Australian & New Zealand Journal of Statistics

Aust. N. Z. J. Stat. 2012 doi: 10.1111/j.1467-842X.2011.00639.x

UNCONSTRAINED PILOT SELECTORS FOR SMOOTHED CROSS-VALIDATION

JOSÉ E. CHACÓN[1,*] AND TARN DUONG[2]

Universidad de Extremadura and Institut Curie

Summary

Two of the most useful multivariate bandwidth selection techniques are the plug-in and cross-validation methods. The smoothed version of the cross-validation method is known to reduce the variability of its non-smoothed counterpart; however, it shares with the plug-in choice the need for a pilot bandwidth matrix. Owing to the mathematical difficulties encountered in the optimal pilot choice, it is common to restrict this pilot matrix to be a scalar multiple of the identity matrix, at the expense of losing the flexibility afforded by the unconstrained approach. Here we show how to overcome these difficulties and propose a smoothed cross-validation selector using an unconstrained pilot matrix. Our numerical results indicate that the unconstrained selector outperforms the constrained one in practice, and is a viable competitor to unconstrained plug-in selectors.

Key words: cross-validation; kernel density estimation; mean integrated squared error; unconstrained bandwidth matrices.

1. Introduction

The ability of kernel density estimators to disclose the structure of multivariate data clouds depends crucially on the selection of the smoothing parameter, commonly known as the bandwidth matrix (see Simonoff 1996 for a review). The parametrization of this bandwidth is an important factor in the performance of kernel density estimators. Constrained parametrizations restrict the kernels to be aligned to the coordinate axes. In contrast, unconstrained bandwidth matrices allow for the most appropriately oriented kernels, so are able to more clearly reveal structures that are oriented away from the coordinate axes. Moreover, Wand & Jones (1993) and Chacón (2009) show that kernel density estimators with unconstrained bandwidths may substantially outperform those using constrained ones.

Optimal unconstrained bandwidth matrices with optimal *scalar* pilot bandwidths have been developed for plug-in selectors (Wand & Jones 1994; Duong & Hazelton 2003) and for smoothed cross-validation (Duong & Hazelton 2005b). The asymptotic and finite-sample performance of these two classes of selectors has as its core the ability to tune the pilot estimators that contribute to the intermediate stages of estimating the optimality criteria. Focusing on

*Author to whom correspondence should be addressed.
[1]Departamento de Matemáticas, Universidad de Extremadura, E-06006, Badajoz, Spain.
e-mail: jechacon@unex.es
[2]Institut Curie, Molecular Mechanisms of Intracellular Transport Laboratory, CNRS, UMR 144, F-75248 Paris, France.
e-mail: tduong@curie.fr

Acknowledgements. We thank an anonymous associate editor and two anonymous referees for their comments, which led to a substantial improvement in the quality of the paper. Grants MTM2009-0730 (J.E.C.) and MTM2010-16660 (both authors) from the Spanish Ministerio de Ciencia e Innovacion, and Mayent-Rothschild and Agence Nationale de Recherche fellowships (T.D.) from the Institut Curie, France have supported this work.

the scalar parametrization of the pilot selectors reduces the complexity of the mathematical analysis, at the price of losing some flexibility. Nevertheless, with some novel matrix analysis results for higher-order derivatives of multivariate functions, these mathematical complexities were surmounted by Chacón & Duong (2010), who presented a plug-in selector using unconstrained matrices for all pilot stages for the first time in the literature, and demonstrated the gains in performance over their constrained counterparts in practice. We present here analogous unconstrained pilot selectors for smoothed cross-validation selectors.

In Section 2, we outline the optimal bandwidth selection problem for kernel density estimators. For smoothed cross-validation, this problem relies in turn on optimal pilot bandwidth selection, as elaborated in Section 3. This section contains our main asymptotic results for unconstrained pilot selectors. This is followed by a numerical study in Section 4 to investigate the performance for finite samples. In a simulation study we show that the theoretical effort made to derive unconstrained pilot selectors is worthwhile, as this new approach consistently outperforms the constrained one in all the considered examples. Finally, we illustrate the differences between various bandwidth selection methods in practice through the analysis of global cellular organization from microscopy image data.

2. Optimal bandwidth selection for kernel density estimation

For a d-variate random sample $\mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_n$ drawn from a common density f, the kernel density estimator is defined as

$$\hat{f}_{\mathbf{H}}(\boldsymbol{x}) = n^{-1} \sum_{i=1}^{n} K_{\mathbf{H}}(\boldsymbol{x} - \mathbf{X}_i),$$

where $\boldsymbol{x} = (x_1, x_2, \ldots, x_d)^\top$ and $\mathbf{X}_i = (X_{i1}, X_{i2}, \ldots, X_{id})^\top$, $i = 1, 2, \ldots, n$. Here $K(\boldsymbol{x})$ is the multivariate kernel, which is a spherically symmetric probability density function. The parameter $\mathbf{H}$ is the bandwidth matrix, which is symmetric and positive-definite, and $K_{\mathbf{H}}(\boldsymbol{x}) = |\mathbf{H}|^{-1/2} K(\mathbf{H}^{-1/2}\boldsymbol{x})$ is the scaled kernel.

We measure the performance of a kernel density estimate using the mean integrated squared error (MISE), defined as

$$\mathrm{MISE}(\mathbf{H}) = \mathrm{MISE}(\hat{f}_{\mathbf{H}}) = \mathbb{E} \int_{\mathbb{R}^d} \{\hat{f}_{\mathbf{H}}(\boldsymbol{x}) - f(\boldsymbol{x})\}^2 \, d\boldsymbol{x},$$

assuming that both K and f are square-integrable. By expanding the integral in the previous equation, the MISE can be rewritten as

$$\begin{aligned} \mathrm{MISE}(\mathbf{H}) \;&= \left\{ n^{-1} |\mathbf{H}|^{-1/2} R(K) - n^{-1} R^*(K_{\mathbf{H}} * K_{\mathbf{H}}, f) \right\} \\ &\quad + \{ R^*(K_{\mathbf{H}} * K_{\mathbf{H}}, f) - 2R^*(K_{\mathbf{H}}, f) + R(f) \}, \end{aligned} \tag{1}$$

where $R(\alpha) = \int_{\mathbb{R}^d} \alpha(\boldsymbol{x})^2 \, d\boldsymbol{x}$ for any square-integrable function α, and $R^*(L_{\mathbf{H}}, f) = \int_{\mathbb{R}^d} (L_{\mathbf{H}} * f)(\boldsymbol{x}) f(\boldsymbol{x}) \, d\boldsymbol{x}$ for any square-integrable kernel L, with $*$ denoting the convolution operator; see, for instance, Chacón, Duong & Wand (2011). The first set of braces contains the contribution from the integrated variance, and the second set that from the integrated squared bias.

This is not the usual expression of the MISE: we use this alternative form because it leads more naturally into our treatment of smoothed cross-validation. The usual asymptotic

approximation to the MISE, for example as found in Wand (1992), is calculated by ignoring the contribution of the second term in the integrated variance, and by keeping only the leading term in the integrated squared bias, resulting in

$$\mathrm{AMISE}(\mathbf{H}) = n^{-1}|\mathbf{H}|^{-1/2}R(K) + \frac{1}{4}m_2(K)^2 \int_{\mathbb{R}^d} \mathrm{tr}^2\{\mathbf{H}\mathsf{D}^2 f(\boldsymbol{x})\}\, d\boldsymbol{x},$$

where $\mathsf{D}^2 f$ is the Hessian matrix of second-order partial derivatives of f, tr denotes the trace operator, and $m_2(K) \in \mathbb{R}$ is such that $\int_{\mathbb{R}^d} \boldsymbol{x}\boldsymbol{x}^\top K(\boldsymbol{x})d\boldsymbol{x} = m_2(K)\mathbf{I}_d$, with $\mathbf{I}_d$ denoting the identity matrix of order d. This AMISE approximation forms the basis of plug-in selectors; see Wand & Jones (1994).

The crucial consideration in kernel density estimation is to select an optimal value for the bandwidth matrix. The MISE-optimal selector is defined to be the minimizer of the MISE:

$$\mathbf{H}_{\mathrm{MISE}} = \operatorname*{argmin}_{\mathbf{H}\in\mathcal{F}} \mathrm{MISE}(\mathbf{H}),$$

where $\mathcal{F}$ is the set of all symmetric and positive-definite $d \times d$ matrices. This ideal selector is mathematically intractable for general densities, so instead it is common to deal with $\mathbf{H}_{\mathrm{AMISE}}$, the minimizer of the AMISE over $\mathcal{F}$, which is asymptotically equivalent to $\mathbf{H}_{\mathrm{MISE}}$. In fact, when f has sufficient smoothness, $\mathbf{H}_{\mathrm{AMISE}}$ can be written as $\mathbf{C}n^{-2/(d+4)}$ for a certain positive-definite symmetric matrix $\mathbf{C}$ (not having an explicit form), so that

$$\mathbf{H}_{\mathrm{MISE}} = \mathbf{C}n^{-2/(d+4)} + o(n^{-2/(d+4)}\mathbf{J}_d),$$

where $\mathbf{J}_d$ is the $d \times d$ matrix having each entry equal to one. Here and hereafter, the asymptotic notation for matrices is to be understood element-wise, as introduced in Duong & Hazelton (2005a).

The smoothed cross-validation (SCV) method of Hall, Marron & Park (1992) attempts to improve the bias estimation in the AMISE expansion. The target for SCV can thus be considered to be

$$\mathrm{MISE2}(\mathbf{H}) = n^{-1}|\mathbf{H}|^{-1/2}R(K) + R^*(K_{\mathbf{H}} * K_{\mathbf{H}}, f) - 2R^*(K_{\mathbf{H}}, f) + R(f), \qquad (2)$$

that is, it keeps the exact integrated squared bias as in the MISE, and only the dominant term in the integrated variance. It is clear that $\mathrm{MISE}(\mathbf{H}) - \mathrm{MISE2}(\mathbf{H}) = O(n^{-1})$, so MISE and MISE2 are asymptotically equivalent. The following theorem establishes that $\mathbf{H}_{\mathrm{MISE2}}$, the minimizer of MISE2, is also asymptotically equivalent to $\mathbf{H}_{\mathrm{MISE}}$, and provides the relative rate of convergence for this equivalence. Here the relative rate of convergence is defined to be $n^{-\alpha}$ if $\mathrm{vec}\,(\mathbf{H}_{\mathrm{MISE2}} - \mathbf{H}_{\mathrm{MISE}}) = O(n^{-\alpha}\mathbf{J}_{d^2})\,\mathrm{vec}\,\mathbf{H}_{\mathrm{MISE}}$, where vec is the operator that concatenates the columns of a matrix into a single vector (see Duong & Hazelton 2005a).

