around these last two modes, there are several identical wages.

A Mincer equation explains log(wages) as a function of years of experience, with the idea that the yield of experience should decrease when approaching retirement. This relation is well depicted by a bivariate contour for those who are tenured, which concerns 1,545 members of the university. In Table 2, we present in a first block the $\boldsymbol{H}$ matrices obtained using the `R` package `ks` of Duong (2022). It yielded unusual values for UCV because of the presence of repeated observations. In a second block of Table 2, we provide the same quantities for our formulae based on $\widehat{h}_{\mathrm{aa}}$ with a multivariate Student kernel having $\nu = 6$. The results are more in accordance with what one would expect, unaffected by repeated observations. UCV corresponds to some under-smoothing, while SCV is between the Student-generalized plug-in and UCV.

The four plots reported in Figure 2 illustrate this relation between log-wages and experience, as the contours are pointing up but flattening when experience increases. This nonlinear relation is also seen from three non-

Table 2: Bandwidth 2×2 matrices for the tenured subsample

Plug-in		UCV		SCV	
Duong `ks`					
2.600	0.035	0.000	0.000	3.050	0.045
0.035	0.009	0.000	0.034	0.045	0.011
Student kernel $\nu = 6$					
4.280	0.024	2.880	0.016	3.518	0.020
0.024	0.006	0.016	0.004	0.020	0.005

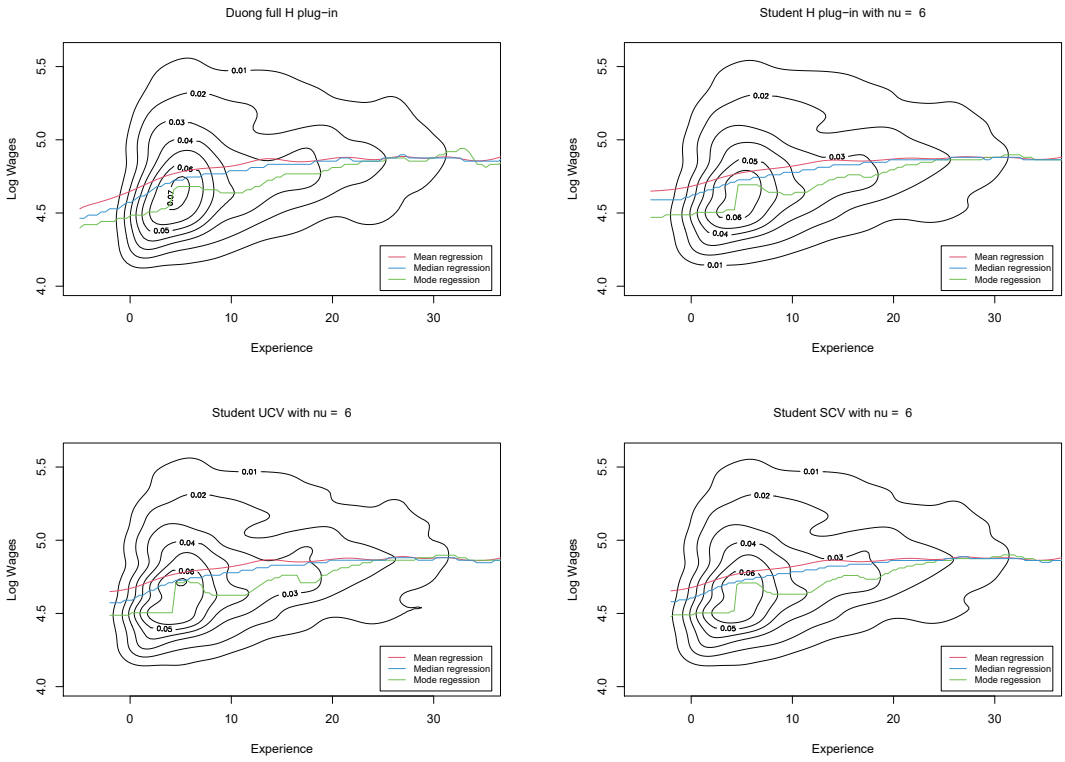

Figure 2: Bivariate density of log-wages and experience, using various methods and a multivariate kernel

parametric regressions for each plot, by taking a sequence of experience levels (vertical lines) and calculating the mean, median (which is less sensitive to outliers than the mean), and peak mode of the conditional densities at each of these experience levels. The Nadaraya-Watson regression with Silverman's bandwidth turns out to be almost the same as the mean regression in the second plot in Figure 2, as expected from it being a conditional expectation. The curves we get from our formulae are less volatile than those obtained by other CV estimates, as indicated earlier, and we can see

this here when we compare them with the standard SCV reported in the first plot in Figure 2. Note that some of the crossovers of curves in Figure 2 can provide counterexamples to the mean-median-mode inequality, in the case of a unimodal (or nearly so) conditional distribution, in addition to the ones in Abadir (2005). We conclude by cautioning that this regression is incomplete because other variables also determine log-wages in academia.

References

Abadir, K. M. (1999), 'An introduction to hypergeometric functions for economists', *Econometric Reviews* **18**, 287–330.

Abadir, K. M. (2005), 'The mean-median-mode inequality: counterexamples', *Econometric Theory* **21**, 477–482.

Abadir, K. M., Heijmans, R. D. H. and Magnus, J. R. (2018), *Statistics*, Cambridge University Press, Cambridge.

Abadir, K. M. and Lawford, S. (2004), 'Optimal asymmetric kernels', *Economics Letters* **83**, 61–68.

Abadir, K. M. and Magnus, J. R. (2005), *Matrix Algebra*, Cambridge University Press, Cambridge.

Aldershof, B., Marron, J., Park, B. and Wand, M. (1995), 'Facts about the gaussian probability density function', *Applicable Analysis* **59**(1-4), 289–306.

Bowman, A. W. (1984), 'An alternative method of cross-validation for the smoothing of density estimates', *Biometrika* **71**, 353–360.

Duong, T. (2004), Bandwidth Selectors for Multivariate Kernel Density Estimation, PhD thesis, School of Mathematics and Statistics, University of Western Australia.

Duong, T. (2007), 'ks: Kernel density estimation and kernel discriminant analysis for multivariate data in R', *Journal of Statistical Software* **21**, 1–16.

Duong, T. (2022), 'ks: Kernel smoothing. r package version 1.14.0.', https://CRAN.R-project.org/package=ks.

Duong, T. and Hazelton, M. L. (2005), 'Cross-validation bandwidth matrices for multivariate kernel density estimation', *Scandinavian Journal of Statistics* **32**, 485–506.

Fan, J. and Marron, J. S. (1992), 'Best possible constant for bandwidth selection', *Annals of Statistics* **20**, 2057–2070.

Faraway, J. and Jhun, M. (1990), 'Bootstrap choice of bandwidth for density estimation', *Journal of the American Statistical Association* **85**, 1119–1122.

Hall, P. and Marron, J. (1987), 'Estimation of integrated squared density derivatives', *Statistics Probability Letters* **6**, 109–115.

Hall, P. and Marron, J. (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.

Hardle, W. and Muller, M. (2000), Multivariate and semiparametric kernel regression, *in* M. G. Schimek, ed., 'Smoothing and Regression', John Wiley & Sons, Ltd, chapter 12, pp. 357–391.

Hayfield, T. and Racine, J. (2008), 'Nonparametric econometrics: The np package', *Journal of Statistical Software* **27**, 1–32.

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* **6**, 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.

Mammen, E., Miranda, M., Nielsen, J. and Sperlich, S. (2011), 'Do-validation for kernel density estimation', *Journal of the American Statistical Association* **106**, 651–660.

Marron, J. S. and Nolan, D. (1989), 'Canonical kernels for density estimation', *Statistics and Probability Letters* **7**, 195–199.

Muller, H.-G. (1987), 'Weighted local regression and kernel methods for nonparametric curve fitting', *Journal of the American Statistical Association* **82**, 231–238.

Newey, W. K. and West, K. D. (1987), 'A simple, positive semi-definite, heteroskedasticity and autocorrelation consistent covariance matrix', *Econometrica* **55**, 703–708.

Robinson, P. M. (2005), 'Robust covariance matrix estimation: Hac estimates with long memory/antipersistence correction', *Econometric Theory* **21**, 171–180.

Robinson, T. and Moyeed, R. (1989), 'Making robust the cross-validatory choice of smoothing parameter in spline smoothing regression', *Communications in Statistics - Theory and Methods* **18**, 523–539.

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.

Sain, S., Baggerly, K. and Scott, D. (1994), 'Cross-validation of multivariate densities', *Journal of the American Statistical Association* **89**, 807–817.

Savchuk, O., Hart, J. and Sheather, S. (2010), 'Indirect cross-validation for density estimation', *Journal of the American Statistical Association* **105**, 415–423.

Scott, D. W. (2015), *Multivariate Density Estimation: Theory, Practice and Visualization*, Wiley Series in Probability and Statistics, Wiley.

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. (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.

Technical appendix

A Solution to univariate UCV, BCV, SCV

The solution details to the UCV, BCV, and SCV bandwidths are given in three subsections.

