*Biometrika* (2024), **111**, 3, *pp.* 809–823 https://doi.org/10.1093/biomet/asae007
Advance Access publication 12 February 2024

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

By KARIM M. ABADIR

*Business School, Imperial College London,*
*53 Prince's Gate, South Kensington Campus, London SW7 2AZ, U.K.*
k.m.abadir@imperial.ac.uk

and MICHEL LUBRANO

*Aix-Marseille Université, CNRS, AMSE, Marseille F-13001, France*
michel.lubrano@univ-amu.fr

## Summary

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

*Some key words*: Academic wage distribution; Bandwidth choice; Cross-validation; Explicit analytical solution; Nonparametric density estimation.

## 1. Introduction

Let $\{x_i\}_{i=1}^n$ be an independent and identically distributed sequence of the scalar variate $x$, drawn from a density $f$ that is a continuous function. The kernel density estimator introduced by Rosenblatt (1956) is

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

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 $\hat{f}(u) = n^{-1}\sum_{i=1}^n K_h(u-x_i)$. The asymptotic expectation and variance of this estimator can be calculated, under the usual regularity conditions, leading

© 2024 Biometrika Trust
This is an Open Access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted reuse, distribution, and reproduction in any medium, provided the original work is properly cited.

Downloaded from https://academic.oup.com/biomet/article/111/3/809/7606771 by guest on 11 September 2024

to the asymptotic mean integrated squared error

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

where

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

with the superscript $(j)$ denoting the $j$th derivative of the function. Minimizing the AMISE leads to

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

and to the Epanechnikov kernel $K_h(t) = 1_{|t|<h\sqrt{5}} 3(h^2 - t^2/5)/(4\sqrt{5}h^3)$, with 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 given in § 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 kernel, the AMISE is multiplied by a factor of $\{6(\pi/125)^{1/2}\}^{-4/5} \approx 1.04$, 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 & Marron (1987) or Jones & 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 method to determine the bandwidth that minimizes the integrated squared error asymptotically. The formula for the integrated squared error is

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

where all three components are assumed finite with probability 1. The first integral in the last line of (3) does not affect the procedure and can be omitted from the optimization. The second integral is in terms of the data and the $h$ over which the optimization is conducted. However, the last integral contains both the unknown density and $h$. Cross-validation 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) that shows that the ISE with its optimal $h$ and the ISE with $h$ obtained by cross-validation 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.

Downloaded from https://academic.oup.com/biomet/article/111/3/809/7606771 by guest on 11 September 2024

This cross-validation criterion is an unbiased estimator of the mean integrated squared error, and we refer to it as unbiased cross-validation to stress this. The biased cross-validation criterion proposed by Scott & Terrell (1987) is a biased estimator of the mean integrated squared error, but it reduces the sample variability of the unbiased cross-validation 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 'biased cross-validation performed poorly for several difficult densities without a very large dataset'.

The biased cross-validation of Scott & Terrell (1987) was followed by a number of alternative biased cross-validations, including the modified cross-validation of Stute (1992), the smoothed cross-validation of Hall et al. (1992) and its extension by Jones et al. (1991). The latter is particularly interesting because it derives the functional form of an additional bandwidth that helps cross-validation achieve the fastest rate of convergence relative to $h_0$, a rate that was established by Hall & Marron (1991) as $n^{1/2}$. Smoothed cross-validation was extensively studied for multivariate density estimation in Duong (2004).

The cross-validation method was applied to contexts other than density estimation. It is the main method for determining $h$ in kernel regression models, as illustrated by Müller (1987) and Li & Racine (2006, pp. 66–72). The Nadaraya–Watson nonparametric regression formula is an estimate of the conditional expectation obtainable from joint densities. Robinson & Moyeed (1989) have investigated the efficiency of various cross-validation methods for spline smoothing regression with the objective of obtaining 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 & West (1987) method that requires the estimation of spectra at the origin, as well as the more recent method by Robinson (2005).

None of the cross-validation methods introduced above give an explicit solution for their optimal $h$. We show that there is a common structure to all these cross-validation methods, and we use this to provide an explicit solution for their bandwidths. Furthermore, we conjecture that this structure extends to other cross-validation 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 define in the next section. The solutions we obtain are explicit and hence also 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 cross-validation methods: their excess variability and their failure in the case of repeated observations.