Theorem 1. *Suppose that the following conditions hold:*

(H) *For the sequence of bandwidth matrices* $\mathbf{H} = \mathbf{H}_n$, *every element of* $\mathbf{H} \to 0$ *and* $n^{-1}|\mathbf{H}|^{-1/2} \to 0$ *as* $n \to \infty$.

(D) *All partial derivatives up to order 6 inclusive of the density function* f *are bounded, continuous and square-integrable.*

(K) *The kernel* K *is a symmetric, square-integrable density function such that* $\int_{\mathbb{R}^d} \boldsymbol{x}\boldsymbol{x}^\top K(\boldsymbol{x})d\boldsymbol{x} = m_2(K)\mathbf{I}_d$ *and all its moments of order 4 are finite.*

Then $\mathrm{vec}\,(\mathbf{H}_{\mathrm{MISE2}} - \mathbf{H}_{\mathrm{MISE}}) = O(n^{-(d+2)/(d+4)}\mathbf{J}_{d^2})\,\mathrm{vec}\,\mathbf{H}_{\mathrm{MISE}}$.

The proof of this result, along with all other proofs and supporting lemmas, is presented in the Appendix. Conditions (H), (D) and (K) are not a minimal set of hypotheses; rather, they serve as a useful starting point. An example of a kernel fulfilling condition (K) is the normal kernel, given by $\phi(\boldsymbol{x}) = (2\pi)^{-d/2} \exp(-\frac{1}{2}\boldsymbol{x}^\top \boldsymbol{x})$ for $\boldsymbol{x} \in \mathbb{R}^d$.

Note that the relative rate in the previous result is faster than $n^{-1/2}$ for all d. Because $n^{-1/2}$ is the fastest relative rate than can be achieved in this bandwidth selection problem (Hall & Marron 1991), it follows that we can replace $\mathbf{H}_{\mathrm{MISE}}$ with $\mathbf{H}_{\mathrm{MISE2}}$ everywhere in our asymptotic analysis. In contrast, it can be shown that $\operatorname{vec}(\mathbf{H}_{\mathrm{AMISE}} - \mathbf{H}_{\mathrm{MISE}})$ is $O(n^{-2/(d+4)}) \mathbf{J}_{d^2}) \operatorname{vec} \mathbf{H}_{\mathrm{MISE}}$, so that $\mathbf{H}_{\mathrm{AMISE}}$ has a slower relative rate of convergence to $\mathbf{H}_{\mathrm{MISE}}$ than $\mathbf{H}_{\mathrm{MISE2}}$. In this sense, $\mathbf{H}_{\mathrm{MISE2}}$ is a more efficient target than $\mathbf{H}_{\mathrm{AMISE}}$ for all d. This can be regarded as a consequence of the improvement in the estimation of the bias. Furthermore, as the dimension d increases, the convergence rate of $\mathbf{H}_{\mathrm{AMISE}}$ becomes slower, whereas for $\mathbf{H}_{\mathrm{MISE2}}$ it becomes faster.

3. Pilot bandwidth selection for smoothed cross-validation

The result of Theorem 1 does not provide a way to choose the bandwidth matrix, as $\mathbf{H}_{\mathrm{MISE2}}$, like $\mathbf{H}_{\mathrm{MISE}}$ and $\mathbf{H}_{\mathrm{AMISE}}$, still depends on the unknown density f. To determine a data-based selector, the SCV method replaces f in equation (2) with a pilot estimator $\tilde{f}_{\mathbf{G}}(\boldsymbol{x}) = n^{-1} \sum_{i=1}^{n} L_{\mathbf{G}}(\boldsymbol{x} - \mathbf{X}_i)$ with kernel L and bandwidth $\mathbf{G}$, possibly different from K and $\mathbf{H}$. To be precise, following Jones, Marron & Park (1991) let $\Delta_{\mathbf{H}} = K_{\mathbf{H}} - K_0$, where K_0 is our notation for the Dirac delta function, and let $\bar{\ell} = \ell * \ell$ be the self-convolution of a function ℓ. Then $\bar{\Delta}_{\mathbf{H}} = \bar{K}_{\mathbf{H}} - 2K_{\mathbf{H}} + K_0$, so that MISE2 admits the expression,

$$\mathrm{MISE2}(\mathbf{H}) = n^{-1}|\mathbf{H}|^{-1/2} R(K) + R^*(\bar{\Delta}_{\mathbf{H}}, f)$$

as $R^*(\bar{\Delta}_{\mathbf{H}}, f) = \int_{\mathbb{R}^d} (\bar{\Delta}_{\mathbf{H}} * f)(\boldsymbol{x}) f(\boldsymbol{x})\, d\boldsymbol{x}$ is the integrated squared bias term from (1). The SCV criterion is a plug-in estimator of this form of MISE2

$$\begin{aligned} \mathrm{SCV}(\mathbf{H}) &= n^{-1}|\mathbf{H}|^{-1/2} R(K) + R^*(\bar{\Delta}_{\mathbf{H}}, \tilde{f}_{\mathbf{G}}) \\ &= n^{-1}|\mathbf{H}|^{-1/2} R(K) + n^{-2} \sum_{i,j=1}^{n} (\bar{\Delta}_{\mathbf{H}} * \bar{L}_{\mathbf{G}})(\mathbf{X}_i - \mathbf{X}_j) \end{aligned}$$

and the SCV selector is the minimizer of the SCV function, $\hat{\mathbf{H}}_{\mathrm{SCV}} = \operatorname{argmin}_{\mathbf{H} \in \mathcal{F}} \mathrm{SCV}(\mathbf{H})$.

As stated in the Introduction, the main goal of this paper is to investigate the choice of the pilot bandwidth $\mathbf{G}$ to be used in the SCV selector. As $\hat{\mathbf{H}}_{\mathrm{SCV}}$ can be considered an estimate of $\mathbf{H}_{\mathrm{MISE}}$, proceeding as in Duong & Hazelton (2005b) we set the optimality criterion for the SCV pilot selector to be

$$\mathrm{MSE}(\mathbf{G}) = \mathrm{MSE}(\hat{\mathbf{H}}_{\mathrm{SCV}}; \mathbf{G}) = \mathbb{E}\left\{ \operatorname{vec}^\top (\hat{\mathbf{H}}_{\mathrm{SCV}} - \mathbf{H}_{\mathrm{MISE}}) \operatorname{vec} (\hat{\mathbf{H}}_{\mathrm{SCV}} - \mathbf{H}_{\mathrm{MISE}}) \right\}.$$

Because this MSE expression is difficult to manage, we derive below its asymptotic approximation, which is easier to interpret for our purposes.

We need the following notation: starting with the d-dimensional vector of first-order differentials $\mathrm{D} = (\partial/\partial x_1, \ldots, \partial/\partial x_d)^\top$, and adopting the convention $(\partial/\partial x_i)(\partial/\partial x_j) = \partial^2/(\partial x_i \partial x_j)$, following Holmquist (1996) our definition of the rth-order derivative of f as a

single vector of length d^r is

$$\mathsf{D}^{\otimes r} f(\boldsymbol{x}) = (\mathsf{D} f)^{\otimes r}(\boldsymbol{x}) = \frac{\partial^r f(\boldsymbol{x})}{\partial \boldsymbol{x}^{\otimes r}},$$

where $\mathbf{A}^{\otimes r}$ denotes the r-fold Kronecker product of a matrix $\mathbf{A}$; see Chacón & Duong (2010) or Chacón *et al.* (2011), where this notation is also used. The arrangement of the rth partial derivatives into a vector allows us to extend the usual scalar integrated density derivative functional (Hall & Marron 1987) to its vector-valued counterpart

$$\psi_r = \int_{\mathbb{R}^d} \mathsf{D}^{\otimes r} f(\boldsymbol{x}) f(\boldsymbol{x})\, d\boldsymbol{x} \in \mathbb{R}^{d^r}.$$

Finally, let us write $\mathbf{R}(\boldsymbol{\ell}) = \int_{\mathbb{R}^d} \boldsymbol{\ell}(\boldsymbol{x})\boldsymbol{\ell}(\boldsymbol{x})^\top\, d\boldsymbol{x}$ for a vector-valued function $\boldsymbol{\ell}$.

The asymptotic approximation of MSE($\mathbf{G}$) is given in the next result.

Theorem 2. *Suppose that (H),(D),(K) from Theorem 1 and the following conditions hold:*

(G) *For the bandwidth sequence $\mathbf{G} = \mathbf{G}_n$, every element of $\mathbf{G}$ and $\mathbf{G}^{-1}\mathbf{H} \to 0$ as $n \to \infty$.*

(L) *The kernel L is a symmetric, square-integrable density function such that $\int_{\mathbb{R}^d} \boldsymbol{x}\boldsymbol{x}^\top L(\boldsymbol{x}) d\boldsymbol{x} = m_2(L)\mathbf{I}_d$. All partial derivatives of L up to order 4 inclusive are continuous, bounded and square-integrable.*

The MSE has the asymptotic representation $\mathrm{MSE}(\mathbf{G}) = \mathrm{AMSE}(\mathbf{G})[1 + o(1)]$, *where*

$$\mathrm{AMSE}(\mathbf{G}) = \frac{1}{4} m_2(K)^4 \mathrm{tr}\left\{\left(\Omega_4 + \boldsymbol{\omega}_4\boldsymbol{\omega}_4^\top\right)\left(\mathrm{vec}\, \mathbf{H}_{\mathrm{AMISE}} \mathrm{vec}^\top \mathbf{H}_{\mathrm{AMISE}} \otimes \mathbf{I}_{d^2}\right)\right\}$$

with

$$\boldsymbol{\Omega}_r = 4n^{-1} var\{\mathsf{D}^{\otimes r} f(\mathbf{X})\} + 2n^{-2} R(f) |\mathbf{G}|^{-1/2} (\mathbf{G}^{-1/2})^{\otimes r} \mathbf{R}(\mathsf{D}^{\otimes r} \bar{L}) (\mathbf{G}^{-1/2})^{\otimes r},$$

$$\boldsymbol{\omega}_r = n^{-1} |\mathbf{G}|^{-1/2} (\mathbf{G}^{-1/2})^{\otimes r} \mathsf{D}^{\otimes r} \bar{L}(0) + \frac{1}{2} m_2(\bar{L}) \left(\mathrm{vec}^\top \mathbf{G} \otimes \mathbf{I}_{d^r}\right) \psi_{r+2}$$

for every even number r, where $\mathbf{X}$ is a random variable having density f.