A.1 UCV

Omitting only the term denoted by $O(1/n)$ in (11), but not the first deterministic term which is now $c_\nu/(nh\sqrt{2})$, similar derivations lead to the first-order condition

$$\begin{aligned} \frac{n}{2\sqrt{2}} &= \nu\widehat{h}_{\mathrm{u}}^{\nu+1}\left[2^{\nu/2}y_n(0;\widehat{h}_{\mathrm{u}}\sqrt{2}) - 2y_n(0;\widehat{h}_{\mathrm{u}})\right] \\ &\quad -2\left(\nu+1\right)\widehat{h}_{\mathrm{u}}^{\nu+3}\left[2^{\nu/2}y_n(1;\widehat{h}_{\mathrm{u}}\sqrt{2}) - y_n(1;\widehat{h}_{\mathrm{u}})\right], \end{aligned} \tag{29}$$

where $\widehat{h}_{\mathrm{u}}$ is the UCV solution. As before, the content of the square brackets can be approximated asymptotically by using $\widehat{h}_{\mathrm{p}}$. This makes (29) an equation of the form $\alpha_1 = \alpha_2\widehat{h}^{\nu+1} + \alpha_3\widehat{h}^{\nu+3}$, 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

$$\widehat{h} = \left(\frac{\alpha_1}{\alpha_2 + \alpha_3\widehat{h}^2}\right)^{1/(\nu+1)}, \tag{30}$$

which can be approximated for $\nu > 2$ by applying $\widehat{h}_{\mathrm{p}}^2$ of (16) to the RHS of (30), giving an explicit asymptotic formula for $\widehat{h}$ which we will call $\widehat{h}_{\mathrm{a}}$:

$$\widehat{h}_{\mathrm{a}} = \left(\frac{\alpha_1}{\alpha_2 + \alpha_3\widehat{h}_{\mathrm{p}}^2}\right)^{1/(\nu+1)}. \tag{31}$$

This solution requires $\alpha_2 + \alpha_3\widehat{h}_{\mathrm{p}}^2 > 0$, which is guaranteed in large samples but might fail in small samples. If so, then the simpler asymptotic approximation (15), reexpressed as

$$\widehat{h}_{\mathrm{aa}} = \left(-\alpha_2/\alpha_3\right)^{1/2}, \tag{32}$$

should be used instead. Iterating (30), instead of using $\widehat{h}_{\mathrm{p}}^2$ in (31), would give the exact UCV solution except for the inconsequential approximation of $1/\left(n-1\right)$ by $1/n$ in the objective function (11).

A.2 BCV

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

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

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 $He_4(a) = a^4 - 6a^2 + 3$ which is calculated from formula (6) for Hermite polynomials, we get

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

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

As before, using the Student t(ν) kernel (9) with $h\sqrt{2}$ instead of h, as required for (33), we get

$$S_{\mathrm{b}} = \frac{k_{02}}{nh} + \frac{c_\nu k_{21}^2}{8\sqrt{2}n^2} \sum_{j=1}^{n} \sum_{i>j} \left(\frac{z_{ij}^4}{4} h^{\nu-4} - 3z_{ij}^2 h^{\nu-2} + 3h^\nu \right) (h^2 + z_{ij}^2/(2\nu))^{-(\nu+1)/2} \tag{34}$$

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

$$\begin{aligned} & \frac{8\sqrt{2}k_{02}n}{c_\nu k_{21}^2} \\ = \; & \widehat{h}_{\mathrm{b}}^{\nu-3} \sum_{j=1}^{n} \sum_{i>j} \left(\left(\frac{\nu}{4} - 1 \right) z_{ij}^4 - 3(\nu-2) z_{ij}^2 \widehat{h}_{\mathrm{b}}^2 + 3\nu \widehat{h}_{\mathrm{b}}^4 \right) \left(\widehat{h}_{\mathrm{b}}^2 + z_{ij}^2/(2\nu) \right)^{-(\nu+1)/2} \\ & -(\nu+1) \widehat{h}_{\mathrm{b}}^{\nu-1} \sum_{j=1}^{n} \sum_{i>j} \left(\frac{z_{ij}^4}{4} - 3z_{ij}^2 \widehat{h}_{\mathrm{b}}^2 + 3\widehat{h}_{\mathrm{b}}^4 \right) \left(\widehat{h}_{\mathrm{b}}^2 + z_{ij}^2/(2\nu) \right)^{-(\nu+3)/2}, \end{aligned} \tag{35}$$

where $\widehat{h}_{\mathrm{b}}$ is the BCV solution. The same approach used in Proposition 1 about $\widehat{h} = O_p(n^{-1/5})$ indicate that this is essentially an equation of the form $\alpha_1 = \alpha_2 \widehat{h}^{\nu-3} + \alpha_3 \widehat{h}^{\nu-1}$, which leads to

$$\widehat{h}_{\mathrm{a}} = \left(\frac{\alpha_1}{\alpha_2 + \alpha_3 \widehat{h}_{\mathrm{p}}^2} \right)^{1/(\nu-3)}. \tag{36}$$

We can make the same remark as before concerning the positivity of $\alpha_2 + \alpha_3 \widehat{h}_{\mathrm{p}}^2$, but this time we have a supplementary restriction on the value of ν

which should be greater than 4. In addition, like (32) was a simplification of (31), here we have the simplifying asymptotic approximation

$$\widehat{h}_{\rm aa} = (-\alpha_2/\alpha_3)^{1/2}\,. \tag{37}$$

We use it instead of (36) whenever $\alpha_2 + \alpha_3 \widehat{h}_{\rm p}^2 < 0$.

A.3 SCV

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

$$\begin{aligned} S_{\rm s} &= \frac{k_{02}}{nh} + \frac{\delta c_\nu}{n\sqrt{2}}\left(\frac{1}{\sqrt{(h^2+g^2)}} - \frac{2^{3/2}}{\sqrt{(h^2+2g^2)}} + \frac{1}{g}\right) \\ &\quad + \frac{2c_\nu}{n^2}\left[\left(2h^2+2g^2\right)^{\nu/2} y_n(0;h\sqrt{2},g) - 2\left(h^2+2g^2\right)^{\nu/2} y_n(0;h,g) + 2^{\nu/2} g^\nu y_n(0;0,g)\right], \end{aligned} \tag{38}$$

where

$$y_n(q;h,g) = \sum_{j=1}^{n}\sum_{i>j}\left(h^2+2g^2+z_{ij}^2/\nu\right)^{-q-(\nu+1)/2}. \tag{39}$$

Since $\partial y_n(0;h\sqrt{2},g)/\partial h = (h/g)\,\partial y_n(0;h\sqrt{2},g)/\partial g$ and

$$\frac{\partial y_n(0;h,g)}{\partial g} = \frac{2g}{h}\frac{\partial y_n(0;h,g)}{\partial h} = -2\,(\nu+1)\,g\; y_n(1;h,g),$$

defining $y_n^\dagger(q;h,g) = \left(h^2+2g^2\right)^{q-1+\nu/2} y_n(q;h,g)$ allows us to write the exact first-order conditions for g and h, respectively, as

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

and

$$\begin{aligned} &\frac{k_{02}n}{4c_\nu\widehat{h}_{\rm s}^3} + \frac{\delta n}{2^{5/2}}\left(\frac{1}{(\widehat{h}_{\rm s}^2+\widehat{g}_{\rm s}^2)^{3/2}} - \frac{2^{3/2}}{(\widehat{h}_{\rm s}^2+2\widehat{g}_{\rm s}^2)^{3/2}}\right) \\ &= \nu\left[y_n^\dagger(0;\widehat{h}_{\rm s}\sqrt{2},\widehat{g}_{\rm s}) - y_n^\dagger(0;\widehat{h}_{\rm s},\widehat{g}_{\rm s})\right] - (\nu+1)\left[y_n^\dagger(1;\widehat{h}_{\rm s}\sqrt{2},\widehat{g}_{\rm s}) - y_n^\dagger(1;\widehat{h}_{\rm s},\widehat{g}_{\rm s})\right], \end{aligned} \tag{41}$$

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

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

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

As in Proposition 1, the asymptotic invariance of the $y_n(q;\cdot,\cdot)$ function here 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 (16), and $\widehat{g}_{\rm p}$ is defined as

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

hence replacing $y_n^{\dagger}(q;a\widehat{h}_{\rm s},\widehat{g}_{\rm s})$ in (41) and (42) by $(a^2\widehat{h}_{\rm s}^2+2\widehat{g}_{\rm s}^2)^{q-1+\nu/2}y_n(q;a\widehat{h}_{\rm p},\widehat{g}_{\rm p})$ for all q and a here, 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, and we get

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

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

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

yielding

$$\widehat{h}_{\rm aa}=\left(\frac{y_n(0;\widehat{h}_{\rm p}\sqrt{2},\widehat{g}_{\rm p})-y_n(0;\widehat{h}_{\rm p},\widehat{g}_{\rm p})}{\left(1+\frac{1}{\nu}\right)\left[y_n(1;\widehat{h}_{\rm p}\sqrt{2},\widehat{g}_{\rm p})-y_n(1;\widehat{h}_{\rm p},\widehat{g}_{\rm p})\right]}-2\widehat{g}_{\rm aa}^2\right)^{1/2}. \tag{46}$$

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

$$\begin{aligned}&\frac{k_{02}n}{4c_\nu}+\frac{\delta n}{2^{5/2}}\left(\frac{1}{(1+n^{1/5})^{3/2}}-\frac{2^{3/2}}{(1+2n^{1/5})^{3/2}}\right) \\ =\;&\widehat{h}^{\nu+1}\nu\left[(2+2n^{1/5})^{(\nu-2)/2}y_n(0;\widehat{h}_{\rm p}\sqrt{2},\widehat{g}_{\rm p})-(1+2n^{1/5})^{(\nu-2)/2}y_n(0;\widehat{h}_{\rm p},\widehat{g}_{\rm p})\right] \\ &-\widehat{h}^{\nu+3}\left(\nu+1\right)\left[(2+2n^{1/5})^{\nu/2}y_n(1;\widehat{h}_{\rm p}\sqrt{2},\widehat{g}_{\rm p})-(1+2n^{1/5})^{\nu/2}y_n(1;\widehat{h}_{\rm p},\widehat{g}_{\rm p})\right],\end{aligned} \tag{47}$$

which is a polynomial of the form $\alpha_1=\alpha_2\widehat{h}^{\nu+1}+\alpha_3\widehat{h}^{\nu+3}$ yielding as before

$$\widehat{h}_{\rm a}=\left(\frac{\alpha_1}{\alpha_2+\alpha_3\widehat{h}_{\rm p}^2}\right)^{1/(\nu+1)} \tag{48}$$

if we use the plug-in $\widehat{h}_{\mathrm{p}}$ on the RHS. However, unlike before, it is not the case that $\widehat{h}_{\mathrm{aa}}$ of (46) equals $\sqrt{(-\alpha_2/\alpha_3)}$, because of the presence of terms like $(a^2+2n^{1/5})^a$ that are due to g.