## 2. Method for the explicit solution of bandwidths

Cross-validation criteria necessitate the calculation of $\int_{-\infty}^{\infty} \hat{f}(u)^2 \, \mathrm{d}u$ 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. Biased and unbiased cross-validation and their variants require calculation of

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

Downloaded from https://academic.oup.com/biomet/article/111/3/809/7606771 by guest on 11 September 2024

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. Smoothed cross-validation 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 $L_g(t) = g^{-1}L(g^{-1}t)$, with the smoothed cross-validation optimal $g$ taking the form $\hat{g} \sim Cn^p/\hat{h}^2$ for $C$ a constant as $n \to \infty$ and $p$ a constant to be detailed in § 3.3 below. The notation $a_n \sim b_n$ means that $\lim_{n\to\infty} a_n/b_n = 1$, while $\hat{h}$ and $\hat{g}$ denote bandwidths that solve the optimization of a cross-validation method. They are stochastic, unlike $h_0$, and hence the hat notation.

There are two components to the solution. The first one is straightforward once we recall that the choice of a Gaussian kernel function $\phi$ 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

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

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 $\mathrm{He}_m$ polynomials and their relation to the other type of Hermite polynomials denoted by $\mathrm{H}_m$. See also Aldershof et al. (1995) for uses of these polynomials.

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

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

$$(D_h * D_h * L_g * L_g)(a) = K_{(2h^2+2g^2)^{1/2}}(a) - 2K_{(h^2+2g^2)^{1/2}}(a) + K_{g\sqrt{2}}(a), \tag{8}$$

*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 $h = 0$,*

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

*has a finite $m \in \mathbb{Z}$. This is a Laurent series, which generalizes the 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 $\phi$, but it holds for another kernel that can be made arbitrarily close to $\phi$ and that can be used instead of $\phi$ now that the

Downloaded from https://academic.oup.com/biomet/article/111/3/809/7606771 by guest on 11 September 2024

convolutions have been worked out. Consider a Student-t($\nu$) kernel, $K(t) = c_\nu/(1 + t^2/\nu)^{(\nu+1)/2}$ with

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

see Lemma S1 within the Supplementary Material. The Gaussian is the limiting t($\infty$) case, but $\nu = 30$ makes the two virtually indistinguishable in practice. The scaled version of t($\nu$) 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 $\hat{h} = O_p(n^{-1/5}) \xrightarrow{\text{P}} 0$, (9) becomes asymptotically separable in $t$ and $h$ since $K_h(t) = c_\nu(t^2/\nu)^{-(\nu+1)/2}h^\nu\{1 + O(h^2)\}$ as $h \to 0$ with $t \neq 0$ and $\nu$ 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 $\nu$. This will allow subsequent derivations to give an explicit asymptotic formula for the cross-validation solutions $\hat{h}$. The only available expansion for the Gaussian kernel is $\exp\{-t^2/(2h^2)\} = 1 - t^2/(2h^2) + \cdots$, which has $m = -\infty$ in Definition 1, 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($\nu$) kernel does not suffer this drawback, even for any arbitrarily large but finite $\nu$.

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

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

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 set-up and illustration of the simplified solution

### 3.1. *Unbiased cross-validation criterion*

The first step of the unbiased cross-validation 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 $\hat{f}_{-j}(u) = (n-1)^{-1}\sum_{i\neq j} K_h(u - x_i)$. The last integral in the ISE in (3) is an expectation which can be estimated by $\hat{f}_{n-1}(x_1, \ldots, x_n; h) = n^{-1}\sum_{j=1}^n \hat{f}_{-j}(x_j) = n^{-1}(n-1)^{-1}\sum_{j=1}^n\sum_{i\neq j} K_h(z_{ij})$, where $z_{ij} = x_j - x_i$. Unbiased cross-validation minimizes,

Downloaded from https://academic.oup.com/biomet/article/111/3/809/7606771 by guest on 11 September 2024

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

$$S_1 = \int_{-\infty}^{\infty} f(u)^2 \, \mathrm{d}u, \qquad S_2 = \int_{-\infty}^{\infty} \hat{f}(u)^2 \, \mathrm{d}u, \qquad S_3 = -2\hat{f}_{n-1}(x_1, \ldots, x_n; h).$$

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

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

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

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 as $\sum_{i \neq j} = 2 \sum_{i>j}$. Using $n/(n-1) = 1 + O(1/n)$,

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

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 unbiased cross-validation*

From (9), $K_{h\sqrt{2}}(0) = c_\nu/(h\sqrt{2})$. Applying (10) to (11), and since the unbiased cross-validation optimal $h$ is $\hat{h} = O_p(n^{-1/5})$, 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 therefore minimize

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

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

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

A substitution inside this double sum leads to the same unbiased cross-validation-optimal asymptotic solution.