The plug-in (PI) method proposed in Chacón & Duong (2010) uses the AMISE instead of the MISE2 function as a target. They show that the AMISE can be rewritten in the form

$$\mathrm{AMISE}(\mathbf{H}) = n^{-1} |\mathbf{H}|^{-1/2} R(K) + \frac{1}{4} m_2(K)^2 \psi_4^\top (\mathrm{vec}\, \mathbf{H})^{\otimes 2}.$$

The only unknown in this AMISE expression is ψ_4, which is estimated by $\hat{\psi}_4(\mathbf{G}) = n^{-2} \sum_{i,j=1}^n \mathsf{D}^{\otimes 4} L_{\mathbf{G}}(\mathbf{X}_i - \mathbf{X}_j)$, so that the optimal pilot bandwidth for the PI selector is given by the matrix $\mathbf{G}$ minimizing the mean squared error $\mathbb{E}[\|\hat{\psi}_4(\mathbf{G}) - \psi_4\|^2]$, where $\|\cdot\|$ denotes the Euclidean norm. Theorem 1 in Chacón & Duong (2010) shows that the dominant part of that error is precisely given by $\mathrm{tr}(\Omega_4 + \boldsymbol{\omega}_4\boldsymbol{\omega}_4^\top)$, but with the unconvolved kernel L in the place of $\bar{L}$.

Therefore, despite the PI and SCV methods having very different target functions, it turns out that there is a deep synchronicity in the two approaches with respect to the problem of pilot bandwidth selection, which was not apparent at all in previous studies with restricted pilot bandwidth parametrizations, but had been noted in the univariate case (see Cao 1993).

This allows us to take advantage of the results already obtained by Chacón & Duong (2010), for example that for the choice of $\mathbf{G}$ the squared bias term dominates the variance term, so

$$\mathrm{AMSE}(\mathbf{G}) = \frac{1}{4} m_2(K)^4 \operatorname{tr}\left\{\boldsymbol{\omega}_4 \boldsymbol{\omega}_4^{\top} \left(\operatorname{vec} \mathbf{H}_{\mathrm{AMISE}} \operatorname{vec}^{\top} \mathbf{H}_{\mathrm{AMISE}} \otimes \mathbf{I}_{d^2}\right)\right\}\{1 + o(1)\}.$$

Moreover, it is clear that the choice of the bandwidth $\mathbf{G}_{\mathrm{AMSE}}$ minimizing the dominant part of the AMSE is unaffected by the terms involving $\mathbf{H}_{\mathrm{AMISE}}$, so from theorem 2 in Chacón & Duong (2010) it follows that $\mathbf{G}_{\mathrm{AMSE}}$ is of order $n^{-2/(d+6)}$ and the minimal MSE is of order $n^{-4/(d+6)}$.

The previous observation is useful for determining the relative convergence rate of $\hat{\mathbf{H}}_{\mathrm{SCV}}$ in a relatively straightforward way.

Theorem 3. *Suppose that the conditions of Theorems 1 and 2 hold. The relative rate of convergence of* $\hat{\mathbf{H}}_{\mathrm{SCV}}$ *to* $\mathbf{H}_{\mathrm{MISE}}$ *is* $n^{-2/(d+6)}$.

In the univariate case, Jones & Sheather (1991) noticed that the pilot bandwidth can be chosen to annihilate the two dominant bias terms in the estimation of ψ_4, leading to a better convergence rate, namely from $n^{-2/7}$ to $n^{-5/14}$. In the multivariate case, this would be equivalent to solving the system of equations $\boldsymbol{\omega}_4(\mathbf{G}) = 0$ for $\mathbf{G}$, but, as mentioned in Chacón & Duong (2010), this bias annihilation is not possible for general densities if $d \geq 2$. As an example, if we consider the bivariate density $f(\boldsymbol{x}) = \{\phi(\boldsymbol{x} - \boldsymbol{\mu}) + \phi(\boldsymbol{x} + \boldsymbol{\mu})\}/2$ with $\boldsymbol{\mu} = (1, 1)$, then the exact normal calculations in section 3.3 of Chacón & Duong (2010) can be used to obtain an explicit formula for $\boldsymbol{\omega}_4(\mathbf{G})$, and numerical minimization leads to $\min_{\mathbf{G} \in \mathcal{F}} \|\boldsymbol{\omega}_4(\mathbf{G})\|^2 = 4.35 \times 10^{-5} > 0$ for such a density when $n = 100$.

Furthermore, in Jones *et al.* (1991) this dominant bias annihilation is the key to obtaining a SCV bandwidth selector with relative convergence rate $n^{-1/2}$ in the univariate case, by carefully choosing the pilot bandwidth as a function of h. Because bias annihilation is not possible in the multivariate case in general, it is unlikely that such a fast rate could be achieved by using the same approach for higher dimensions.

On the other hand, the SCV selector of Duong & Hazelton (2005b) used a scalar parametrization $\mathbf{G} = g^2 \mathbf{I}_d$ for the pilot bandwidth. To use a scalar bandwidth with multivariate data, the usual procedure is to pre-transform the data so that all dimensions have (at least approximately) the same marginal dispersion; this is called pre-sphering. For unimodal distributions, pre-sphering does lead to an approximately spherically symmetric transformed distribution. In other cases, however, such as the separated bimodal example from Chacón & Duong (2010), pre-sphering does not achieve such a goal, even if the marginal variances are equal. In these latter cases, a scalar pilot with the pre-transformed data will be markedly less efficient than an unconstrained selector.

We can also use Theorem 2 to derive known results about scalar pilot selectors (such as those shown in Duong & Hazelton 2005b) in a very simple way. Substituting $\mathbf{G} = g^2 \mathbf{I}_d$ into the AMSE from Theorem 2 we obtain the following expression for the optimal scalar pilot bandwidth.

Corollary 1. *Suppose the conditions for Theorem 2 hold. If the pilot matrix is parametrized as* $\mathbf{G} = g^2 \mathbf{I}_d$, *then the AMSE simplifies to*

$$\mathrm{AMSE}(g) = \frac{1}{4} m_2(K)^4 \left(n^{-2} g^{-2d-8} A_1 + 2n^{-1} g^{-d-2} A_2 + g^4 A_3\right)\{1 + o(1)\},$$

where

$$A_1 = \mathsf{D}^{\otimes 4}\bar{L}(0)^\top \big(\operatorname{vec}\mathbf{H}_{\mathrm{AMISE}} \operatorname{vec}^\top \mathbf{H}_{\mathrm{AMISE}} \otimes \mathbf{I}_{d^2}\big)\mathsf{D}^{\otimes 4}\bar{L}(0),$$
$$A_2 = \mathsf{D}^{\otimes 4}\bar{L}(0)^\top \big(\operatorname{vec}\mathbf{H}_{\mathrm{AMISE}} \operatorname{vec}^\top \mathbf{H}_{\mathrm{AMISE}} \otimes \mathbf{I}_{d^2}\big)\big(\operatorname{vec}^\top \mathbf{I}_d \otimes \mathbf{I}_{d^4}\big)\psi_6,$$
$$A_3 = \psi_6^\top (\operatorname{vec}\mathbf{I}_d \otimes \mathbf{I}_{d^4})\big(\operatorname{vec}\mathbf{H}_{\mathrm{AMISE}} \operatorname{vec}^\top \mathbf{H}_{\mathrm{AMISE}} \otimes \mathbf{I}_{d^2}\big)\big(\operatorname{vec}^\top \mathbf{I}_d \otimes \mathbf{I}_{d^4}\big)\psi_6,$$

whose minimizer is given by

$$g_{\mathrm{AMSE}} = \left\{ \frac{2(d+4)A_1}{\big[-(d+2)A_2 + \big\{(d+2)^2A_2^2 + 8(d+4)A_1A_3\big\}^{1/2}\big]n} \right\}^{1/(d+6)}.$$

Next we show that this optimal pilot has the same form as the optimal scalar pilot from Duong & Hazelton (2005b), except for some constant matrix factors, as they arrange the sixth-order integrated derivatives into in a $d \times d$ matrix, whereas we characterize these derivatives as the vector ψ_6.

Corollary 2. *Suppose that the conditions for Theorem 2 hold and that the pilot matrix is parametrized as* $\mathbf{G} = g^2\mathbf{I}_d$. *Further suppose that* $K = L = \phi$ *and* $\mathbf{H}_{\mathrm{AMISE}} = \mathbf{C}n^{-2/(d+4)}$. *In this case the coefficients from Corollary 1 admit the alternative forms*

$$A_1 = \frac{9}{16}(4\pi)^{-d}n^{-4/(d+4)}(\operatorname{tr}\mathbf{C})\operatorname{vec}^\top(\mathbf{C}\otimes\mathbf{I}_d)\boldsymbol{\mathcal{S}}_{d,4}\operatorname{vec}\mathbf{I}_{d^2},$$
$$A_2 = \frac{3}{4}(4\pi)^{-d/2}n^{-4/(d+4)}(\operatorname{tr}\mathbf{C})\big(\operatorname{vec}^\top\mathbf{C}\otimes\operatorname{vec}^\top\mathbf{I}_d\big)\boldsymbol{\mathcal{S}}_{d,4}\big(\operatorname{vec}^\top\mathbf{I}_d\otimes\mathbf{I}_{d^4}\big)\psi_6,$$
$$A_3 = n^{-4/(d+4)}\psi_6^\top(\operatorname{vec}\mathbf{I}_d\otimes\mathbf{I}_{d^4})\big(\operatorname{vec}\mathbf{C}\operatorname{vec}^\top\mathbf{C}\otimes\mathbf{I}_{d^2}\big)\big(\operatorname{vec}^\top\mathbf{I}_d\otimes\mathbf{I}_{d^4}\big)\psi_6.$$

The equivalent coefficients for the optimal pilot selector from Duong & Hazelton (2005b) are

$$A_1' = \frac{1}{16}(4\pi)^{-d}n^{-4/(d+4)}(\operatorname{tr}\mathbf{C})\{4+(d+4)\operatorname{tr}\mathbf{C}\},$$
$$A_2' = \frac{1}{16}(4\pi)^{-d/2}n^{-4/(d+4)}\operatorname{vec}^\top\big[\{2\mathbf{C}^2+(\operatorname{tr}\mathbf{C})\mathbf{C}\}\otimes\mathbf{I}_{d^2}\big]\psi_6,$$
$$A_3' = \frac{1}{4}n^{-4/(d+4)}\psi_6^\top\big(\mathbf{I}_d\otimes\mathbf{C}^2\otimes\operatorname{vec}\mathbf{I}_{d^2}\operatorname{vec}^\top\mathbf{I}_{d^2}\big)\psi_6.$$

Remark 1. The previous methodology for selecting the pilot bandwidth $\mathbf{G}$ follows the guidelines of the univariate case as described in the paper by Hall *et al.* (1992). A different approach, suggested by an anonymous referee, could be based on the minimization of the SCV criterion over $\mathbf{G}$ and $\mathbf{H}$. This is an interesting idea, and is closely related to another univariate procedure, the double kernel-double h method, introduced in the L_1 context by Berlinet & Devroye (1994) and studied in the L_2 context by Jones (1998) and Abdous (1999). Minimizing SCV over $\mathbf{G}$ and $\mathbf{H}$ is equivalent to finding the closest pair of estimators $\hat{f}_{\mathbf{H}}$ and $\tilde{f}_{\mathbf{G}}$, so a different kernel L (usually a higher-order kernel) is needed for the estimator $\tilde{f}_{\mathbf{G}}$ in order to avoid degeneracy of the problem. Therefore, this would be a completely different procedure, and its study would be worth a separate paper.