B Other kernels

There are three subsections here. First, we derive our method's bandwidths when one uses the separable Epanechnikov kernel. Second, we report a simulation experiment to show that the Epanechnikov kernel is best for estimating a density that is Gaussian, but that a Student kernel is better in all other cases. This is why the main text reports the application of our method to the Student kernel, and this Supplementary Material the rest.

Third, we derive Student kernels' AMISE, which is needed to translate AMISE-optimal bandwidths from one kernel to another. This result applies to Epanechnikov and other kernels.

B.1 Epanechnikov kernel

UCV with Epanechnikov kernel. Derivations similar to (15) yield:

$$\widehat{h}=\left(\frac{3\sum_{j=1}^{n}\sum_{i>j}\left(1_{|z_{ij}|<\sqrt{5}\widehat{h}_{\mathrm{p}}}4\sqrt{2}-1_{|z_{ij}|<\sqrt{10}\widehat{h}_{\mathrm{p}}}\right)z_{ij}^2}{10\sum_{j=1}^{n}\sum_{i>j}\left(1_{|z_{ij}|<\sqrt{5}\widehat{h}_{\mathrm{p}}}2\sqrt{2}-1_{|z_{ij}|<\sqrt{10}\widehat{h}_{\mathrm{p}}}\right)}\right)^{1/2}. \tag{49}$$

For $\widehat{h}_{\mathrm{p}}$ based on Silverman's rule but for the Epanechnikov case, we get $\widehat{h}_{\mathrm{p}}=1.05\widehat{\sigma}n^{-1/5}$, which is almost the same as the normal's.

Derivations for approximating UCV with the Epanechnikov kernel yield:

$$\widehat{h}_{\mathrm{a}}=\left(\frac{\frac{3}{5}\sum_{j=1}^{n}\sum_{i>j}\left(1_{|z_{ij}|<\sqrt{5}\widehat{h}_{\mathrm{p}}}4\sqrt{2}-1_{|z_{ij}|<\sqrt{10}\widehat{h}_{\mathrm{p}}}\right)z_{ij}^2}{n+2\sum_{j=1}^{n}\sum_{i>j}\left(1_{|z_{ij}|<\sqrt{5}\widehat{h}_{\mathrm{p}}}2\sqrt{2}-1_{|z_{ij}|<\sqrt{10}\widehat{h}_{\mathrm{p}}}\right)}\right)^{1/2}, \tag{50}$$

where n is replaced by 0 for the corresponding $\widehat{h}_{\mathrm{aa}}$.

BCV with Epanechnikov kernel. Derivations for approximating BCV with the Epanechnikov kernel yield the first-order condition

$$\widehat{h}^6=\frac{1}{2^{17/2}n}\sum_{j=1}^{n}\sum_{i>j}1_{|z_{ij}|<\sqrt{10}\widehat{h}}\left(7z_{ij}^6-110z_{ij}^4\widehat{h}^2+396z_{ij}^2\widehat{h}^4-120\widehat{h}^6\right), \tag{51}$$

where we have used $k_{21}^2=1$ and $k_{02}=3/5^{3/2}$. Terms can be collected as

$$\widehat{h}=\left(\frac{\sum_{j=1}^{n}\sum_{i>j}1_{|z_{ij}|<\sqrt{10}\widehat{h}}\left(7z_{ij}^4-110z_{ij}^2\widehat{h}^2+396\widehat{h}^4\right)z_{ij}^2}{8\left(2^{11/2}n+15\sum_{j=1}^{n}\sum_{i>j}1_{|z_{ij}|<\sqrt{10}\widehat{h}}\right)}\right)^{1/6}, \tag{52}$$

and the corresponding solution $\widehat{h}_{\rm a}$ is with $\widehat{h}_{\rm p}$ plugged into the RHS.

SCV with Epanechnikov kernel. Derivations for approximating SCV with the Epanechnikov kernel (with $k_{02} = 3/5^{3/2}$) yield:

$$\widehat{g}_{\rm aa} = \left(\frac{3}{10} \frac{z_n(\widehat{h}_{\rm p}, \widehat{g}_{\rm p}) - z_n(0, \widehat{g}_{\rm p})}{a_{-1} + \zeta_n(\widehat{h}_{\rm p}, \widehat{g}_{\rm p}) - \zeta_n(0, \widehat{g}_{\rm p})} \right)^{1/2}, \tag{53}$$

$$\widehat{h}_{\rm aa} = \left(\frac{3}{5} \frac{z_n(\widehat{h}_{\rm p}\sqrt{2}, \widehat{g}_{\rm p}) - z_n(\widehat{h}_{\rm p}, \widehat{g}_{\rm p})}{a_0 + \zeta_n(\widehat{h}_{\rm p}\sqrt{2}, \widehat{g}_{\rm p}) - \zeta_n(\widehat{h}_{\rm p}, \widehat{g}_{\rm p})} - 2\widehat{g}_{\rm aa}^2 \right)^{1/2}, \tag{54}$$

$$\widehat{g}_{\rm a} = \left(\frac{3}{10} \frac{z_n(\widehat{h}_{\rm p}, \widehat{g}_{\rm p}) - z_n(0, \widehat{g}_{\rm p})}{a_0 + \zeta_n(\widehat{h}_{\rm p}, \widehat{g}_{\rm p}) - \zeta_n(0, \widehat{g}_{\rm p})} \right)^{1/2}, \tag{55}$$

$$\widehat{h}_{\rm a} = \left(\frac{3}{10\left(1 + n^{1/5}\right)} \frac{z_n(\widehat{h}_{\rm p}\sqrt{2}, \widehat{g}_{\rm p}) - z_n(\widehat{h}_{\rm p}, \widehat{g}_{\rm p})}{a_1 + \zeta_n(\widehat{h}_{\rm p}\sqrt{2}, \widehat{g}_{\rm p}) - \zeta_n(\widehat{h}_{\rm p}, \widehat{g}_{\rm p})} \right)^{1/2}, \tag{56}$$

with $a_j = \left(j + 1 + 2n^{1/5}\right)^{3/2} n/5$, $\zeta_n(h, g) = \sum_{j=1}^{n} \sum_{i>j} 1_{|z_{ij}|<\sqrt{5}\sqrt{(h^2+2g^2)}}$, $z_n(h, g) = \sum_{j=1}^{n} \sum_{i>j} 1_{|z_{ij}|<\sqrt{5}\sqrt{(h^2+2g^2)}} z_{ij}^2$.

B.2 Epanechnikov versus Student kernels in finite samples

The asymptotically-best kernel for optimizing the AMISE is the Epanechnikov kernel which is exactly separable. The following experiment shows that this kernel is easily dominated in finite samples as soon as we depart from the Gaussian case. Within our Monte Carlo framework of Section E below, let us compute for each generating process: the ISE using an Epanechnikov kernel with $\widehat{h} = 1.05\,\widehat{\sigma}\,n^{-1/5}$, and the ISE using a Student kernel with $\widehat{h}_{\rm S}(\widehat{\nu})$ of (16) with $\widehat{\nu}$ obtained using our empirical rule of Table 6. Table 3 gives the Monte Carlo mean (and standard deviation) of the ratio of these two ISEs. The Epanechnikov kernel is the best kernel for estimating a Gaussian density. However, as soon as we consider more complex processes, the Student kernel becomes substantially more efficient and this gain persists even as n increases.

This result demonstrates three things. First, our simple sample-based procedure for choosing ν is efficient. Second, the improvements are increasing as soon as we depart from the Gaussian case, these departures being the most relevant cases when dealing with applications. Every try we made with the Epanechnikov kernel proved to be inefficient, either with UCV or SCV. Finally, and as a side-product, if the Student kernel with $\widehat{h}_{\rm S}(\widehat{\nu})$ of (16) manages to beat the Epanechnikov kernel in most cases, it means that $\widehat{h}_{\rm S}$ of (16) is a good initial plug-in to use as $\widehat{h}_{\rm p}$. In an unreported experiment, we show that the generalized Jones and Sheather rule $\widehat{h}_{\rm JS}$ of (76) does not bring in any improvement, confirming the result of Proposition 1 which implies that both give the same asymptotic outcome.

Table 3: Comparing relative efficiency of Epanechnikov and Student kernels

Sample size	150	450	1000	1500
Gaussian $\mathrm{N}(0,1)$	0.91 (0.12)	0.92 (0.08)	0.92 (0.07)	0.92 (0.07)
Bimodal mix $0.5\mathrm{N}(-1,4/9)+0.5\mathrm{N}(1,4/9)$	1.34 (0.34)	1.35 (0.28)	1.31 (0.22)	1.31 (0.22)
Skewed mix $0.75\mathrm{N}(0,1)+0.25\mathrm{N}(3/2,1/9)$	1.47 (0.37)	1.58 (0.35)	1.66 (0.36)	1.66 (0.29)
Student $\mathrm{t}(3)$	1.62 (1.28)	1.69 (1.01)	1.68 (0.99)	1.61 (0.66)
Lognormal $\mathrm{LN}(0,1)$	5.23 (2.04)	6.80 (2.26)	8.06 (2.18)	8.67 (2.19)

Figures represent the Monte Carlo mean of the ratio $\mathrm{ISE}_{\mathrm{Epan}}/\mathrm{ISE}_{\mathrm{Stud}}$ and its standard deviation in small numbers below it. The $\widehat{\nu}$ for the Student kernel was determined using our rule of thumb of Table 6 of Section E below. The bandwidths follow Silverman's rule, generalized for each kernel.