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

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

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

Downloaded from https://academic.oup.com/biomet/article/111/3/809/7606771 by guest on 11 September 2024

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

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

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

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

We now derive a plug-in to use as $\hat{h}_{\mathrm{p}}$. We could substitute the rule of thumb $\hat{h} = 1.06\hat{\sigma}n^{-1/5}$ of Silverman (1986) mentioned before (3), with $\hat{\sigma}^2$ denoting the sample variance of $\{x_i\}_{i=1}^n$. A more elaborate version would again use (2), but with $f$ replaced by a Student density instead of the Gaussian density. The ingredients for this are given in Lemma S1 within the Supplementary Material, yielding, for $\nu > 2$,

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

with $\lim_{\nu\to\infty} \hat{h}_{\mathrm{S}}/(\hat{\sigma}n^{-1/5}) = (4/3)^{1/5} \approx 1.06$ implying Silverman's rule as a special case.

By $R/n^2 \xrightarrow{\mathrm{P}} -S_1 < 0$, the numerator and denominator in (15) should both be negative at the optimum, thus restricting the allowable solutions for $h$. Note also that $z_{ij}^2/\nu = (x_j - x_i)^2/\nu$, appearing in $y_n(q;\hat{h})$ of (14), is a measure of distance between the data points. It is quadratic because of the adoption of a spherical probability density function 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 given in the Supplementary Material.

The combination of plug-in and cross-validation approaches has also been used by Mammen et al. (2011). They introduced a bandwidth based on the weighted average of a plug-in method and a fully iterated cross-validation, using Epanechnikov, quartic and Gaussian kernels. The empirical intuition is that plug-in methods oversmooth, while cross-validation methods undersmooth, and their argument for considering their combination was the important observation that practical implementation is crucial in achieving the theoretical potential of a method. However, they showed that their asymptotic best weighted-average solution does not perform as well as hoped in small samples, both in terms of average ISE and variability. In the Supplementary Material, our simulations show that both our Student plug-in $\hat{h}_{\mathrm{S}}$ and our cross-validation solution manage to beat the usual methods available in standard packages, both in terms of ISE and variability. Proposition 1 assessed the nonlinear combination of plug-in and cross-validation approaches, where the asymptotic optimality

Downloaded from https://academic.oup.com/biomet/article/111/3/809/7606771 by guest on 11 September 2024

of our combination is now proved for unbiased cross-validation and will apply similarly for smoothed cross-validation below.

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) managed to reach the improved speed of $n^{-1/4}$. Their kernel is robust to rounding and to 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-t($\nu$) being an infinite mixture of Gaussian processes by a $\chi^2$ mixing density, but with only one parameter $\nu$ to determine. Our kernel is also usable for smoothed cross-validation with its optimal $n^{-1/2}$ rate of convergence, as we shall see. In addition, the applications in § 5 below will show that our method is robust to rounding and ties.

### 3.3. *Smoothed cross-validation criterion*

Having analysed unbiased cross-validation, we now introduce smoothed cross-validation. Jones et al. (1991) estimated 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 combined this with the option of using the idea of Jones & Sheather (1991), in which case they set an indicator function $\delta = 1$ below and $\delta = 0$ otherwise. The result is the smoothed cross-validation objective function

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

where 0 and $z_{ij}$ are the arguments of the respective convolutions. They showed that the asymptotically optimal $p$ in $g \sim Cn^p/h^2$ is $\hat{p} = -23/45$ if $\delta = 1$ or $\hat{p} = -44/85$ if $\delta = 0$, but the constant $C$ depends on the unknown $f$ again. They experimented 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 in which $\delta = 1$ achieves the best $n^{-1/2}$ rate for the relative distance between the values of $h$ minimizing mean integrated squared error and $S_{\rm s}$, while it is the slightly slower rate of $n^{-8/17}$ that is obtained if $\delta = 0$. Note that $\hat{g}_{\rm s}$ dominates $\hat{h}_{\rm s}$, where these are the optimizers of $S_{\rm s}$; e.g., if we take $\hat{p}$ to be $-\frac{1}{2}$ henceforth then $\hat{g}_{\rm s} = O_p(n^{-1/10})$ dominates $\hat{h}_{\rm s} = O_p(n^{-1/5})$. Nevertheless, the argument used for $\hat{h}$ in connection with the Student kernel in § 2 applies to $\hat{g}_{\rm s}$ as well.