4. Numerical results

4.1. Practical implementation of the new method

We took $K = L = \phi$ for all of our numerical work, owing to the simplification induced in the SCV criterion because of its good convolution properties. It is easy to show (Duong & Hazelton 2005b) that for this choice of K and L we have

$$\mathrm{SCV}(\mathbf{H}) = n^{-1}|\mathbf{H}|^{-1/2}(4\pi)^{-d/2} + n^{-2}\sum_{i,j=1}^{n}(\phi_{2\mathbf{H}+2\mathbf{G}} - 2\phi_{\mathbf{H}+2\mathbf{G}} + \phi_{2\mathbf{G}})(\mathbf{X}_i - \mathbf{X}_j).$$

For the practical implementation of the SCV bandwidth we propose a two-stage approach as in Chacón & Duong (2010), which can be described as follows.

(i) Compute

$$\hat{\psi}_8^{\mathrm{NR}} = \frac{8!}{4!2^{d+8}\pi^{d/2}}|\mathbf{S}|^{-1/2}\boldsymbol{\mathcal{S}}_{d,8}(\mathrm{vec}\,\mathbf{S}^{-1})^{\otimes 4},$$

which is the value of ψ_8 in the case where f is the $N(0, \boldsymbol{\Sigma})$ density, but with $\boldsymbol{\Sigma}$ replaced by $\mathbf{S}$, the sample covariance matrix. Plug this estimate in the formula of $\|\boldsymbol{\omega}_6\|^2$ and numerically minimize in $\mathbf{G} \in \mathcal{F}$ to obtain $\hat{\mathbf{G}}_6$.

(ii) Use $\mathbf{G} = \hat{\mathbf{G}}_6$ to compute

$$\hat{\psi}_6(\mathbf{G}) = n^{-2}\sum_{i,j=1}^{n}\mathsf{D}^{\otimes 6}\phi_{\mathbf{G}}(\mathbf{X}_i - \mathbf{X}_j),$$

plug $\hat{\psi}_6(\hat{\mathbf{G}}_6)$ in the formula of $\|\boldsymbol{\omega}_4\|^2$ and numerically minimize in $\mathbf{G} \in \mathcal{F}$ to obtain $\hat{\mathbf{G}}_4$.

(iii) Finally, employ $\mathbf{G} = \hat{\mathbf{G}}_4$ in the SCV criterion and numerically minimize in $\mathbf{H} \in \mathcal{F}$ to obtain $\hat{\mathbf{H}}_{\mathrm{SCV}}$.

4.2. Simulation study

In this section we explore the finite-sample performance of the unconstrained SCV selector in comparison to that of other selectors. The selectors that we compare are

(i) the plug-in with a scalar pilot from Duong & Hazelton (2003), labelled PIS.
(ii) the plug-in with an unconstrained pilot from Chacón & Duong (2010), labelled PIU.
(iii) the smoothed cross-validation with a scalar pilot from Duong & Hazelton (2005b), labelled SCVS.
(iv) our proposed smoothed cross-validation with an unconstrained pilot, labelled SCVU.
(v) the unconstrained cross-validation (UCV) of Sain, Baggerly & Scott (1994), labelled UCV; this selector does not rely on asymptotic arguments, and does not require a pilot bandwidth.

All these selectors are implemented in the ks library (Duong 2007) in R.

For the bivariate study, we examine six normal mixture target densities from Chacón (2009). Their contour plots are depicted in Figure 1. Target density '1' is a single normal density and so it can be considered a base case. Densities '2', '6', '7', '8' and '11' have

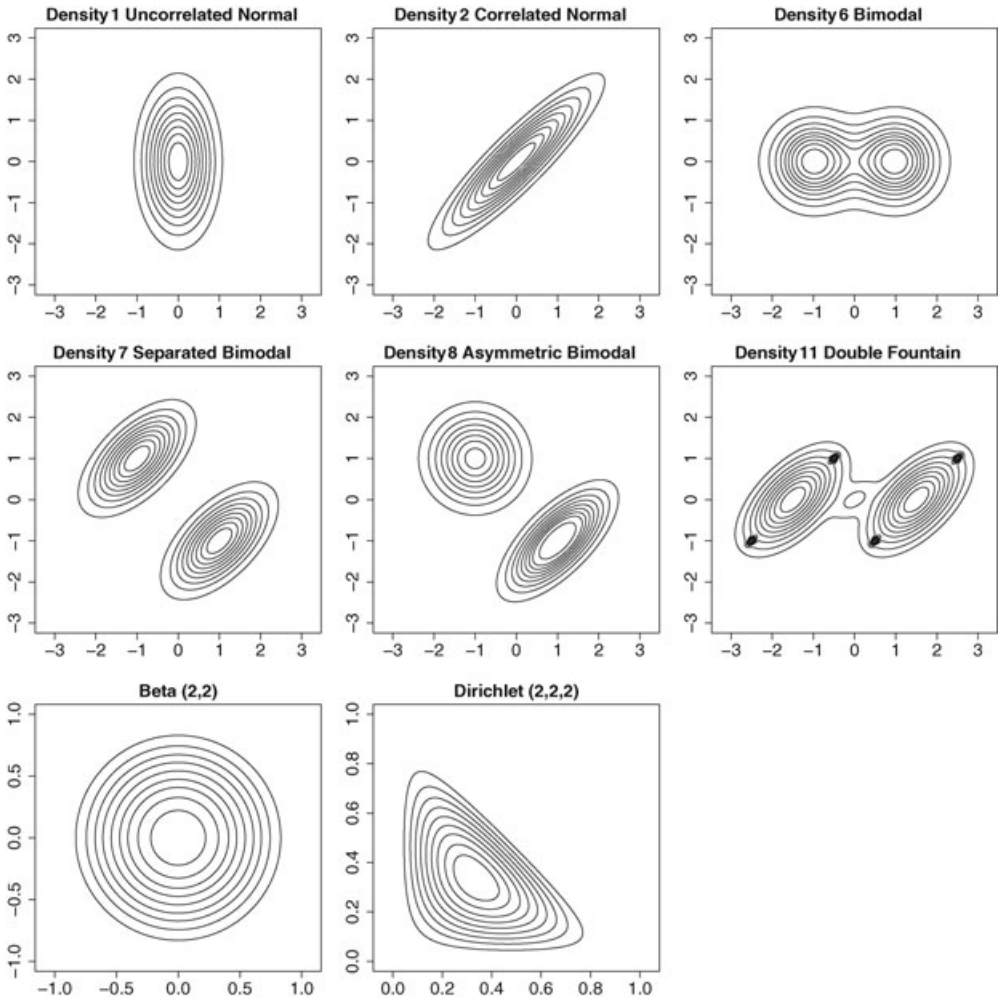

Figure 1. Contour plots for the bivariate target densities.

varying degrees of intricate structure. In addition to the normal mixture densities, we include two bivariate densities having bounded support, namely the symmetric beta density (see Devroye 1996, example 2 with $c = 2$, for random variate generation from this distribution) and the Dirichlet(2, 2, 2) distribution.

For each target density, we take 100 replicates for two representative sample sizes, $n = 100$ and $n = 1000$. To measure the accuracy of a bandwidth selection method, say $\hat{\mathbf{H}}$, we compute the integrated squared error of the kernel density estimate, $\mathrm{ISE}(\hat{\mathbf{H}}) = \int_{\mathbb{R}^d} \{\hat{f}_{n\hat{\mathbf{H}}}(\boldsymbol{x}) - f(\boldsymbol{x})\}^2 d\boldsymbol{x}$. The logarithms of the ISEs are given in Figures 2 and 3 for $n = 100$ and $n = 1000$, respectively.

The relative performances of the five selectors considered are similar for the two sample sizes, so the following comments apply to both situations, with $n = 100$ and $n = 1000$. Comparing the two versions of the SCV approach, we see that for densities '1', '2', '6', beta

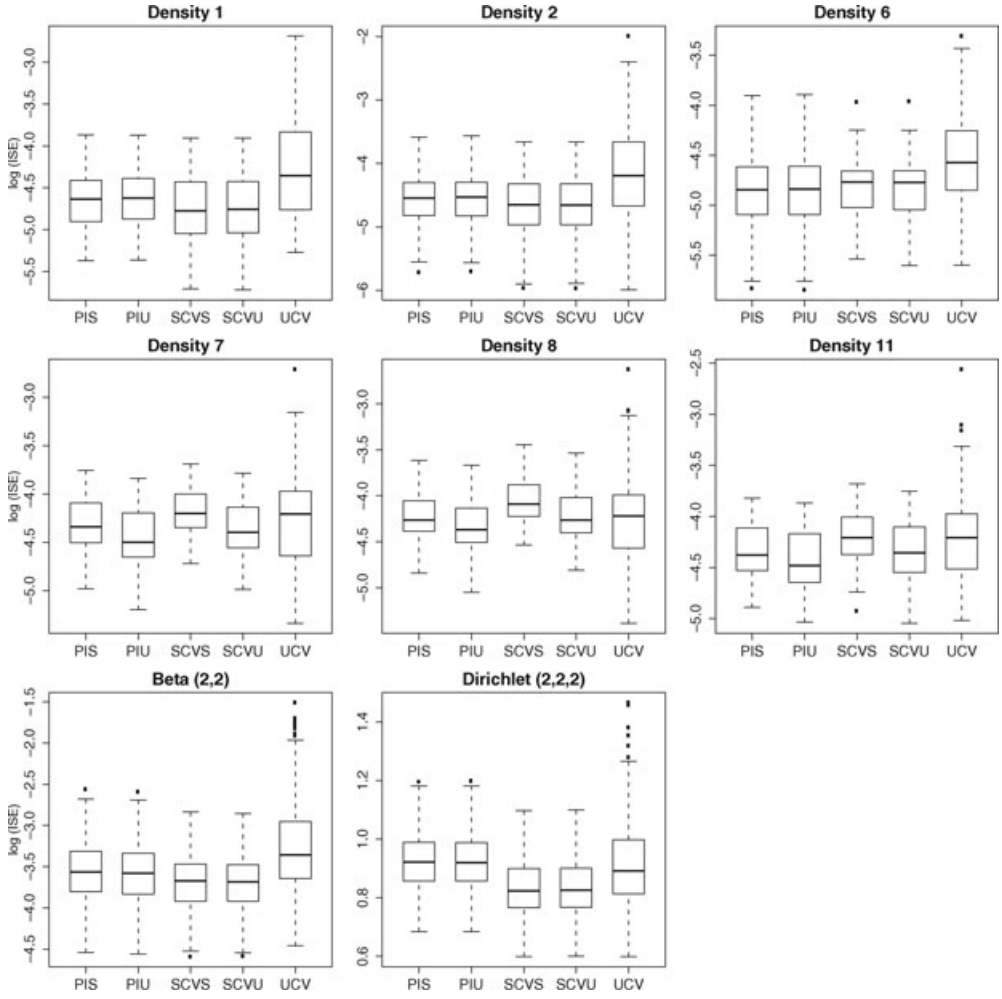

Figure 2. Boxplots for the logarithm of the integrated squared error of the kernel estimator using the plug-in bandwidth with a scalar pilot (PIS) and with an unconstrained pilot (PIU), the smoothed cross-validation bandwidth with a scalar pilot (SCVS) and with an unconstrained pilot (SCVU), and the unconstrained cross-validation bandwidth (UCV) for the bivariate target densities for sample size $n = 100$.