B.3 AMISE for Student kernel, and implied exchange-rate for other kernels

The following proposition works out the implication on AMISE of using a Student $\mathrm{t}(\nu)$ kernel.

Proposition 3 *Within the class of Student* $\mathrm{t}(\nu)$ *kernels, the AMISE is*

$$\left(\frac{4^{1-\nu}}{\nu^2\left(\nu-2\right)}\Gamma\left(\nu+\frac{1}{2}\right)^2\frac{\Gamma\left(\frac{1}{2}\nu+\frac{1}{2}\right)^2}{\Gamma\left(\frac{1}{2}\nu\right)^6}\right)^{2/5}\left(\frac{h^4}{4}I_2+\frac{1}{nh}\right), \tag{57}$$

whose leading term with respect to ν *is*

$$\left(\frac{\nu\left(1-\frac{3}{16\nu}\right)^4\left(1-\frac{1}{4\nu}\right)^2}{4\pi\left(\nu-2\right)}\right)^{2/5}\left(\frac{h^4}{4}I_2+\frac{1}{nh}\right).$$

Here, h *denotes the bandwidth of the canonical kernel of Marron and Nolan (1989).*

Proof 1 *This follows by using the canonical kernels of Marron and Nolan (1989), corrected by Abadir and Lawford (2004) for a typo, to write the AMISE as*

$$\left(k_{02}^2k_{21}\right)^{2/5}\left(\frac{h^4}{4}I_2+\frac{1}{nh}\right). \tag{58}$$

The term $\left(k_{02}^2k_{21}\right)^{2/5}$ *varies with the kernel, but the subsequent term does not. The result follows from* k_{02} *and* k_{21} *of Lemma 2.*

The plot of the first factor of (57) is given in Figure 3, since the second factor does not vary with ν. It shows that, for $\nu>4$, there is little loss of

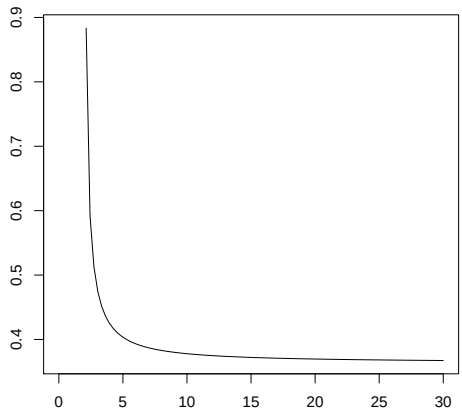

Figure 3: AMISE as a function of ν

relative efficiency from using any t(ν) including the normal. But this result is obtained under the assumption that the choice of ν does not affect the estimation of the unknown h in the second factor of (57), which is not the case, especially in finite samples. Furthermore, as shown in Hall and Marron (1987), minimizing the MISE leads to the deterministic bandwidth (2) containing an unknown I_2 whose estimation introduces an error which is dominated by minimizing the ISE instead. To assess the net effect of introducing a Student kernel on finite-sample performance, we needed to resort to the simulations that we present below.

Proposition 3 implies the analytic formula for an exchange-rate table that translates AMISE-optimal bandwidths from one kernel to another, which we have calculated. We have conducted unreported simulations that show that this exchange rate, however, does not extend to translating ISE-optimal h's, not even in large samples. This is another reason for us to require simulations to compare CV-optimal bandwidths.

C Multivariate and product kernels

We do not pursue BCV solutions here because they underperformed in the univariate case. We focus on the traditional UCV and the more modern SCV. In the bivariate application to the Michigan dataset, we also applied our formulae for product kernels and the results we got were only slightly different from those for our multivariate kernels. The reason is that we are dealing with the small dimension of $d = 2$; see the discussion of rates in Subsection C.2 below. Note that both methods use the orthonormalization procedure that we discuss before Theorem 2 in the main text, implying a bandwidth matrix $\boldsymbol{H}$ that is proportional to the sample variance matrix.

The derivations of Subsections C.1 and C.3 below contain the results for this Theorem 2, with the stated plug-ins obtained from Subsection 4.2 of the main text.

C.1 UCV, multivariate kernel

With $\boldsymbol{z}_{ij} = \boldsymbol{x}_j - \boldsymbol{x}_i$ (whose elements are denoted by $z_{ij,m}$), we now have

$$\begin{aligned} S_2 &= n^{-1}K_{h\sqrt{2}}(0) + 2n^{-2}\sum_{j=1}^{n}\sum_{i>j} K_{h\sqrt{2}}(\boldsymbol{z}_{ij}), \\ S_2 + S_3 &= \frac{K_{h\sqrt{2}}(0)}{n} + \frac{2+O(1/n)}{n^2}\sum_{j=1}^{n}\sum_{i>j}\left[K_{h\sqrt{2}}(\boldsymbol{z}_{ij}) - 2K_h(\boldsymbol{z}_{ij})\right] \end{aligned}$$

and we will drop $O(1/n)$ as before but keep the diagonal element $K_{h\sqrt{2}}(0)/n$. The scaled t(ν) kernel gives $K_{h\sqrt{2}}(0) = 2^{-d/2}c_{\nu,d}h^{-d}$ and the objective function (up to constant scale)

$$2^{-1-d/2}nh^{-d} + h^{\nu}\sum_{j=1}^{n}\sum_{i>j}\left[2^{\nu/2}\left(2h^2 + \frac{1}{\nu}\boldsymbol{z}_{ij}^{\top}\boldsymbol{z}_{ij}\right)^{-(\nu+d)/2} - 2\left(h^2 + \frac{1}{\nu}\boldsymbol{z}_{ij}^{\top}\boldsymbol{z}_{ij}\right)^{-(\nu+d)/2}\right]$$

and exact first-order condition

$$\begin{aligned} 2^{-1-d/2}nd &= \nu\widehat{h}_{\mathrm{u}}^{\nu+d}\left[2^{\nu/2}y_n(0;\widehat{h}_{\mathrm{u}}\sqrt{2}) - 2y_n(0;\widehat{h}_{\mathrm{u}})\right] \\ &\quad -2(\nu+d)\widehat{h}_{\mathrm{u}}^{\nu+d+2}\left[2^{\nu/2}y_n(1;\widehat{h}_{\mathrm{u}}\sqrt{2}) - y_n(1;\widehat{h}_{\mathrm{u}})\right], \end{aligned} \tag{59}$$

where $y_n(q;h) = \sum_{j=1}^{n}\sum_{i>j}(h^2 + \frac{1}{\nu}\boldsymbol{z}_{ij}^{\top}\boldsymbol{z}_{ij})^{-q-(\nu+d)/2}$. As before, the same asymptotic invariance applies to $y_n(q;h)$, but with $\widehat{h}_{\mathrm{u}} = O_p(n^{-1/(4+d)})$ now, allowing us to use $\widehat{h}_{\mathrm{p}}$ in $y_n(q;h)$ in (59) whose form is now $\alpha_1 = \alpha_2\widehat{h}^{\nu+d} + \alpha_3\widehat{h}^{\nu+d+2}$ and having the explicit asymptotic solution

$$\widehat{h}_{\mathrm{a}} = (\alpha_1/(\alpha_2 + \alpha_3\widehat{h}_{\mathrm{p}}^2))^{1/(\nu+d)} \tag{60}$$

and its asymptotic approximation is $\widehat{h}_{\mathrm{aa}} = \sqrt{(-\alpha_2/\alpha_3)}$.

C.2 UCV, product kernel

Here we consider the case of multiplicative marginal t(ν) kernels. Sain et al. (1994) show that this case of product kernels with common h for all dimensions, $K_h(\boldsymbol{t}) = h^{-d}\prod_{m=1}^{d}K(h^{-1}t_m)$, leads to faster convergence in n when the dimension d is larger: compared to the AMISE-minimizing h_0, they show that $\widehat{h}_{\mathrm{u}} - h_0 = O_p(n^{-(2+d)/(8+2d)})$ and $\widehat{h}_{\mathrm{u}}/h_0 = 1 + O_p(n^{-d/(8+2d)})$, both rates improving and approaching $n^{-1/2}$ for larger d.

We now have the objective function (up to constant scale)

$$2^{-1-d/2}nh^{-d}+h^{\nu d}\sum_{j=1}^{n}\sum_{i>j}\left[2^{\nu d/2}\prod_{m=1}^{d}\left(2h^2+z_{ij,m}^2/\nu\right)^{-(\nu+1)/2}-2\prod_{m=1}^{d}\left(h^2+z_{ij,m}^2/\nu\right)^{-(\nu+1)/2}\right]$$

and first-order condition

$$\begin{aligned}2^{-1-d/2}nd &= \nu\widehat{h}_{\mathrm{u}}^{(\nu+1)d}\left[2^{\nu d/2}y_n(0;\widehat{h}_{\mathrm{u}}\sqrt{2})-2y_n(0;\widehat{h}_{\mathrm{u}})\right] \\ &\quad -2(\nu+1)\widehat{h}_{\mathrm{u}}^{(\nu+1)d+2}\left[2^{\nu d/2}y_n(1;\widehat{h}_{\mathrm{u}}\sqrt{2})-y_n(1;\widehat{h}_{\mathrm{u}})\right],\end{aligned} \tag{61}$$

where $y_n(q;h)=\sum_{j=1}^{n}\sum_{i>j}\sum_{m=1}^{d}\left(h^2+z_{ij,m}^2/\nu\right)^{-q}/\prod_{m=1}^{d}\left(h^2+z_{ij,m}^2/\nu\right)^{(\nu+1)/2}$. The asymptotic invariance of $y_n(q;h)$ allows us to use $\widehat{h}_{\mathrm{p}}$ in these, to get an explicit asymptotic solution in (61) whose form is now $\alpha_1=\alpha_2\widehat{h}^{(\nu+1)d}+\alpha_3\widehat{h}^{(\nu+1)d+2}$. We get

$$\widehat{h}_{\mathrm{a}}=(\alpha_1/(\alpha_2+\alpha_3\widehat{h}_{\mathrm{p}}^2))^{1/(\nu+1)d} \tag{62}$$

and $\widehat{h}_{\mathrm{aa}}=\sqrt{(-\alpha_2/\alpha_3)}$.