Although the $n^{-1/2}$ rate is achieved by smoothed cross-validation, the best possible multiplicative constant established by Fan & Marron (1992) is not quite reached by the limiting variance of the normalized $\hat{h}_{\rm s}$. Kim et al. (1994) modified the method to achieve this lower bound, but their results showed that samples as large as $n = 1000$ are not large enough to reach these asymptotics and they say on page 120 that their method is 'mostly of theoretical interest'. We therefore do not include their extension.

## 4. GENERAL SOLUTION FOR UNBIASED AND SMOOTHED CROSS-VALIDATION

### 4.1. *Solution of unbiased and smoothed cross-validation bandwidths for multivariate kernels*

Let the bandwidth matrix be $H = h^2 I$, where $I$ is the identity matrix without a subscript so that it is distinct from the earlier use of $I$. We do not tackle directly the case of $H$ positive definite in full generality, which would require additional $\frac{1}{2}d(d+1) - 1$ bandwidths to be

Downloaded from https://academic.oup.com/biomet/article/111/3/809/7606771 by guest on 11 September 2024

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 § 5 below, and we discuss both there and here the generalization that it implies for $H$. In this section, $x$ now refers to the $d \times 1$ variate, and its $n \times d$ sample matrix is $X = (x_1, \ldots, x_n)^{\mathrm{T}}$; we define the vector $z_{ij} = x_j - x_i$ whose elements are denoted by $z_{ij,m}$. The scaled kernel defined as $K_h(t) = h^{-d}K(h^{-1}t)$ is used to write $\hat{f}(u) = n^{-1}\sum_{i=1}^{n} K_h(u - x_i)$, where $t, u$ are now vectors.

The procedure for orthonormalization is as follows. Since the sample variance matrix $S$ is positive definite, $S = Q\Lambda Q^{\mathrm{T}}$ and the square root of the matrix is the symmetric $S^{1/2} = Q\Lambda^{1/2}Q^{\mathrm{T}}$, where $\Lambda$ is the diagonal matrix of positive eigenvalues of $S$, the columns of $Q$ contain the orthonormal eigenvectors of $S$ and $QQ^{\mathrm{T}} = I$. The orthonormalization is then $y = S^{-1/2}x$, which has $\hat{\mathrm{var}}(y) = S^{-1/2}\hat{\mathrm{var}}(x)S^{-1/2} = I$, where $S^{-1/2} = Q\Lambda^{-1/2}Q^{\mathrm{T}}$; see Abadir & Magnus (2005) for matrix functions. In general, the components of $y$ are uncorrelated, but mutually dependent. Under general conditions, the sample variance is a consistent estimator of $\mathrm{var}(x)$ 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 those below that are for multivariate kernels.

The scaled multivariate $\mathrm{t}(\nu)$ kernel is

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

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

THEOREM 1. *Let $\hat{h}$ denote the solution of a cross-validation-optimal bandwidth. Then we use $\hat{h}_{\mathrm{a}}$ to denote our asymptotic solution satisfying $\lim_{n\to\infty}\hat{h}_{\mathrm{a}}/\hat{h} = 1$ and $\hat{h}_{\mathrm{aa}}$ to denote the leading term of its asymptotic expansion. Take plug-ins $\hat{h}_{\mathrm{p}}, \hat{g}_{\mathrm{p}}$ satisfying $\hat{h}_{\mathrm{p}} = O(n^{-1/(4+d)})$ and $\hat{g}_{\mathrm{p}} = O(n^{-1/(6+d)})$.*

(a) *For unbiased cross-validation, with $y_n(q;h) = \sum_{j=1}^{n}\sum_{i>j}(h^2 + \nu^{-1}z_{ij}^{\mathrm{T}}z_{ij})^{-q-(\nu+d)/2}$, letting*

$$\alpha_1 = 2^{-1-d/2}dn,$$
$$\alpha_2 = \nu\{2^{\nu/2}y_n(0;\hat{h}_{\mathrm{p}}\sqrt{2}) - 2y_n(0;\hat{h}_{\mathrm{p}})\},$$
$$\alpha_3 = -2(\nu + d)\{2^{\nu/2}y_n(1;\hat{h}_{\mathrm{p}}\sqrt{2}) - y_n(1;\hat{h}_{\mathrm{p}})\},$$

*we have*

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

Downloaded from https://academic.oup.com/biomet/article/111/3/809/7606771 by guest on 11 September 2024

(b) *For smoothed cross-validation, with* $y_n(q;h,g)=\sum_{j=1}^{n}\sum_{i>j}(h^2+2g^2+\nu^{-1}z_{ij}^{\mathrm{T}}z_{ij})^{-q-(\nu+d)/2}$, *letting*