and Dirichlet, where we know that the constrained SCV pilot selectors are optimal, there is no loss in ISE performance when using unconstrained pilots. This indicates that our more general, unconstrained approach remains useful even in situations where simpler methods succeed. For the other three multimodal densities, however, the SCVU selector consistently outperforms SCVS, thus justifying the additional mathematical calculations involved in the new method. On the other hand, SCVU achieves the goal of having much less variability than UCV, and also outperforms UCV in terms of median ISE for all densities. These observations mirror the comparison between the scalar and unconstrained pilot selectors for plug-in selectors (Chacón & Duong 2010). When comparing SCV and plug-in, it is found that they both have quite similar performances. SCV is known to produce smoother density estimates (see Cao,

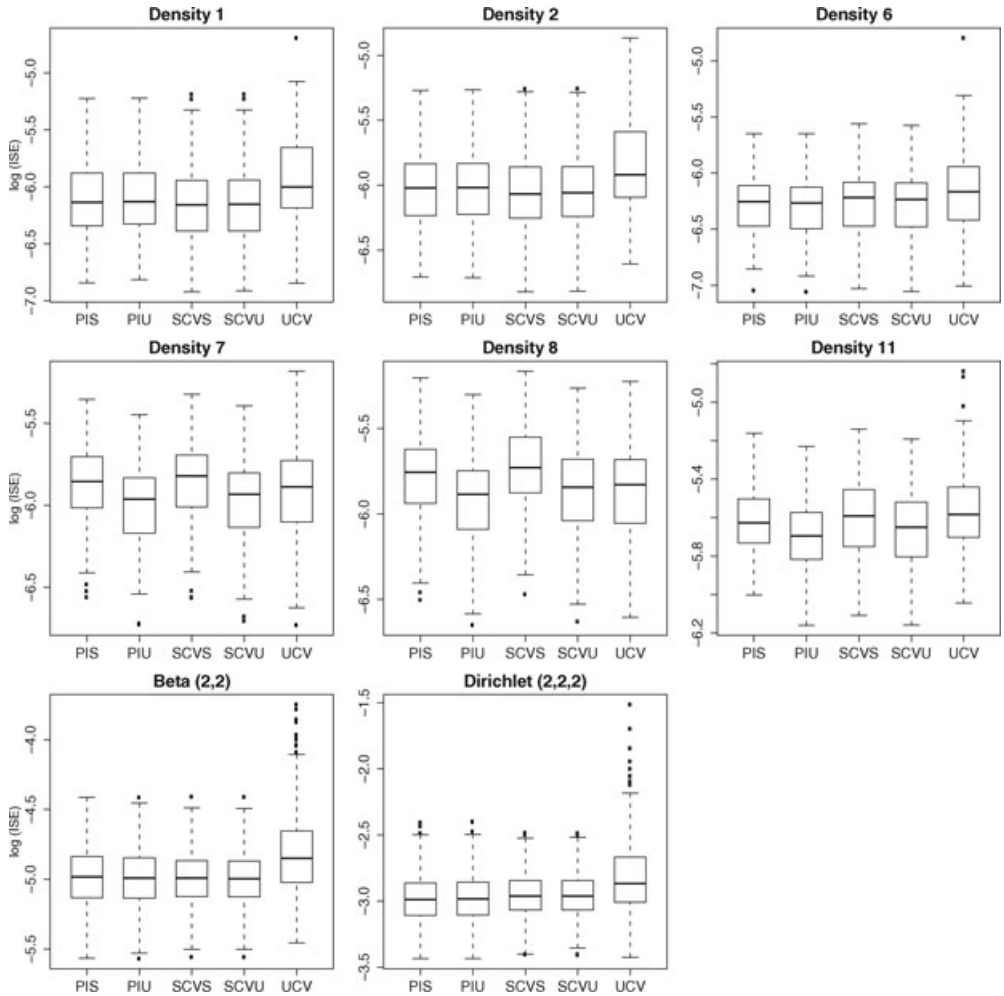

Figure 3. Boxplots for the logarithm of the integrated squared error of the kernel estimator using the plug-in bandwidth with a scalar pilot (PIS) and with an unconstrained pilot (PIU), the smoothed cross-validation bandwidth with a scalar pilot (SCVS) and with an unconstrained pilot (SCVU), and the unconstrained cross-validation bandwidth (UCV) for the bivariate target densities for sample size $n = 1000$.

Cuevas & González-Manteiga 1994), so it performs better for density functions with smoother features, for example densities '1', '2', beta and Dirichlet, and vice versa for sharper features, for example densities '6', '7', '8', '11', but with only a slight advantage of one method over the other.

For the multivariate study, we focus on the multi-dimensional generalization of target density '7' introduced by Chacón & Duong (2010). This density is an equal two-component normal mixture $\frac{1}{2}N(\Lambda(d, 0, \ldots, 0)^\top, \Lambda\mathbf{\Pi}\Lambda^\top) + \frac{1}{2}N(\Lambda(-d, 0, \ldots, 0)^\top, \Lambda\mathbf{\Pi}\Lambda^\top)$, where $\Lambda = \Lambda_d\Lambda_{d-1}\cdots\Lambda_2$, with Λ_i the 45° rotation matrix in the plane of $\mathbb{R}^d$ defined by the coordinates x_1 and x_i, and $\mathbf{\Pi} = \text{diag}(4^{-(d-1)}, 4^{-(d-2)}, \ldots, 4^{-1}, 1)$. Each bivariate projection of this density thus consists of a separated bimodal density. The ISEs for $d = 2, 3, 4$ are

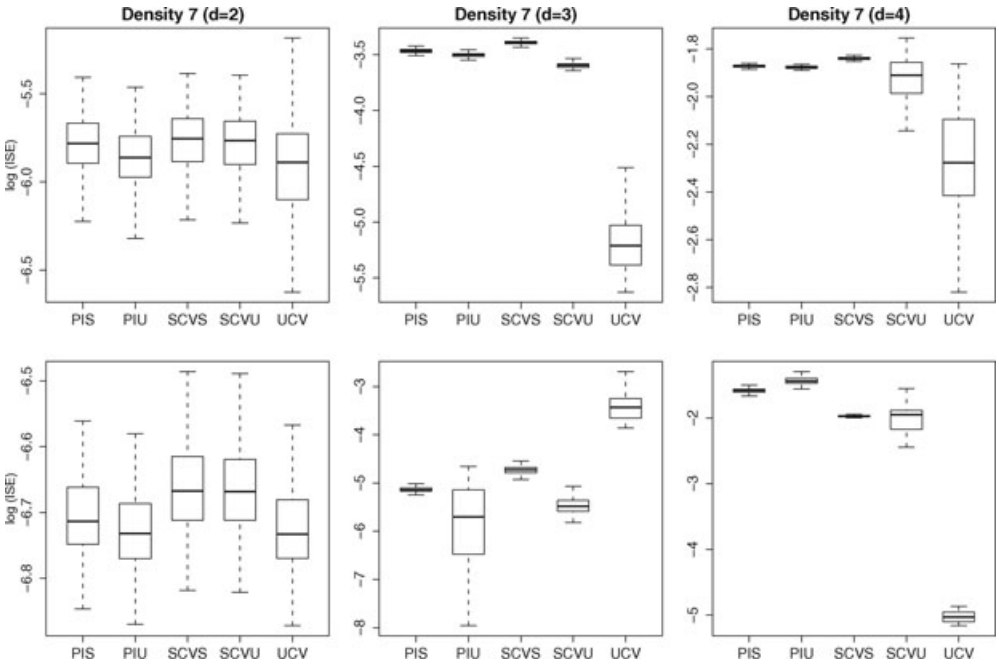

Figure 4. Boxplots for the logarithm of the integrated squared error of the kernel estimator using the plug-in bandwidth with a scalar pilot (PIS) and with an unconstrained pilot (PIU), the smoothed cross-validation bandwidth with a scalar pilot (SCVS) and with an unconstrained pilot (SCVU), and the unconstrained cross-validation bandwidth (UCV) for the multivariate target densities for sample size $n = 1000$ (upper row) and $n = 10\,000$ (lower row), for $d = 2, 3, 4$.

shown in Figure 4 for $n = 1000$ and $n = 10\,000$. As in the bivariate case, SCVU is uniformly better than SCVS. Moreover, the marginal improvement of the unconstrained pilots over scalar pilots seems to increase with d, providing additional justification for the study of unconstrained pilot selectors. On comparing SCVU with its unconstrained plug-in counterpart PIU, the former outperforms the latter in all cases except $d = 3$ and $n = 10\,000$, although we note for this case that, even though the median ISEs are similar, the SCVU is considerably less variable. Comparing SCVU with UCV, the situation is more blurred, with UCV generally exhibiting lower median ISEs coupled with more dispersed ISEs, although for $d = 3$ and $n = 10\,000$ SCVU outperforms UCV. This variable ISE performance of UCV has been noted previously, for example by Scott (1992, pp. 166–170).

4.3. Real data analysis

A key feature of eukaryotic cells is the compartmentalization of cellular functions into complex, membrane-surrounded organelles. The advent of modern fluorescent microscopy technologies has allowed for the visualization of a variety of these sub-cellular organelles, using fluorescent markers attached to proteins of interest, which serve as a useful proxy for studying organelles and their behaviour. The 3-dimensional spatial organization of these organelles poses difficult challenges for their analysis, thus requiring statistical approaches historically not used in cell biology.