Notice the power of $\widehat{h}_{\mathrm{a}}$ in the cases of multivariate kernel versus product kernel: $1/(\nu+d)$ vs $1/(\nu+1)d$, the latter leading to a fraction closer to zero when d is large. Although the $y_n(\cdot)$ functions (and the alphas) are not the same, they are of comparable orders of magnitudes as they are effectively related to the value of the density at a given point.

C.3 SCV, multivariate kernel

As before, but with

$$k_{02,d}=\left(\frac{\nu}{2\nu+d}\right)^{d/2}\frac{\left((\pi\nu)^{-d/2}\,\Gamma\left(\frac{\nu+d}{2}\right)/\Gamma\left(\frac{\nu}{2}\right)\right)^2}{\left(\pi\,(2\nu+d)\right)^{-d/2}\,\Gamma\left(\nu+d\right)/\Gamma\left(\nu+\frac{d}{2}\right)}$$

of Lemma 2(ii),

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

where $y_n(q;h,g)=\sum_{j=1}^{n}\sum_{i>j}(h^2+2g^2+\frac{1}{\nu}\boldsymbol{z}_{ij}^{\top}\boldsymbol{z}_{ij})^{-q-(\nu+d)/2}$. Since $\partial y_n(0;h\sqrt{2},g)/\partial h=(h/g)\,\partial y_n(0;h\sqrt{2},g)/\partial g$ and

$$\frac{\partial y_n(0;h,g)}{\partial g}=\frac{2g}{h}\frac{\partial y_n(0;h,g)}{\partial h}=-2\,(\nu+d)\,g\;y_n(1;h,g),$$

defining $y_n^\dagger(q;h,g) = \left(h^2+2g^2\right)^{q-1+\nu/2} y_n(q;h,g)$ allows us to write the exact first-order conditions for g and h, respectively, as

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

and

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

Simplifying the former by subtracting the latter,

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

We now consider joint solutions of the last two equations.

As before, replacing $y_n^\dagger(q;a\widehat{h}_{\rm s},\widehat{g}_{\rm s})$ by $(a^2\widehat{h}_{\rm s}^2+2\widehat{g}_{\rm s}^2)^{q-1+\nu/2}y_n(q;a\widehat{h}_{\rm p},\widehat{g}_{\rm p})$, the RHS of the last equation gives

$$\widehat{g}_{\rm aa}=\left(\frac{y_n(0;\widehat{h}_{\rm p},\widehat{g}_{\rm p})-y_n(0;0,\widehat{g}_{\rm p})}{2\left(1+\frac{d}{\nu}\right)\left[y_n(1;\widehat{h}_{\rm p},\widehat{g}_{\rm p})-y_n(1;0,\widehat{g}_{\rm p})\right]}\right)^{1/2} \tag{63}$$

and the one before it (without LHS)

$$\widehat{h}_{\rm aa}=\left(\frac{y_n(0;\widehat{h}_{\rm p}\sqrt{2},\widehat{g}_{\rm p})-y_n(0;\widehat{h}_{\rm p},\widehat{g}_{\rm p})}{\left(1+\frac{d}{\nu}\right)\left[y_n(1;\widehat{h}_{\rm p}\sqrt{2},\widehat{g}_{\rm p})-y_n(1;\widehat{h}_{\rm p},\widehat{g}_{\rm p})\right]}-2\widehat{g}_{\rm aa}^2\right)^{1/2} \tag{64}$$

and (keeping its LHS)

$$\begin{aligned}
&\frac{k_{02,d}dn}{4c_{\nu,d}}+\frac{\delta dn}{2}\left((2+2n^{1/5})^{-1-d/2}-(1+2n^{1/5})^{-1-d/2}\right) \qquad (65)\\
= \;& \nu\widehat{h}^{\nu+d}\left[(2+2n^{1/5})^{(\nu-2)/2}y_n(0;\widehat{h}_{\rm p}\sqrt{2},\widehat{g}_{\rm p})-(1+2n^{1/5})^{(\nu-2)/2}y_n(0;\widehat{h}_{\rm p},\widehat{g}_{\rm p})\right]\\
& -(\nu+d)\,\widehat{h}^{\nu+d+2}\left[(2+2n^{1/5})^{\nu/2}y_n(1;\widehat{h}_{\rm p}\sqrt{2},\widehat{g}_{\rm p})-(1+2n^{1/5})^{\nu/2}y_n(1;\widehat{h}_{\rm p},\widehat{g}_{\rm p})\right]
\end{aligned}$$

which is a polynomial of the form $\alpha_1=\alpha_2\widehat{h}^{\nu+d}+\alpha_3\widehat{h}^{\nu+d+2}$ yielding as before

$$\widehat{h}_{\rm a}=(\alpha_1/(\alpha_2+\alpha_3\widehat{h}_{\rm p}^2))^{1/(\nu+d)} \tag{66}$$

if we use the plug-in $\widehat{h}_{\rm p}$ on the RHS.

C.4 SCV, product kernel

We now have

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

where $y_n(q;h,g) = \sum_{j=1}^n \sum_{i>j} \sum_{m=1}^d (h^2 + 2g^2 + z_{ij,m}^2/\nu)^{-q} / \prod_{m=1}^d (h^2 + 2g^2 + z_{ij,m}^2/\nu)^{(\nu+1)/2}$. Since $\partial y_n(0;h\sqrt{2},g)/\partial h = (h/g)\,\partial y_n(0;h\sqrt{2},g)/\partial g$ and

$$\frac{\partial y_n(0;h,g)}{\partial g} = \frac{2g}{h}\frac{\partial y_n(0;h,g)}{\partial h} = -2d\,(\nu+1)\,g\; y_n(1;h,g),$$

defining $y_n^\dagger(q;h,g) = \left(h^2+2g^2\right)^{q-1+\nu d/2} y_n(q;h,g)$ allows us to write the exact first-order conditions for g and h, respectively, as

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

and

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

Simplifying the former by subtracting the latter,

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

We now consider joint solutions of the last two equations.

As before, replacing $y_n^\dagger(q;a\widehat{h}_{\rm s},\widehat{g}_{\rm s}) = (a^2\widehat{h}_{\rm s}^2 + 2\widehat{g}_{\rm s}^2)^{q-1+\nu d/2} y_n(q;a\widehat{h}_{\rm p},\widehat{g}_{\rm p})$ the RHS of the last equation gives

$$\widehat{g}_{\rm aa} = \left(\frac{y_n(0;\widehat{h}_{\rm p},\widehat{g}_{\rm p}) - y_n(0;0,\widehat{g}_{\rm p})}{2\left(1+\frac{1}{\nu}\right)\left[y_n(1;\widehat{h}_{\rm p},\widehat{g}_{\rm p}) - y_n(1;0,\widehat{g}_{\rm p})\right]}\right)^{1/2} \tag{67}$$

and the one before it (without LHS)

$$\widehat{h}_{\rm aa} = \left(\frac{y_n(0;\widehat{h}_{\rm p}\sqrt{2},\widehat{g}_{\rm p}) - y_n(0;\widehat{h}_{\rm p},\widehat{g}_{\rm p})}{\left(1+\frac{1}{\nu}\right)\left[y_n(1;\widehat{h}_{\rm p}\sqrt{2},\widehat{g}_{\rm p}) - y_n(1;\widehat{h}_{\rm p},\widehat{g}_{\rm p})\right]} - 2\widehat{g}_{\rm aa}^2\right)^{1/2} \tag{68}$$

and (keeping its LHS)

$$\begin{aligned} & \frac{k_{02}^d dn}{4c_\nu^d} + \frac{\delta dn}{2}\left((2+2n^{1/5})^{-1-d/2} - (1+2n^{1/5})^{-1-d/2}\right) \qquad (69) \\ = \;& \nu\widehat{h}^{(\nu+1)d}\left[(2+2n^{1/5})^{-1+\nu d/2} y_n(0;\widehat{h}_{\mathrm{p}}\sqrt{2},\widehat{g}_{\mathrm{p}}) - (1+2n^{1/5})^{-1+\nu d/2} y_n(0;\widehat{h}_{\mathrm{p}},\widehat{g}_{\mathrm{p}})\right] \\ & -(\nu+1)\,\widehat{h}^{(\nu+1)d+2}\left[(2+2n^{1/5})^{\nu d/2} y_n(1;\widehat{h}_{\mathrm{p}}\sqrt{2},\widehat{g}_{\mathrm{p}}) - (1+2n^{1/5})^{\nu d/2} y_n(1;\widehat{h}_{\mathrm{p}},\widehat{g}_{\mathrm{p}})\right] \end{aligned}$$

which is a polynomial of the form $\alpha_1 = \alpha_2\widehat{h}^{(\nu+1)d} + \alpha_3\widehat{h}^{(\nu+1)d+2}$ yielding as before

$$\widehat{h}_{\mathrm{a}} = (\alpha_1/(\alpha_2 + \alpha_3\widehat{h}_{\mathrm{p}}^2))^{1/(\nu+1)d} \qquad (70)$$

if we use the plug-in $\widehat{h}_{\mathrm{p}}$ on the RHS.