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

$$\alpha_1=\frac{k_{02,d}dn}{4c_{\nu,d}}+\frac{\delta dn}{2}\{(2+2n^{1/5})^{-1-d/2}-(1+2n^{1/5})^{-1-d/2}\},$$

$$\alpha_2=\nu\{(2+2n^{1/5})^{(\nu-2)/2}y_n(0;\hat{h}_{\mathrm{p}}\sqrt{2},\hat{g}_{\mathrm{p}})-(1+2n^{1/5})^{(\nu-2)/2}y_n(0;\hat{h}_{\mathrm{p}},\hat{g}_{\mathrm{p}})\},$$

$$\alpha_3=-(\nu+d)\{(2+2n^{1/5})^{\nu/2}y_n(1;\hat{h}_{\mathrm{p}}\sqrt{2},\hat{g}_{\mathrm{p}})-(1+2n^{1/5})^{\nu/2}y_n(1;\hat{h}_{\mathrm{p}},\hat{g}_{\mathrm{p}})\},$$

*we have*

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

$$\textit{and}\quad \hat{h}_{\mathrm{aa}}=\left[\frac{y_n(0;\hat{h}_{\mathrm{p}}\sqrt{2},\hat{g}_{\mathrm{p}})-y_n(0;\hat{h}_{\mathrm{p}},\hat{g}_{\mathrm{p}})}{(1+d/\nu)\{y_n(1;\hat{h}_{\mathrm{p}}\sqrt{2},\hat{g}_{\mathrm{p}})-y_n(1;\hat{h}_{\mathrm{p}},\hat{g}_{\mathrm{p}})\}}-2\hat{g}_{\mathrm{aa}}^2\right]^{1/2} \tag{20}$$

*with*

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

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

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

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

### 4.2. *Multivariate plug-ins*

We consider the multivariate version of $h_0=k_{02}^{1/5}(nk_{21}^2I_2)^{-1/5}$ of (2) and recalculate its components to get $\hat{h}_{\mathrm{p}}$ in the case of a multivariate Student-t($\nu$) kernel. 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

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

Downloaded from https://academic.oup.com/biomet/article/111/3/809/7606771 by guest on 11 September 2024

leading to the generalization of (2) as

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

where $k_{21} = \nu/(\nu-2)$ as before, $k_{02,d}$ is given in (19) and $I_2 = \int_{\mathbb{R}^d} \{\sum_{j=1}^d \partial^2 f(u)/\partial u_j^2\}^2 \, \mathrm{d}u$ now; see, e.g., Härdle & Müller (2000). It remains for us to work out $I_2$ for a multivariate Student-t($\nu$) plug-in density, that is, a multivariate version of our generalized Silverman rule. From Lemma S1(iv) within the Supplementary Material,

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

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 $\nu$ follows from now having all the ingredients for (25) as

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

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

Theorem 3 of Duong & Hazelton (2005) implies a $\hat{g}_{\mathrm{p}}$ for smoothed cross-validation, 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) gave a numerical way to compute these, which yields a $\hat{g}_{\mathrm{p}}$ 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

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

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

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

However, $d = 1$ here would give $\hat{g}_{\mathrm{p}} = \hat{h}_{\mathrm{p}} n^{2/35}$, an overestimate by $n^{3/70}$ of $\hat{g}_{\mathrm{p}}$ compared to $\hat{g}_{\mathrm{p}} = \hat{h}_{\mathrm{p}} n^{1/10}$ of (22). Since the notation for orders of magnitude is an inequality relation, we adopt the larger order $\hat{g}_{\mathrm{p}} = O(n^{-1/(6+d)})$ used in the optimality derivations of Duong & 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).

Downloaded from https://academic.oup.com/biomet/article/111/3/809/7606771 by guest on 11 September 2024

Table 1. *Bandwidths for the wage dataset of Michigan State University*

| | Student kernel | | | | Gaussian kernel | | |
|---|---|---|---|---|---|---|---|
| $\nu$ | $\hat{h}_{\mathrm{S}}$ | UCV $\hat{h}_{\mathrm{a}}$ | SCV $\hat{h}_{\mathrm{a}}$ | SCV $\hat{h}_{\mathrm{aa}}$ | 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 | | | |

UCV, unbiased cross-validation; SCV, smoothed cross-validation.

## 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 database for 2012. An additional practical advantage of our explicit formulae is to avoid the troubles faced by existing cross-validation approaches when there are some ties in the data, in this case some equal salaries and/or experience.