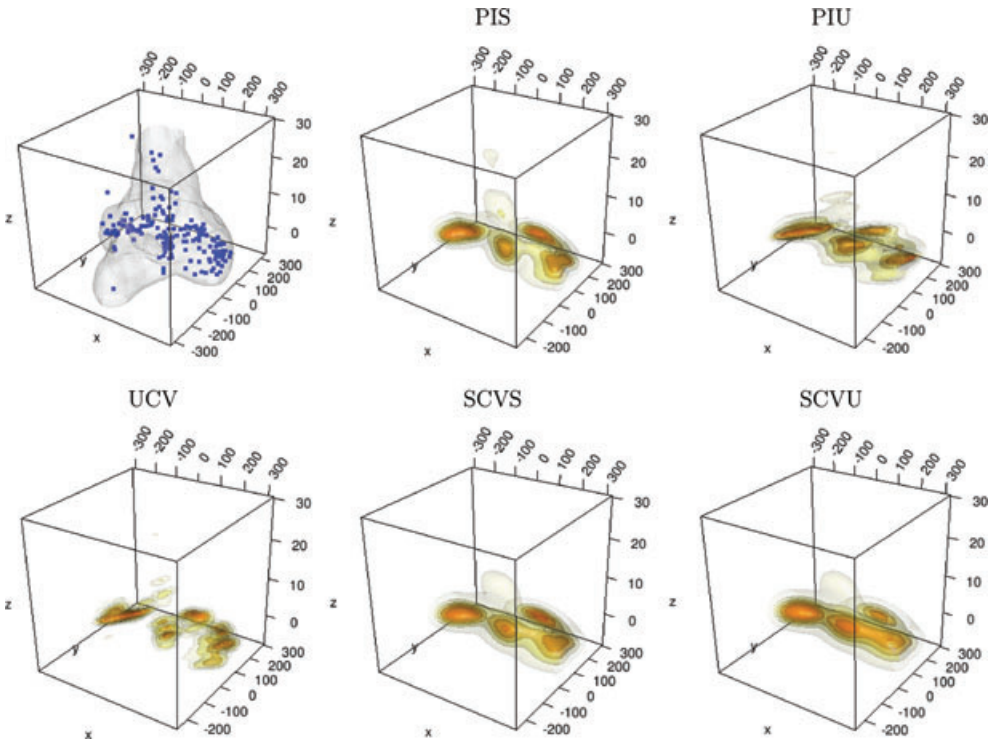

Figure 5. Density estimates of Rab6-positive organelles from microscopy image data. The scatter plot is the upper left plot, with the approximate cell outline as the grey shell; the constrained and unconstrained plug-in pilots (PIS, PIU) are the other two plots in the upper row; the unbiased cross-validation and the constrained and unconstrained SCV pilots are in the lower row (UCV, SCVS, SCVU). The plotted contours are the 10%, 30%, 50% 70% and 90% levels. The axes indicate pixels from the microscopy images.

In a recent paper, Schauer *et al.* (2010) showed that density estimation is a powerful technique for studying global cellular organization and for the quantification of cell biology. The data set that we consider here is taken from these authors, and consists of the trivariate locations of 181 organelles with a Rab6-positive fluorescence signal detected from a single cell. We compute each of the five bandwidths considered in the previous simulation study:

$$\hat{\mathbf{H}}_{\text{PIS}} = \begin{bmatrix} 2455.07 & -85.3 & 0.1 \\ -85.3 & 777.3 & -12.4 \\ 0.1 & -12.4 & 3.0 \end{bmatrix}, \hat{\mathbf{H}}_{\text{PIU}} = \begin{bmatrix} 6072.53 & 82.9 & 149.0 \\ 82.9 & 1465.5 & -31.9 \\ 149.0 & -31.9 & 5.4 \end{bmatrix},$$

$$\hat{\mathbf{H}}_{\text{SCVS}} = \begin{bmatrix} 4484.8 & -254.5 & 0.6 \\ -254.5 & 1163.9 & -16.3 \\ 0.6 & -16.3 & 3.9 \end{bmatrix}, \hat{\mathbf{H}}_{\text{SCVU}} = \begin{bmatrix} 6089.4 & 80.0 & -6.3 \\ 80.0 & 1470.5 & -33.3 \\ -6.3 & -33.3 & 7.5 \end{bmatrix},$$

$$\hat{\mathbf{H}}_{\text{UCV}} = \begin{bmatrix} 885.8 & 293.4 & 13.5 \\ 293.4 & 892.8 & -1.8 \\ 13.5 & -1.8 & 0.8 \end{bmatrix}$$

and the resulting density estimates are displayed in Figure 5. The constrained pilot estimates (PIS and SCVS) are similar to each other. The PIU estimate emphasizes the oblique orientation of the data, whereas the SCVU estimate exhibits less marked trimodality in the central region. The UCV selector yields the noisiest density estimate, indicating that undersmoothing is quite likely here. Schauer *et al.* (2010) note that Rab6-positive endosomes move along from the Golgi to the cell periphery. The Golgi is a larger endosome that is located at approximately (0, 100) in the (x, y)-plane, and the cell periphery is roughly demarcated by the grey contour shell in the figure. As expected, these density estimates exhibit local peaks in density near the Golgi and the cell periphery. The UCV estimates were considered to be too variable when applied to other similar experimental data (not shown). The SCVU estimate highlights a smoother density of Rab6-positive endosomes from the Golgi to the periphery, compared with the more clustered densities for the SCVS, PIU and PIS estimates. The results are leading to further studies to quantify this movement, using density estimation at discrete snapshots in time-lapse experiments.

Appendix: Proofs

In the following proofs, integrals without any integration limits are integrated over the appropriate Euclidean space.

Proof of Theorem 1. Using previously established results, for example Duong & Hazelton (2005b), we can show that $\operatorname{vec}(\mathbf{H}_{\mathrm{MISE2}} - \mathbf{H}_{\mathrm{MISE}})$ is of the same order as

$$\mathsf{D}_{\mathbf{H}}(\mathrm{MISE2} - \mathrm{MISE})(\mathbf{H}_{\mathrm{MISE}}) = n^{-1}\mathsf{D}_{\mathbf{H}}\{R^*(\bar{K}_{\mathbf{H}_{\mathrm{MISE}}}, f)\},$$

where $\mathsf{D}_{\mathbf{H}} = \partial/(\partial \operatorname{vec} \mathbf{H})$ denotes the gradient operator with respect to $\operatorname{vec} \mathbf{H}$. Notice that a change of variables allows us to write $R^*(\bar{K}_{\mathbf{H}}, f) = \int \bar{K}(\boldsymbol{x}) p(\mathbf{H}^{1/2}\boldsymbol{x}) d\boldsymbol{x}$, where the function $p = f * f_-$, with $f_-(\boldsymbol{x}) = f(-\boldsymbol{x})$, is such that $\mathsf{D}^{\otimes r} p(0) = \psi_r$. Moreover, under the smoothness conditions on f and K, we can commute $\mathsf{D}_{\mathbf{H}}$ and the integral appearing in the R^* functional, so that $\mathsf{D}_{\mathbf{H}}\{R^*(\bar{K}_{\mathbf{H}}, f)\} = \int \bar{K}(\boldsymbol{x}) \mathsf{D}_{\mathbf{H}}\{p(\mathbf{H}^{1/2}\boldsymbol{x})\} d\boldsymbol{x}$. Next, if we denote $\rho(\mathbf{H}) = \mathbf{H}^{1/2} \otimes \mathbf{I}_d + \mathbf{I}_d \otimes \mathbf{H}^{1/2}$ then straightforward matrix differential calculus (see Magnus & Neudecker 1999) gives the differential

$$d\{p(\mathbf{H}^{1/2}\boldsymbol{x})\} = \mathsf{D}p(\mathbf{H}^{1/2}\boldsymbol{x})^{\top}\big(\boldsymbol{x}^{\top} \otimes \mathbf{I}_d\big)\rho(\mathbf{H})^{-1} d \operatorname{vec} \mathbf{H},$$

which yields $\mathsf{D}_{\mathbf{H}}\{R^*(\bar{K}_{\mathbf{H}}, f)\} = \rho(\mathbf{H})^{-1} \int \bar{K}(\boldsymbol{x})(\boldsymbol{x} \otimes \mathbf{I}_d) \mathsf{D}p(\mathbf{H}^{1/2}\boldsymbol{x}) d\boldsymbol{x}$. Now using the Taylor expansion of the vector-valued function $\boldsymbol{x} \mapsto \mathsf{D}p(\mathbf{H}^{1/2}\boldsymbol{x})$ around $\boldsymbol{x} = \mathbf{0}$, in the formulation of Chacón *et al.* (2010), it follows that

$$\begin{aligned}&\mathsf{D}_{\mathbf{H}}\{R^*(\bar{K}_{\mathbf{H}}, f)\}\\ &\quad = \rho(\mathbf{H})^{-1}\left[\sum_{j=0}^{1} \int \bar{K}(\boldsymbol{x})(\boldsymbol{x} \otimes \mathbf{I}_d)\big\{\mathbf{I}_d \otimes (\boldsymbol{x}^{\top}\mathbf{H}^{1/2})^{\otimes j}\big\} d\boldsymbol{x}\right] \mathsf{D}^{\otimes(j+1)} p(0)\{1 + o(1)\}.\end{aligned}$$

Because the first of the two summands in the previous expression vanishes, owing to the symmetry of $\bar{K}$, after some more matrix manipulations we finally obtain that $\mathsf{D}_{\mathbf{H}}\{R^*(\bar{K}_{\mathbf{H}}, f)\}$ is asymptotically equivalent to $m_2(\bar{K})\rho(\mathbf{H})^{-1}(\mathbf{I}_d \otimes \mathbf{H}^{1/2})\psi_2$, which is a bounded sequence.

Thus, the components of the vector vec $(\mathbf{H}_{\mathrm{MISE2}} - \mathbf{H}_{\mathrm{MISE}})$ are of order n^{-1}, so recalling that $\mathbf{H}_{\mathrm{MISE}} = O(n^{-2/(d+4)}\mathbf{J}_d)$ the proof is complete. □

In the following, we define the vector raw moment of f as $\boldsymbol{\mu}_r(f) = \int \boldsymbol{x}^{\otimes r} f(\boldsymbol{x})\, d\boldsymbol{x}$ (see Holmquist 1988, or Jammalamadaka, Rao & Terdik 2006). This vector moment is just the vectorization of the usual matrix moment for $r = 2$; that is, $\boldsymbol{\mu}_2(f) = \mathrm{vec} \int \boldsymbol{x}\boldsymbol{x}^\top f(\boldsymbol{x})\, d\boldsymbol{x}$.