C.5 Multivariate plug-in $\widehat{h}_{\mathrm{p}}$ for product kernels

In the case of the product kernel, replace $k_{02,d}$ in (27) by

$$k_{02}^d = \left(\frac{\sqrt{2}\Gamma\left(\frac{\nu}{2}+\frac{1}{2}\right)\Gamma\left(\frac{\nu}{2}+\frac{1}{4}\right)\Gamma\left(\frac{\nu}{2}+\frac{3}{4}\right)}{\sqrt{\pi}\nu^{\frac{3}{2}}\left(\Gamma\left(\frac{\nu}{2}\right)\right)^3}\right)^d \sim \left(\frac{\left(1-\frac{3}{16\nu}\right)^2\left(1-\frac{1}{4\nu}\right)}{2\sqrt{\pi}}\right)^d$$

from Lemma 2.

D Further proofs

D.1 Proof of Lemma 1

Proof 2 *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,*

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

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 $He_0 = 1$,

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

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.

D.2 Proof of Proposition 1

Proof 3 *Differentiating (13) with respect to h, we get the first-order condition*

$$\begin{aligned} &\nu\sum_{j=1}^{n}\sum_{i>j}\left[2^{\nu/2}(2\widehat{h}^2 + z_{ij}^2/\nu)^{-(\nu+1)/2} - 2(\widehat{h}^2 + z_{ij}^2/\nu)^{-(\nu+1)/2}\right] \\ = \;& 2(\nu+1)\,\widehat{h}^2\sum_{j=1}^{n}\sum_{i>j}\left[2^{\nu/2}(2\widehat{h}^2 + z_{ij}^2/\nu)^{-(\nu+3)/2} - (\widehat{h}^2 + z_{ij}^2/\nu)^{-(\nu+3)/2}\right] \end{aligned}$$

or

$$\nu\left[2^{\nu/2}y_n(0;\widehat{h}\sqrt{2}) - 2y_n(0;\widehat{h})\right] = 2(\nu+1)\,\widehat{h}^2\left[2^{\nu/2}y_n(1;\widehat{h}\sqrt{2}) - y_n(1;\widehat{h})\right]. \tag{71}$$

Since $R = O_p(n^2)$ and $\widehat{h} = O_p(n^{-1/5})$, the leading term of each double sum defining a $y_n(\cdot;\cdot)$ is the one not containing $\widehat{h}$, and the asymptotic solution is unaffected by using $\widehat{h}_{\mathrm{p}}$ in $y_n(q;\widehat{h}_{\mathrm{p}})$.

D.3 Lemmas for generalized plug-in bandwidths

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*

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

where $K(t) = c_\nu/(1+t^2/\nu)^{(\nu+1)/2}$ *with* $c_\nu = \Gamma(\frac{\nu+1}{2})/(\sqrt{(\pi\nu)}\Gamma(\frac{\nu}{2}))$, *and* $k_{02} \sim a(\nu)$ *means 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 multivariate Student $\mathrm{t}(\nu)$ *kernel,*

$$\begin{aligned} k_{02,d} &= \int_{\mathbb{R}^d} c_{\nu,d}^2 \left(1+\frac{1}{\nu}\boldsymbol{t}^\top\boldsymbol{t}\right)^{-\nu-d}\mathrm{d}\boldsymbol{t} \\ &= \left(\frac{\nu}{2\nu+d}\right)^{d/2} \frac{\left((\pi\nu)^{-d/2}\,\Gamma\left(\frac{\nu+d}{2}\right)/\Gamma\left(\frac{\nu}{2}\right)\right)^2}{\left(\pi\,(2\nu+d)\right)^{-d/2}\Gamma\left(\nu+d\right)/\Gamma\left(\nu+\frac{d}{2}\right)} \sim (4\pi)^{-d/2}\,\frac{\left(1+\frac{d(d-2)}{4\nu}\right)^2}{\left(1+\frac{d(3d-2)}{8\nu}\right)}, \end{aligned}$$

where $c_{\nu,d} = (\pi\nu)^{-d/2}\,\Gamma(\frac{\nu+d}{2})/\Gamma(\frac{\nu}{2})$ *generalizes the univariate* $c_\nu = c_{\nu,1}$.
(iii) For a scaled Student $\mathrm{t}(\nu)$ *density with variance* σ^2,

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

(iv) For a scaled multivariate Student $\mathrm{t}(\nu)$ *density with variance* $\sigma^2\boldsymbol{I}$,

$$\begin{aligned} I_2 &= \int_{\mathbb{R}^d} (\sum_{j=1}^{d} \partial^2 f(\boldsymbol{u})/\partial u_j^2)^2\,\mathrm{d}\boldsymbol{u} = \frac{d\,(2+d)}{2^{\nu+d+1}\pi^{(d-1)/2}b^{d+4}\nu^{2+d/2}}\frac{\Gamma\left(\frac{\nu+d}{2}+2\right)\Gamma\left(\nu+\frac{d}{2}+2\right)}{\left(\Gamma\left(\frac{\nu}{2}\right)\right)^2\Gamma\left(\frac{\nu+d+5}{2}\right)} \\ &\sim \frac{d\,(2+d)}{2^{d+2}\pi^{d/2}b^{d+4}}\left(1+\frac{(d+4)\,(d+2)}{4\nu}\right)\left(1+\frac{d\,(d+4)}{16\nu}\right)\left(1-\frac{(d+4)\,(3d+12)}{16\nu}\right), \end{aligned}$$

where we have the scale factor $b = \sigma\sqrt{(1-2/\nu)}$.

Proof 4 *(i) For* k_{21}, *the result is simply the usual variance of a t(*ν*). For* k_{02}, *the integrating constant* $c_{2\nu+1}$ *of the* $t(2\nu+1)$ *density implies that*

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

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 subsequent asymptotic expansion is due to the general approximation for the ratio of two gamma functions

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

e.g., see pp.358,368,711 of Abadir et al. (2018).
(ii) The scaled multivariate $t(\nu)$ kernel is

$$K_h(\boldsymbol{t}) = c_{\nu,d}\left|\boldsymbol{H}\right|^{-1/2}\left(1+\frac{1}{\nu}\boldsymbol{t}^\top\boldsymbol{H}^{-1}\boldsymbol{t}\right)^{-(\nu+d)/2} = c_{\nu,d}h^\nu\left(h^2+\frac{1}{\nu}\sum_{m=1}^{d}t_m^2\right)^{-(\nu+d)/2}.$$

Then,

$$\begin{aligned}k_{02,d} &= \left(\frac{\nu}{2\nu+d}\right)^{d/2}\frac{c_{\nu,d}^2}{c_{2\nu+d,d}}\int_{\mathbb{R}^d}c_{2\nu+d,d}\left(1+\frac{1}{2\nu+d}\boldsymbol{t}^\top\boldsymbol{t}\right)^{-\nu-d}\mathrm{d}\boldsymbol{t}\\ &= \left(\frac{\nu}{2\nu+d}\right)^{d/2}\frac{c_{\nu,d}^2}{c_{2\nu+d,d}}\\ &= \left(\frac{\nu}{2\nu+d}\right)^{d/2}\frac{\left((\pi\nu)^{-d/2}\Gamma\left(\frac{\nu+d}{2}\right)/\Gamma\left(\frac{\nu}{2}\right)\right)^2}{\left(\pi\left(2\nu+d\right)\right)^{-d/2}\Gamma\left(\nu+d\right)/\Gamma\left(\nu+\frac{d}{2}\right)} \sim (4\pi)^{-d/2}\frac{\left(1+\frac{d(d-2)}{4\nu}\right)^2}{\left(1+\frac{d(3d-2)}{8\nu}\right)}.\end{aligned}$$

(iii) The Student $t(\nu)$ 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

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

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

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

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

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

Using

$$c_\nu = \frac{\Gamma\left(\frac{\nu+1}{2}\right)}{\sqrt{(\pi\nu)}\Gamma\left(\frac{\nu}{2}\right)} \sim \frac{1-\frac{1}{4\nu}}{\sqrt{(2\pi)}} \tag{72}$$

and

$$c_{2\nu+9} = \frac{\Gamma(\nu+5)}{\sqrt{\pi}\sqrt{(2\nu+9)}\Gamma\left(\nu+\frac{9}{2}\right)} \sim \frac{1+\frac{17}{8\nu}}{\sqrt{\pi}\sqrt{\left(2+\frac{9}{\nu}\right)}}$$

gives the required asymptotic result.

(iv) Letting $b = \sigma\sqrt{(1-2/\nu)}$, the Student t density with variance $\sigma^2\boldsymbol{I}$ (for $\nu > 2$) is

$$f(\boldsymbol{u}) = \frac{c_{\nu,d}}{b^d}\left(1 + \frac{1}{\nu b^2}\sum_{m=1}^d u_m^2\right)^{-(\nu+d)/2}$$