The database contains 6402 entries, after deleting 22 lines that corresponded to a null wage. Deleting duplicate names, as the same person can be appointed by several departments, we were left with $n = 5050$ distinct individuals earning 4070 different salaries. The minimum yearly wage is \$3600, due to being 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 of 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.

This very asymmetric wage distribution has a Kolmogorov–Smirnov measured complexity equal to $d_n(x_1, \ldots, x_n) = 0.120$, 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 unbiased cross-validation 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, that is those with the same qualification and experience, tend to have similar wages. The value obtained for standard unbiased cross-validation corresponds to the lower bound of the grid search of the Brent algorithm of `bw.ucv` in `R` (R Development Core Team, 2024). 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 $\hat{h} = 1.06\hat{\sigma}n^{-1/5}$ produces an even larger value of 10.41. In both cases, these represent unreliable window sizes, the effects of which are depicted in Fig. 1(b), where we see undersmoothing and oversmoothing respectively.

None of our formulae suffer these drawbacks. Our unbiased cross-validation's $\hat{h}_{\mathrm{a}}$ helps to identify small details of the wage distribution, while our two integral-free smoothed cross-validations $\hat{h}_{\mathrm{a}}$ and $\hat{h}_{\mathrm{aa}}$ give a smoother density, and similarly for our generalized Silverman rule $\hat{h}_{\mathrm{S}}$. We use $\nu = 6$ in Fig. 1(a) and see the following features. The wage density presents several bumps that are well identified when using smoothed cross-validation $\hat{h}_{\mathrm{aa}}$, 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 S6 within the Supplementary Material. The two main modes of the distribution are well identified with all our methods. Our unbiased cross-validation's $\hat{h}_{\mathrm{a}}$, even if it provides slightly more variability than our other bandwidths, helps to identify the first very small mode of the distribution that corresponds to 49 teachers with a fixed-term contract and who are all paid \$22 870 a year. The second and main mode

Downloaded from https://academic.oup.com/biomet/article/111/3/809/7606771 by guest on 11 September 2024

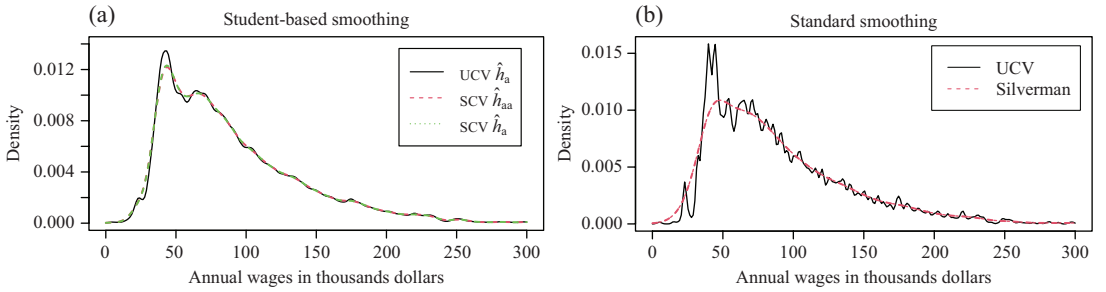


Fig. 1. Wage density estimation for Michigan State University in 2012.

Table 2. *Bandwidth* $2 \times 2$ *matrices for the tenured subsample*

| Plug-in | | UCV | | SCV | |
|---|---|---|---|---|---|
| | | `ks` (Duong, 2022) | | | |
| 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 with $\nu = 6$ | | | |
| 4.280 | 0.024 | 2.880 | 0.016 | 3.518 | 0.020 |
| 0.024 | 0.006 | 0.016 | 0.004 | 0.020 | 0.005 |

UCV, unbiased cross-validation; SCV, smoothed cross-validation.

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 to either specialists or assistant professors with a labour contract that 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 1545 members of the university. In the top panel of Table 2, we present the $H$ matrices obtained using the `R` package `ks` of Duong (2022). It yielded unusual values for unbiased cross-validation because of the presence of repeated observations. In the bottom panel of Table 2, we provide the same quantities for our formulae based on $\hat{h}_{aa}$ with a multivariate Student kernel having $\nu = 6$. The results are more in accordance with what one would expect, unaffected by repeated observations. Unbiased cross-validation corresponds to some undersmoothing, while smoothed cross-validation is between the Student-generalized plug-in and unbiased cross-validation.