We also denote by $\boldsymbol{\mathcal{S}}_{d,r}$ the symmetrizer matrix of order r (see Holmquist 1985). This matrix is characterized by the fact that pre-multiplying a Kronecker product of any r vectors by $\boldsymbol{\mathcal{S}}_{d,r}$ results in the normalized sum of all possible permutations of the r-fold product; for example, for three d-vectors $\boldsymbol{x}_1$, $\boldsymbol{x}_2$ and $\boldsymbol{x}_3$, we have $\boldsymbol{\mathcal{S}}_{d,3}(\boldsymbol{x}_1 \otimes \boldsymbol{x}_2 \otimes \boldsymbol{x}_3) = \frac{1}{3!}(\boldsymbol{x}_1 \otimes \boldsymbol{x}_2 \otimes \boldsymbol{x}_3 + \boldsymbol{x}_1 \otimes \boldsymbol{x}_3 \otimes \boldsymbol{x}_2 + \boldsymbol{x}_2 \otimes \boldsymbol{x}_1 \otimes \boldsymbol{x}_3 + \boldsymbol{x}_2 \otimes \boldsymbol{x}_3 \otimes \boldsymbol{x}_1 + \boldsymbol{x}_3 \otimes \boldsymbol{x}_1 \otimes \boldsymbol{x}_2 + \boldsymbol{x}_3 \otimes \boldsymbol{x}_2 \otimes \boldsymbol{x}_1)$.

Using formula (7.4) in Holmquist (1996), which gives the vector binomial expansion $(\boldsymbol{x} + \boldsymbol{y})^{\otimes r} = \boldsymbol{\mathcal{S}}_{d,r} \sum_{j=0}^{r} \binom{r}{j} \boldsymbol{x}^{\otimes j} \otimes \boldsymbol{y}^{\otimes(r-j)}$ for any $\boldsymbol{x}$, $\boldsymbol{y} \in \mathbb{R}^d$, it is straightforward to prove that the $2r$th moment of the self-convolution $\bar{\ell} = \ell * \ell$ of a symmetric function $\ell : \mathbb{R}^d \to \mathbb{R}$ can be calculated as

$$\boldsymbol{\mu}_{2r}(\bar{\ell}) = \boldsymbol{\mathcal{S}}_{d,2r} \sum_{j=0}^{r} \binom{2r}{2j} \boldsymbol{\mu}_{2j}(\ell) \otimes \boldsymbol{\mu}_{2r-2j}(\ell).$$

Thus, for the particular case of $\ell = \Delta$ we obtain $\boldsymbol{\mu}_j(\bar{\Delta}) = 0$ for $j = 0, 1, 2, 3$ and $\boldsymbol{\mu}_4(\bar{\Delta}) = \boldsymbol{\mu}_4(\bar{K}) - 2\boldsymbol{\mu}_4(K) = 6\,\boldsymbol{\mathcal{S}}_{d,4}\boldsymbol{\mu}_2(K)^{\otimes 2} = 6m_2(K)^2\boldsymbol{\mathcal{S}}_{d,4}(\mathrm{vec}\,\mathbf{I}_d)^{\otimes 2}$.

The proof of Theorem 2 needs two separate lemmas, which we state and prove next.

Lemma 1. *Under the conditions of Theorem 2,*

$$\mathbb{E}\,\mathrm{SCV}(\mathbf{H}) - \mathrm{MISE2}(\mathbf{H}) = \frac{1}{4} m_2(K)^2 \boldsymbol{\omega}_4^\top (\mathrm{vec}\,\mathbf{H})^{\otimes 2}\{1 + o(1)\}.$$

Proof. From the definitions,

$$\begin{aligned}
&\mathbb{E}\,\mathrm{SCV}(\mathbf{H}) - \mathrm{MISE2}(\mathbf{H}) \\
&\quad = n^{-1}\bar{\Delta}_{\mathbf{H}} * \bar{L}_{\mathbf{G}}(0) + (1 - n^{-1})\mathbb{E}\{\bar{\Delta}_{\mathbf{H}} * \bar{L}_{\mathbf{G}}(\mathbf{X}_1 - \mathbf{X}_2)\} - \int (\bar{\Delta}_{\mathbf{H}} * f)(\boldsymbol{x}) f(\boldsymbol{x})\, d\boldsymbol{x}.
\end{aligned}$$

For the difference between the second and third terms, denoting by L_0 the Dirac delta function here, we have

$$\begin{aligned}
&\mathbb{E}\{\bar{\Delta}_{\mathbf{H}} * \bar{L}_{\mathbf{G}}(\mathbf{X}_1 - \mathbf{X}_2)\} - \int (\bar{\Delta}_{\mathbf{H}} * f)(\boldsymbol{x}) f(\boldsymbol{x})\, d\boldsymbol{x} \\
&\quad = \int \{\bar{\Delta}_{\mathbf{H}} * (\bar{L}_{\mathbf{G}} - L_0) * f\}(\boldsymbol{x}) f(\boldsymbol{x})\, d\boldsymbol{x} \\
&\quad = \iiint \bar{\Delta}(\boldsymbol{y})(\bar{L} - L_0)(z) f(\boldsymbol{x} - \mathbf{H}^{1/2}\boldsymbol{y} - \mathbf{G}^{1/2}z) f(\boldsymbol{x})\, d\boldsymbol{x}\, d\boldsymbol{y}\, dz.
\end{aligned}$$

To simplify this, we note that $\mu_0(\bar{L} - L_0) = 0$, $\boldsymbol{\mu}_1(\bar{L} - L_0) = 0$ and $\boldsymbol{\mu}_2(\bar{L} - L_0) = \boldsymbol{\mu}_2(\bar{L}) = 2\boldsymbol{\mu}_2(L) = 2m_2(L)\,\mathrm{vec}\,\mathbf{I}_d$, so taking into account the Taylor expansion

$$f(\boldsymbol{x} - \mathbf{H}^{1/2}\boldsymbol{y} - \mathbf{G}^{1/2}z) = \sum_{j=0}^{2} \frac{(-1)^j}{j!} (z^\top \mathbf{G}^{1/2})^{\otimes j} \mathsf{D}^{\otimes j} f(\boldsymbol{x} - \mathbf{H}^{1/2}\boldsymbol{y})\{1 + o(1)\}$$

and integrating with respect to z we obtain

$$\begin{aligned}
&\mathbb{E}\{\bar{\Delta}_{\mathbf{H}} * \bar{L}_{\mathbf{G}}(\mathbf{X}_1 - \mathbf{X}_2)\} - \int (\bar{\Delta}_{\mathbf{H}} * f) f(\boldsymbol{x}) f(\boldsymbol{x})\, d\boldsymbol{x} \\
&\quad = \iint \bar{\Delta}(\boldsymbol{y})\big\{\boldsymbol{\mu}_2(L)^\top (\mathbf{G}^{1/2})^{\otimes 2} \mathsf{D}^{\otimes 2} f(\boldsymbol{x} - \mathbf{H}^{1/2}\boldsymbol{y})\big\} f(\boldsymbol{x})\{1 + o(1)\}\, d\boldsymbol{x} d\boldsymbol{y} \\
&\quad = m_2(L) \mathrm{vec}^\top \mathbf{G} \iint \bar{\Delta}(\boldsymbol{y}) \mathsf{D}^{\otimes 2} f(\boldsymbol{x} - \mathbf{H}^{1/2}\boldsymbol{y}) f(\boldsymbol{x})\{1 + o(1)\}\, d\boldsymbol{x} d\boldsymbol{y},
\end{aligned}$$

where we have used $(\mathrm{vec}^\top \mathbf{I}_d)(\mathbf{G}^{1/2})^{\otimes 2} = \mathrm{vec}^\top \mathbf{G}$. Next we apply the Taylor expansion of the vector-valued function $\mathsf{D}^{\otimes 2} f(\boldsymbol{x} - \mathbf{H}^{1/2}\boldsymbol{y})$, for example from Chacón *et al.* (2010),

$$\mathsf{D}^{\otimes 2} f(\boldsymbol{x} - \mathbf{H}^{1/2}\boldsymbol{y}) = \sum_{j=0}^{4} \frac{(-1)^j}{j!} \big\{\mathbf{I}_{d^2} \otimes (\boldsymbol{y}^\top \mathbf{H}^{1/2})^{\otimes j}\big\} \mathsf{D}^{\otimes (j+2)} f(\boldsymbol{x})\{1 + o(\mathrm{vec}\, \mathbf{J}_d)\}.$$

Because the first three moments of $\bar{\Delta}$ are zero and $\boldsymbol{\mu}_4(\bar{\Delta}) = 6 m_2(K)^2 \boldsymbol{\mathcal{S}}_{d,4} (\mathrm{vec}\, \mathbf{I}_d)^{\otimes 2}$, we arrive at

$$\begin{aligned}
&\mathbb{E}\{\bar{\Delta}_{\mathbf{H}} * \bar{L}_{\mathbf{G}}(\mathbf{X}_1 - \mathbf{X}_2)\} - \int (\bar{\Delta}_{\mathbf{H}} * f)(\boldsymbol{x}) f(\boldsymbol{x})\, d\boldsymbol{x} \\
&\quad = \frac{1}{4!} m_2(L) \mathrm{vec}^\top \mathbf{G} \big[\mathbf{I}_{d^2} \otimes \big\{\boldsymbol{\mu}_4(\bar{\Delta})^\top (\mathbf{H}^{1/2})^{\otimes 4}\big\}\big] \int \mathsf{D}^{\otimes 6} f(\boldsymbol{x}) f(\boldsymbol{x})\, d\boldsymbol{x}\{1 + o(1)\} \\
&\quad = \frac{1}{4} m_2(K)^2 m_2(L) (\mathrm{vec}^\top \mathbf{G}) \big[\mathbf{I}_{d^2} \otimes \big\{\big(\mathrm{vec}^\top \mathbf{I}_d\big)^{\otimes 2} \boldsymbol{\mathcal{S}}_{d,4} (\mathbf{H}^{1/2})^{\otimes 4}\big\}\big] \psi_6 \{1 + o(1)\} \\
&\quad = \frac{1}{4} m_2(K)^2 m_2(L) \big[(\mathrm{vec}^\top \mathbf{G}) \otimes \{(\mathrm{vec}^\top \mathbf{H})^{\otimes 2}\}\big] \psi_6 \{1 + o(1)\} \\
&\quad = \frac{1}{8} m_2(K)^2 m_2(\bar{L}) \psi_6^\top [(\mathrm{vec}\, \mathbf{G}) \otimes \mathbf{I}_{d^4}] (\mathrm{vec}\, \mathbf{H})^{\otimes 2} \{1 + o(1)\},
\end{aligned}$$

where we have used $\boldsymbol{\mathcal{S}}_{d,4} (\mathbf{H}^{1/2})^{\otimes 4} = (\mathbf{H}^{1/2})^{\otimes 4} \boldsymbol{\mathcal{S}}_{d,4}$ and $(\mathbf{I}_{d^2} \otimes \boldsymbol{\mathcal{S}}_{d,4}) \mathsf{D}^{\otimes 6} = \mathsf{D}^{\otimes 6}$ (see Schott 2003).