(where the components are uncorrelated but mutually dependent) and

$$\frac{\partial^2 f(\boldsymbol{u})}{\partial u_j^2} = -\frac{(\nu+d)\left(\nu b^2 + \sum_{m=1}^d u_m^2 - (\nu+2+d)u_j^2\right)}{\nu^2 b^4 \left(1 + \sum_{m=1}^d u_m^2/(\nu b^2)\right)^2} f(\boldsymbol{u}),$$

hence

$$\begin{aligned} I_2 &= \frac{(\nu+d)^2(\nu+2)^2 c_{\nu,d}^2}{\nu^2 b^{2d+4}} \int_{\mathbb{R}^d} \left(\frac{d}{\nu+2} - \frac{1}{\nu b^2}\sum_{m=1}^d u_m^2\right)^2 \left(1 + \frac{1}{\nu b^2}\sum_{m=1}^d u_m^2\right)^{-\nu-d-4} \mathrm{d}\boldsymbol{u} \\ &= \frac{(\nu+d)^2(\nu+2)^2 c_{\nu,d}^2}{\nu^{2-d/2} b^{d+4}} \sum_{j=0}^2 \binom{2}{j}\left(-1-\frac{d}{\nu+2}\right)^j \int_{\mathbb{R}^d} \left(1 + \frac{1}{\nu b^2}\sum_{m=1}^d u_m^2\right)^{-j-\nu-d-2} \frac{\mathrm{d}\boldsymbol{u}}{(\nu b^2)^{d/2}} \\ &= \frac{(\nu+d)^2(\nu+2)^2 c_{\nu,d}^2}{\nu^{2-d/2} b^{d+4}} \sum_{j=0}^2 \binom{2}{j}\left(-1-\frac{d}{\nu+2}\right)^j \frac{c_{2(j+\nu)+d+4,d}^{-1}}{(2(j+\nu)+d+4)^{d/2}} \end{aligned}$$

by the integral of the multivariate Student $t(2(j+\nu)+d+4)$ and a change of scale $b\sqrt{\nu} \leftrightarrow b\sqrt{(2(j+\nu)+d+4)}$. By

$$\frac{c_{\nu,d}^2 c_{2(j+\nu)+d+4,d}^{-1}}{(2(j+\nu)+d+4)^{d/2}} = (\pi\nu)^{-d}\left(\frac{\Gamma\left(\frac{\nu+d}{2}\right)}{\Gamma\left(\frac{\nu}{2}\right)}\right)^2 \frac{\pi^{d/2}\Gamma\left(j+\nu+\frac{d}{2}+2\right)}{\Gamma(j+\nu+d+2)}$$

and the recurrence relation $\Gamma(a+1) = a\Gamma(a)$ *giving*

$$\sum_{j=0}^{2} \binom{2}{j} \frac{\Gamma\left(j+\nu+\frac{d}{2}+2\right)}{\Gamma\left(j+\nu+d+2\right)} \left(-1-\frac{d}{\nu+2}\right)^j = \frac{d\left(2+d\right)}{4} \frac{\left(\nu+d+2\right)^2}{\left(\nu+2\right)^2} \frac{\Gamma\left(\nu+\frac{d}{2}+2\right)}{\Gamma\left(\nu+d+4\right)},$$

we get

$$\begin{aligned} I_2 &= \frac{4d\left(2+d\right)}{\pi^{d/2} b^{d+4} \nu^{2+d/2}} \left(\frac{\Gamma\left(\frac{\nu+d}{2}+2\right)}{\Gamma\left(\frac{\nu}{2}\right)}\right)^2 \frac{\Gamma\left(\nu+\frac{d}{2}+2\right)}{\Gamma\left(\nu+d+4\right)} \qquad (73) \\ &= \frac{d\left(2+d\right)}{2^{\nu+d+1} \pi^{(d-1)/2} b^{d+4} \nu^{2+d/2}} \frac{\Gamma\left(\frac{\nu+d}{2}+2\right) \Gamma\left(\nu+\frac{d}{2}+2\right)}{\left(\Gamma\left(\frac{\nu}{2}\right)\right)^2 \Gamma\left(\frac{\nu+d+5}{2}\right)} \end{aligned}$$

by Legendre's duplication formula as on pp.358,368 of Abadir et al. (2018). If we use p.711 of Abadir et al. (2018) on the first line of (73), then $\nu \to \infty$ *gives*

$$I_2 \sim \frac{d\left(2+d\right)}{2^{d+2} \pi^{d/2} b^{d+4}} \left(1+\frac{d}{\nu}\right)^2 \left(1+\frac{d+2}{\nu}\right)^2 \left(1+\frac{d\left(d-2\right)}{4\nu}\right)^2 \left(1-\frac{\left(d+4\right)\left(3d+10\right)}{4\nu}\right),$$

while if we apply Legendre's duplication formula on the second line of (73) before expanding for $\nu \to \infty$*, we get*

$$I_2 \sim \frac{d\left(2+d\right)}{2^{d+2} \pi^{d/2} b^{d+4}} \left(1+\frac{\left(d+4\right)\left(d+2\right)}{4\nu}\right) \left(1+\frac{d\left(d+4\right)}{16\nu}\right) \left(1-\frac{\left(d+4\right)\left(3d+12\right)}{16\nu}\right);$$

the two forms being equivalent if we expand to the same degree in $1/\nu$*, and both giving*

$$I_2 \to \frac{d\left(2+d\right)}{2^{d+2} \pi^{d/2} \sigma^{d+4}}$$

in the limit.

The next lemma allows us to generalize the popular plug-in method of Jones and Sheather (1991) for univariate densities, now using a Student (rather than Gaussian) kernel and plug-in density.

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

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

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

Proof 5 *(i) This follows directly from* $K(t) = c_\nu/(1+t^2/\nu)^{(\nu+1)/2}$ *and (72).*
(ii) From the Student $t(\nu)$ *density with variance* σ^2 *(see Lemma 2(iii)),*

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

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

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

From the Student $t(2\nu+13)$ *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 (72) for c_ν *and*

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

gives the required asymptotic result.

This lemma provides us with the ingredients for the estimate $\widehat{I_2} = n^{-2}\widehat{\lambda}^{-5}\sum_{i,j} K^{(4)}(z_{ij}/\widehat{\lambda})$ as

$$\widehat{I_2} = \frac{(4\nu-1)\,(\nu+1)\,(\nu+3)}{4\sqrt{(2\pi)}n^2\widehat{\lambda}^5\nu^5} \sum_{i,j} \frac{(\nu+2)\,(\nu+4)\,z_{ij}^4/\widehat{\lambda}^4 - 6\nu\,(\nu+4)\,z_{ij}^2/\widehat{\lambda}^2 + 3\nu^2}{(1+z_{ij}^2/(\widehat{\lambda}^2\nu))^{(\nu+9)/2}} \tag{74}$$

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

$$\widehat{\lambda} = \left(\frac{\sqrt{2}\left(\nu-2\right)^{9/2}\left(2\nu+7\right)\left(2\nu+9\right)\left(2\nu+11\right)\left(8\nu+25\right)}{5\nu^{7/2}\left(\nu+1\right)\left(\nu+3\right)\left(\nu+5\right)^2\left(4\nu-1\right)}\right)^{1/7}\widehat{\sigma}n^{-1/7} \quad (75)$$

leading to

$$\widehat{h}_{\mathrm{JS}} = \left(\frac{\left(\nu-2\right)^2\left(16\nu-3\right)^2\left(4\nu-1\right)}{\sqrt{\pi}2^{11}\nu^5\widehat{I}_2}\right)^{1/5} n^{-1/5}. \quad (76)$$

E Investigating finite-sample performance

We study the small sample performance of our UCV formulae, then the ones for SCV. We do not report the results for BCV, as they are dominated by the other methods in the case of non-Gaussian densities, while improving only slightly over the usual UCV in the Gaussian case. We propose an empirical method to determine an optimal choice for the degrees of freedom ν of the Student kernel, based on a measure of complexity of the data's distribution.

E.1 General simulation framework

We have selected five generating processes: Gaussian, symmetric bimodal and asymmetric mixtures of normals, Student's t(3), and lognormal in order to cover various degrees of complexity. We report these in Table 4, together

Table 4: Generating processes and a measure of their intrinsic complexity

n		150	450	1000
Gaussian	$\mathrm{N}(0,1)$	0.065	0.039	0.026
Bimodal mix	$0.5\mathrm{N}(-1,4/9)+0.5\mathrm{N}(1,4/9)$	0.105	0.086	0.077
Skewed mix	$0.75\mathrm{N}(0,1)+0.25\mathrm{N}(3/2,1/9)$	0.105	0.090	0.083
Student	t(3)	0.146	0.127	0.115
Lognormal	$\mathrm{LN}(0,1)$	0.266	0.256	0.252

We report the 0.9 quantile of $d_n(\boldsymbol{x})$ obtained with 1,500 simulations.

with a statistical measure of complexity based on the Kolmogorov-Smirnov statistics for departing from normality, $d_n(\boldsymbol{x}) = \max_{\tilde{x}_i}|\widehat{F}_n(\tilde{x}_i) - \Phi(\tilde{x}_i)|$, where $\tilde{x}_i$ are the standardized values of the observed sample x_i, and $\widehat{F}_n(.)$ is the empirical distribution function.[1]

[1]There exists various measures of complexity in the literature, mainly based on I_2 or I_4. They are simple to evaluate analytically, but are much more difficult to estimate empirically as they rely on the delicate choice of a bandwidth. The Kolmogorov-Smirnov statistic measures departure from normality without requiring the choice of a bandwidth and is used here as a measure of complexity, as the Gaussian density is the simplest to estimate. In the Gaussian case $\sqrt{n}d_n(\boldsymbol{x})$ converges in distribution and $d_n(\boldsymbol{x}) \xrightarrow{p} 0$, while otherwise $d_n(\boldsymbol{x})$ converges to a positive constant summarizing our degree of complexity.

In order to study the performance of our different suggested methods, we first compute an optimal ISE for each process that corresponds to the minimum integrated square error when using a Gaussian-kernel density estimator, knowing the true density as done in Faraway and Jhun (1990) for instance. This is the best ISE that can be reached when using a Kernel density estimation. We then compute the ISE of each method and divide it by this optimal ISE_{opt}. Results displayed in tables are thus scale free. For each experiment, we generate 1,500 replications.[2] All computations are done in `R`.

E.2 UCV and choice of degree of freedom ν

The optimal ν is found for each process by minimizing the true ISE when f is estimated using a Student kernel, where h is determined either using the generalized Silverman rule $\widehat{h}_{\text{S}}$ of (16) or according to the UCV approximation (31). From Table 5, we see that the optimal ν decreases as complexity

Table 5: Optimal median value for $\widehat{\nu}$

	$\widehat{h}_{\text{S}}$ of (16)			$\widehat{h}_{\text{a}}$ of (31)		
	150	450	1000	150	450	1000
Gaussian	30.0	30.0	30.0	30.0	30.0	30.0
Bimodal mix	5.7	4.9	4.7	8.4	7.2	6.8
Skewed mix	4.5	3.9	3.6	6.1	5.1	4.8
Student t(3)	4.6	4.1	3.9	7.6	6.1	6.1
Lognormal	2.5	2.5	2.5	2.6	2.6	2.7

increases. For a Gaussian process we found an optimal $\nu = 30$, while for a lognormal process it was $\nu = 2.5$. For the other processes, the median optimal value could be between $\nu = 4$ and $\nu = 6$, depending on the method chosen and the sample size.