The four plots reported in Fig. 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 nonparametric regressions for each plot, by taking a sequence of vertical lines at various experience levels, then calculating the mean, median 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 Fig. 2(b), as expected from it being a conditional expectation. The curves we get from our formulae are less volatile than those obtained by other cross-validation estimates, as indicated earlier, and we can see this here when we compare them with the standard smoothed cross-validation reported in Fig. 2(a). Some of the crossovers of curves in Fig. 2 can provide counterexamples to the mean-median-mode inequality, in the case of

Downloaded from https://academic.oup.com/biomet/article/111/3/809/7606771 by guest on 11 September 2024

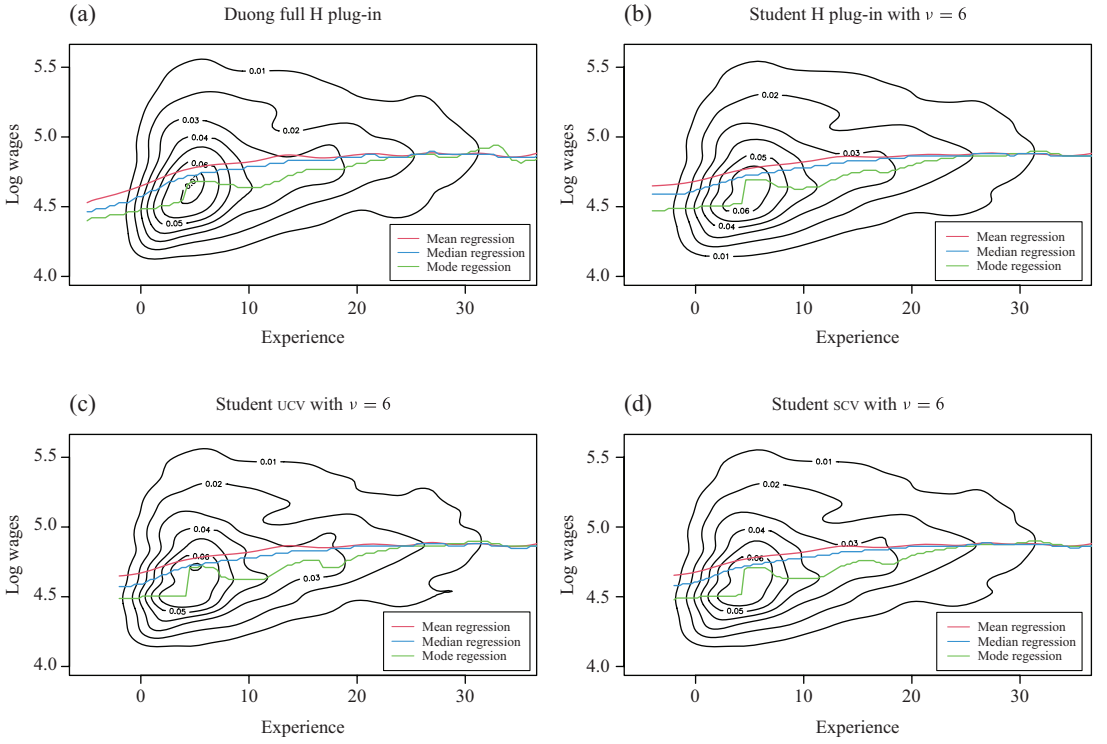


Fig. 2. Bivariate density of log wages and experience, using various methods and a multivariate kernel.

a unimodal, or nearly so, conditional distribution, in addition to those of Abadir (2005). We conclude by cautioning that this regression is incomplete because other variables also determine log wages in academia.

## ACKNOWLEDGEMENT

We are grateful for comments received at several conferences. We thank Jeff Racine for useful discussions and suggestions, Omiros Papaspiliopoulos for suggesting the multivariate extension of our earlier univariate results, Paul Fearnhead, the associate editor and referee. Support from the ESRC is gratefully acknowledged. Abadir is also affiliated with the American University in Cairo. Lubrano is also affiliated with Jiangxi University of Finance and Economics, China. This project was supported by the French National Research Agency (ANR-17-EURE-0020) and the Excellence Initiative of Aix-Marseille University - A*MIDEX.

## SUPPLEMENTARY MATERIAL

The Supplementary Material includes proofs, as well as our solutions for biased cross-validation and for other kernels, including Epanechnikov and product kernels. We have also included results of a Monte Carlo experiment showing that the Epanechnikov kernel is best for estimating a density if the latter is Gaussian, but that a Student kernel is better in all other cases. We also give details on execution times, showing that our method can be around 20 times faster than other cross-validation methods, especially in large samples.