But of course, we do not know the true process in real life. We just know that Gaussian samples are characterized by a measured complexity not exceeding 0.065, while for lognormal samples it always exceeds 0.15. Other processes are characterized by in-between measured complexity. So we can decide to select ν according to the measured complexity, following the rule of thumb given in Table 6. We could have designed a response surface, but it would have introduced more variability.

Equipped with this rule, let us measure the performance of $\widehat{h}_{\text{a}}$ of (31) with $\widehat{h}_{\text{S}}$ as plug-in, and compare it to the usual UCV as proposed in standard packages using a Gaussian kernel, which we denote by $\widehat{h}_{\text{ucv}}$. Results are in

[2] In order to reduce the variance of the Monte Carlo experiments, first we impose the same starting seed for each experiment, and second we take the smaller samples (150, 450) as sub-samples of the largest sample of size 1,000.

Table 6: Selection rule for ν

$d_n(\boldsymbol{x})$	$\widehat{\nu}$
$d_n \leq 0.06$	30
$0.06 < d_n \leq 0.15$	4 or 6
$d_n > 0.15$	2.5

Table 7: UCV with Student kernel, and standard UCV with Gaussian kernel: ratio of their ISEs to optimal $\mathrm{ISE}_{\mathrm{opt}}$

	$\widehat{h}_{\mathrm{a}}(\widehat{\nu})$			Gaussian $\widehat{h}_{\mathrm{ucv}}$		
	150	450	1000	150	450	1000
Gaussian	*1.48* (1.26)	*1.26* (0.64)	*1.16* (0.30)	2.09 (3.25)	1.79 (2.71)	1.49 (1.08)
Bimodal mix	*1.22* (0.26)	*1.23* (0.24)	*1.20* (0.18)	1.60 (1.32)	1.47 (0.95)	1.37 (0.69)
Skewed mix	*1.17* (0.24)	*1.18* (0.17)	*1.18* (0.15)	1.44 (0.75)	1.32 (0.65)	1.28 (0.54)
Student t(3)	*1.59* (0.80)	*1.34* (0.55)	*1.27* (0.26)	2.01 (2.96)	1.86 (2.71)	1.98 (2.99)
Lognormal	*1.24* (0.26)	1.22 (0.19)	1.20 (0.13)	1.33 (0.79)	*1.21* (0.40)	*1.14* (0.21)

We choose $\widehat{\nu}$ according to the rule of Table 6 that is based on measured complexity. We use $\widehat{h}_{\mathrm{S}}$ as a plug-in for $\widehat{h}_{\mathrm{a}}$. The $\widehat{h}_{\mathrm{ucv}}$ corresponds to the standard command `bw.ucv` of `R`. Small numbers between parentheses are Monte Carlo standard deviations. Figures in italics indicate best results.

Table 7. They show that our $\widehat{h}_{\mathrm{a}}$ is much better than the traditional UCV with a Gaussian kernel, and its variability is much lower. This advantage tends to decrease with the sample size. In small samples, the difference can be quite large. A substantial advantage of $\widehat{h}_{\mathrm{a}}$ is that it does not require iterations that would increase variability.

E.3 SCV

SCV was designed to improve over other CV methods. The simulations of Jones et al. (1991) show that SCV should gives better convergence results when not far from the Gaussian but, when far from it, UCV should give better results despite its variability. This is also the case when we compare the last 3-column block of Table 7 with the same in Table 8. They show that, with a Gaussian kernel, the `R` implementation of the SCV of Jones et al. (1991) written by Tarn Duong (`hscv(x)` of the `R` package `ks`, Duong 2007) is better than the `R` implementation of UCV for all processes except for the lognormal. However, this implementation is said not to be always stable for large sample sizes when the binning option is used (see the release notes of the package). And without binning, the procedure is just infeasible in large samples.

Because of the presence of a pilot bandwidth g of (43), the choice of ν will be different from the UCV case. We have to determine a first ν to compute the starting value $\widehat{h}_{\mathrm{S}}$ for which the initial selection rule given in Table 6 can be applied. We then have to determine a second ν to compute $\widehat{h}_{\mathrm{a}}$ or $\widehat{h}_{\mathrm{aa}}$. Unreported Monte Carlo experiments indicate that the best choice is $\nu = 30$ (essentially a Gaussian kernel) for this second step. An intuitive explanation can be that as SCV is based on the convolution of two kernels, there is already enough smoothing. Results given in Table 8

Table 8: SCV with Student kernel, its asymptotic approximation, and Jones-Marron-Park SCV: ratio of their ISEs to optimal $\mathrm{ISE}_{\mathrm{opt}}$

	$\widehat{h}_{\mathrm{a}}(\widehat{\nu})$			$\widehat{h}_{\mathrm{aa}}(\widehat{\nu})$			JMP $\widehat{h}_{\mathrm{scv}}$		
	150	450	1000	150	450	1000	150	450	1000
Gaussian	*1.49* (1.42)	*1.23* (0.64)	*1.14* (0.32)	1.62 (1.71)	1.29 (0.77)	1.18 (0.40)	1.50 (1.31)	1.27 (0.68)	1.16 (0.35)
Bimodal mix	*1.08* (0.12)	*1.09* (0.14)	*1.08* (0.13)	1.09 (0.15)	1.09 (0.15)	1.09 (0.15)	1.19 (0.27)	1.14 (0.21)	1.10 (0.16)
Skew mix	1.11 (0.14)	1.15 (0.16)	1.16 (0.16)	*1.07* (0.10)	*1.09* (0.11)	*1.10* (0.11)	1.21 (0.26)	1.16 (0.17)	1.12 (0.14)
Student t(3)	1.46 (1.01)	1.25 (0.80)	1.16 (0.23)	1.56 (1.26)	1.27 (0.93)	*1.15* (0.24)	*1.34* (0.69)	*1.24* (0.46)	1.17 (0.30)
Lognormal	1.20 (0.37)	1.28 (0.32)	1.45 (0.45)	*1.19* (0.37)	*1.23* (0.29)	*1.38* (0.42)	1.69 (0.61)	1.75 (0.52)	1.87 (0.56)

We choose $\widehat{\nu}$ according to the rule of Table 6 for the starting value $\widehat{h}_{\mathrm{S}}$, then $\nu = 30$ is used for the second step. The $\widehat{h}_{\mathrm{scv}}$ corresponds to the standard command `hscv(x)` of the `R` package `ks`. The small number below the ratio of ISEs is Monte Carlo standard deviation of this ratio. Figures in italics indicate best results.

confirm that with these choices, both $\widehat{h}_{\mathrm{a}}(\widehat{\nu})$ and $\widehat{h}_{\mathrm{aa}}(\widehat{\nu})$ manage to beat the `R` implementation of SCV for the lognormal process, a case where SCV is less at ease than UCV. For the Gaussian process where SCV is expected to beat UCV, $\widehat{h}_{\mathrm{a}}(\widehat{\nu})$ is superior while $\widehat{h}_{\mathrm{aa}}(\widehat{\nu})$ is not far. In this Table, standard SCV beats our proposed methods only when n is small and we have a t(3) density. A general rule of thumb emerges, that one should switch from using $\widehat{h}_{\mathrm{a}}(\widehat{\nu})$ to preferring $\widehat{h}_{\mathrm{aa}}(\widehat{\nu})$ when the density is skewed (skewed mixture and lognormal in our table).

E.4 Numerical efficiency

We know that, in large samples, the usual cross validation method can be time consuming when no binning techniques are used and that binning techniques can be problematic for SCV. Table 9 displays absolute execution times in seconds for various methods and various sample sizes. Our method can be between 18 and 24 times quicker. The $\widehat{h}_{\mathrm{a}}$ of SCV has similar execution times as the $\widehat{h}_{\mathrm{a}}$ of UCV, but the $\widehat{h}_{\mathrm{aa}}$ formula is of course more time-consuming because of the pilot bandwidth. Our gain in efficiency starts to be sig-

nificant for large sample sizes (especially above 1,000) or for Monte Carlo experiments. This gain can be very important for large financial datasets or household surveys.

Table 9: Execution time for computing a bandwidth

	450	1000	1500	5000
Student kernel and integral-free methods				
UCV $\widehat{h}_{\mathrm{a}}$	0.084	0.368	0.922	9.322
SCV $\widehat{h}_{\mathrm{a}}$	0.082	0.378	0.894	9.762
SCV $\widehat{h}_{\mathrm{aa}}$	0.166	0.676	1.530	16.862
Gaussian kernel and usual cross validation				
Gauss UCV	2.194	8.466	18.658	185.792
Hayfield and Racine	1.922	7.858	17.132	207.678
Duong SCV	1.490	5.405	11.740	118.665

Time is measured in seconds. All computations are done in `R` on a laptop equipped with an Intel Core i3 at 2.40 GHz. The bandwidth UCV $\widehat{h}_{\mathrm{a}}$ corresponds to (31), while SCV $\widehat{h}_{\mathrm{a}}$ and $\widehat{h}_{\mathrm{aa}}$ correspond to (48) and (46), respectively. Gauss UCV corresponds to the minimization of function (11) for a Gaussian kernel using a Brent algorithm as the `R` command `bw.ucv` includes binning automatically. Hayfield and Racine corresponds to `npudensbw(x,bwmethod="cv.ls")` in the `R` package `np`, Hayfield and Racine (2008). Duong SCV corresponds to `hscv(x, binned=F)` from the `R` package `ks` of Duong (2022). Time for $\widehat{h}_{\mathrm{S}}$ was too quick to be measured.