Downloaded from https://academic.oup.com/biomet/article/111/3/809/7606771 by guest on 11 September 2024

## References


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

Abadir, K. M. (2005). The mean-median-mode inequality: counterexamples. *Economet. Theory* **21**, 477–82.

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

Aldershof, B., Marron, J., Park, B. & Wand, M. (1995). Facts about the Gaussian probability density function. *Appl. Anal.* **59**, 289–306.

Bowman, A. W. (1984). An alternative method of cross-validation for the smoothing of density estimates. *Biometrika* **71**, 353–60.

Duong, T. (2004). *Bandwidth Selectors for Multivariate Kernel Density Estimation.* Ph.D. 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. *J. Statist. 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. & Hazelton, M. L. (2005). Cross-validation bandwidth matrices for multivariate kernel density estimation. *Scand. J. Statist.* **32**, 485–506.

Fan, J. & Marron, J. S. (1992). Best possible constant for bandwidth selection. *Ann. Statist.* **20**, 2057–70.

Hall, P. & Marron, J. (1987). Estimation of integrated squared density derivatives. *Statist. Prob. Lett.* **6**, 109–15.

Hall, P. & Marron, J. (1991). Lower bounds for bandwidth selection in density estimation. *Prob. Theory Rel. Fields* **90**, 149–73.

Hall, P., Marron, J. S. & Park, B. U. (1992). Smoothed cross-validation. *Prob. Theory Rel. Fields* **92**, 1–20.

Härdle, W. & Müller, M. (2000). Multivariate and semiparametric kernel regression. In *Smoothing and Regression*, Ed. M. G. Schimek, pp. 357–91. New York: John Wiley and Sons.

Jones, M. C. & Sheather, S. J. (1991). Using nonstochastic terms to advantage in kernel-based estimation of integrated squared density estimates. *Statist. Prob. Lett.* **6**, 511–4.

Jones, M. C., Marron, J. S. & Park, B. U. (1991). A simple root *n* bandwidth selector. *Ann. Statist.* **19**, 1919–32.

Kim, W. C., Park, B. U. & Marron, J. S. (1994). Asymptotically best bandwidth selectors in kernel density estimation. *Statist. Prob. Lett.* **19**, 119–27.

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

Mammen, E., Miranda, M., Nielsen, J. & Sperlich, S. (2011). Do-validation for kernel density estimation. *J. Am. Statist. Assoc.* **106**, 651–60.

Müller, H.-G. (1987). Weighted local regression and kernel methods for nonparametric curve fitting. *J. Am. Statist. Assoc.* **82**, 231–8.

Newey, W. K. & West, K. D. (1987). A simple, positive semi-definite, heteroskedasticity and autocorrelation consistent covariance matrix. *Econometrica* **55**, 703–8.

R Development Core Team (2024). *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. ISBN 3-900051-07-0, http://www.R-project.org.

Robinson, P. M. (2005). Robust covariance matrix estimation: HAC estimates with long memory/antipersistence correction. *Economet. Theory* **21**, 171–80.

Robinson, T. & Moyeed, R. (1989). Making robust the cross-validatory choice of smoothing parameter in spline smoothing regression. *Commun. Statist.* A **18**, 523–39.

Rosenblatt, M. (1956). Remarks on some nonparametric estimates of a density function. *Ann. Math. Statist.* **27**, 832–7.

Rudemo, M. (1982). Empirical choice of histograms and kernel density estimators. *Scand. J. Statist.* **9**, 65–78.

Savchuk, O., Hart, J. & Sheather, S. (2010). Indirect cross-validation for density estimation. *J. Am. Statist. Assoc.* **105**, 415–23.

Scott, D. W. (2015). *Multivariate Density Estimation: Theory, Practice and Visualization*. New York: John Wiley and Sons.

Scott, D. W. & Terrell, G. R. (1987). Biased and unbiased cross-validation in density estimation. *J. Am. Statist. Assoc.* **82**, 1131–46.

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

Stone, C. (1984). An asymptotically optimal window selection rule for kernel density estimates. *Ann. Statist.* **12**, 1285–97.

Stute, W. (1992). Modified cross-validation in density estimation. *J. Statist. Plan. Infer.* **30**, 293–305.

Velasco, C. (2000). Local cross-validation for spectrum bandwidth choice. *J. Time Ser. Anal.* **21**, 329–61.


[*Received on* 15 *March* 2023. *Editorial decision on* 30 *January* 2024]

Downloaded from https://academic.oup.com/biomet/article/111/3/809/7606771 by guest on 11 September 2024

Downloaded from https://academic.oup.com/biomet/article/111/3/809/7606771 by guest on 11 September 2024