For the other summand, $\bar{\Delta}_{\mathbf{H}} * \bar{L}_{\mathbf{G}}(0) = |\mathbf{G}|^{-1/2} \int \bar{\Delta}(\boldsymbol{z}) \bar{L}(\mathbf{G}^{-1/2}\mathbf{H}^{1/2}\boldsymbol{z}) d\boldsymbol{z}$, so that a Taylor expansion of order 4 of $\bar{L}$ around $\boldsymbol{z} = \mathbf{0}$ gives

$$\begin{aligned}
\bar{\Delta}_{\mathbf{H}} * \bar{L}_{\mathbf{G}}(0) &= \frac{1}{4!} |\mathbf{G}|^{-1/2} \boldsymbol{\mu}_4(\bar{\Delta})^\top (\mathbf{H}^{1/2})^{\otimes 4} (\mathbf{G}^{-1/2})^{\otimes 4} \mathsf{D}^{\otimes 4} \bar{L}(0)\{1 + o(1)\} \\
&= \frac{1}{4} m_2(K)^2 |\mathbf{G}|^{-1/2} (\mathrm{vec}^\top \mathbf{H})^{\otimes 2} (\mathbf{G}^{-1/2})^{\otimes 4} \mathsf{D}^{\otimes 4} \bar{L}(0)\{1 + o(1)\}
\end{aligned}$$

and the proof is complete. □

Lemma 2. *Under the conditions of Theorem 2,*

$$\mathrm{var}\, \mathsf{D}_{\mathbf{H}} \mathrm{SCV}(\mathbf{H}) = \frac{1}{4} m_2(K)^4 \big(\mathrm{vec}^\top \mathbf{H} \otimes \mathbf{I}_{d^2}\big) \Omega_4 (\mathrm{vec}\, \mathbf{H} \otimes \mathbf{I}_{d^2}) \{1 + o(1)\}.$$

Proof. By standard U-statistics theory, we are interested in the dominant terms of

$$\operatorname{var} \mathsf{D}_{\mathbf{H}}\mathrm{SCV}(\mathbf{H}) = \operatorname{var}\left\{ n^{-2} \sum_{i \neq j}^{n} (\mathsf{D}_{\mathbf{H}} \bar{\Delta}_{\mathbf{H}}) * \bar{L}_{\mathbf{G}}(\mathbf{X}_i - \mathbf{X}_j) \right\} \sim 4n^{-1}(\Xi_1 - \Xi_0) + 2n^{-2}\Xi_2,$$

where

$$\begin{aligned}
\Xi_1 &= \mathbb{E}\left\{ (\mathsf{D}_{\mathbf{H}} \bar{\Delta}_{\mathbf{H}}) * \bar{L}_{\mathbf{G}}(\mathbf{X}_1 - \mathbf{X}_2)(\mathsf{D}_{\mathbf{H}} \bar{\Delta}_{\mathbf{H}}) * \bar{L}_{\mathbf{G}}(\mathbf{X}_1 - \mathbf{X}_3)^{\top} \right\} \\
\Xi_2 &= \mathbb{E}\left\{ (\mathsf{D}_{\mathbf{H}} \bar{\Delta}_{\mathbf{H}}) * \bar{L}_{\mathbf{G}}(\mathbf{X}_1 - \mathbf{X}_2)(\mathsf{D}_{\mathbf{H}} \bar{\Delta}_{\mathbf{H}}) * \bar{L}_{\mathbf{G}}(\mathbf{X}_1 - \mathbf{X}_2)^{\top} \right\} \\
\Xi_0 &= \mathbb{E}\{ (\mathsf{D}_{\mathbf{H}} \bar{\Delta}_{\mathbf{H}}) * \bar{L}_{\mathbf{G}}(\mathbf{X}_1 - \mathbf{X}_2) \} \mathbb{E}\{ (\mathsf{D}_{\mathbf{H}} \bar{\Delta}_{\mathbf{H}}) * \bar{L}_{\mathbf{G}}(\mathbf{X}_1 - \mathbf{X}_2) \}^{\top}.
\end{aligned}$$

Regarding Ξ_0, taking into account the moment properties of $\bar{\Delta}$ it follows that

$$\begin{aligned}
\mathbb{E}\{ \bar{\Delta}_{\mathbf{H}} * \bar{L}_{\mathbf{G}}(\mathbf{X}_1 - \mathbf{X}_2) \} &= \iint \bar{\Delta}(\boldsymbol{y}) f(\boldsymbol{x} - \mathbf{H}^{1/2}\boldsymbol{y}) f(\boldsymbol{x}) \{1 + o(1)\} \, d\boldsymbol{x} d\boldsymbol{y} \\
&= \frac{1}{4} m_2(K)^2 (\operatorname{vec}^{\top} \mathbf{H})^{\otimes 2} \psi_4 \{1 + o(1)\}.
\end{aligned}$$

It is straightforward to obtain the differential $d(\operatorname{vec} \mathbf{H})^{\otimes 2} = \Gamma_2(\operatorname{vec} \mathbf{H} \otimes \mathbf{I}_{d^2}) d \operatorname{vec} \mathbf{H}$, where $\Gamma_2 = \mathbf{I}_{d^2} + \mathbf{K}_{d^2,d^2}$, with $\mathbf{K}_{m,n}$ denoting the commutation matrix of order $mn \times mn$ (Magnus & Neudecker 1979). So taking into account that $\Gamma_2^{\top} \mathsf{D}^{\otimes 4} = 2\mathsf{D}^{\otimes 4}$ (hence $\Gamma_2^{\top} \psi_4 = 2\psi_4$) and swapping the order of expectation and differentiation,

$$\mathbb{E}\{ (\mathsf{D}_{\mathbf{H}} \bar{\Delta}_{\mathbf{H}}) * \bar{L}_{\mathbf{G}}(\mathbf{X}_1 - \mathbf{X}_2) \} = \frac{1}{2} m_2(K)^2 \left(\operatorname{vec}^{\top} \mathbf{H} \otimes \mathbf{I}_{d^2} \right) \psi_4 \{1 + o(1)\}.$$

Overall,

$$\Xi_0 = \frac{1}{4} m_2(K)^4 \left(\operatorname{vec}^{\top} \mathbf{H} \otimes \mathbf{I}_{d^2} \right) \psi_4 \psi_4^{\top} (\operatorname{vec} \mathbf{H} \otimes \mathbf{I}_{d^2}) \{1 + o(1)\}.$$

For the other two matrices Ξ_1 and Ξ_2, using the Taylor expansion

$$\bar{L}(\mathbf{G}^{-1/2}\boldsymbol{x} - \mathbf{G}^{-1/2}\mathbf{H}^{1/2}\boldsymbol{z}) = \sum_{j=0}^{4} \frac{(-1)^j}{j!} (\boldsymbol{z}^{\top} \mathbf{H}^{1/2} \mathbf{G}^{-1/2})^{\otimes j} \mathsf{D}^{\otimes j} \bar{L}(\mathbf{G}^{-1/2}\boldsymbol{x}) \{1 + o(1)\}$$

gives, for every fixed $\boldsymbol{x}$,

$$\begin{aligned}
\bar{\Delta}_{\mathbf{H}} * \bar{L}_{\mathbf{G}}(\boldsymbol{x}) &= |\mathbf{G}|^{-1/2} \int \bar{\Delta}(\boldsymbol{z}) \bar{L}(\mathbf{G}^{-1/2}\boldsymbol{x} - \mathbf{G}^{-1/2}\mathbf{H}^{1/2}\boldsymbol{z}) d\boldsymbol{z} \\
&= \frac{1}{4} m_2(K)^2 |\mathbf{G}|^{-1/2} (\operatorname{vec}^{\top} \mathbf{H})^{\otimes 2} (\mathbf{G}^{-1/2})^{\otimes 4} \mathsf{D}^{\otimes 4} \bar{L}(\mathbf{G}^{-1/2}\boldsymbol{x}) \{1 + o(1)\} \\
&= \frac{1}{4} m_2(K)^2 (\operatorname{vec}^{\top} \mathbf{H})^{\otimes 2} \mathsf{D}^{\otimes 4} \bar{L}_{\mathbf{G}}(\boldsymbol{x}) \{1 + o(1)\}.
\end{aligned}$$

Differentiating the previous expression, we obtain

$$(\mathsf{D}_{\mathbf{H}} \bar{\Delta}_{\mathbf{H}}) * \bar{L}_{\mathbf{G}}(\boldsymbol{x}) = \frac{1}{2} m_2(K)^2 \left(\operatorname{vec}^{\top} \mathbf{H} \otimes \mathbf{I}_{d^2} \right) \mathsf{D}^{\otimes 4} \bar{L}_{\mathbf{G}}(\boldsymbol{x}) \{1 + o(1)\}.$$

Therefore, integrating out, using the change of variables $z = \mathbf{G}^{-1/2}(x - y)$ and a Taylor expansion of f,

$$\begin{aligned}
\Xi_2 &= \iint (\mathsf{D}_{\mathbf{H}}\bar{\Delta}_{\mathbf{H}}) * \bar{L}_{\mathbf{G}}(x-y)(\mathsf{D}_{\mathbf{H}}\bar{\Delta}_{\mathbf{H}}) * \bar{L}_{\mathbf{G}}(x-y)^{\top} f(x) f(y) dx dy \\
&= \frac{1}{4} m_2(K)^4 \big(\mathrm{vec}^{\top}\mathbf{H} \otimes \mathbf{I}_{d^2}\big) \iint \mathsf{D}^{\otimes 4}\bar{L}_{\mathbf{G}}(x-y)\mathsf{D}^{\otimes 4}\bar{L}_{\mathbf{G}}(x-y)^{\top} f(x) f(y) dx dy \\
&\quad \times (\mathrm{vec}\,\mathbf{H} \otimes \mathbf{I}_{d^2})\{1 + o(1)\} \\
&= \frac{1}{4} m_2(K)^4 R(f)\big(\mathrm{vec}^{\top}\mathbf{H} \otimes \mathbf{I}_{d^2}\big)|\mathbf{G}|^{-1/2}(\mathbf{G}^{-1/2})^{\otimes 4}\mathbf{R}(\mathsf{D}^{\otimes 4}\bar{L})(\mathbf{G}^{-1/2})^{\otimes 4}(\mathrm{vec}\,\mathbf{H} \otimes \mathbf{I}_{d^2}) \\
&\quad \times \{1 + o(1)\}
\end{aligned}$$

and similarly

$$\begin{aligned}
\Xi_1 &= \iiint (\mathsf{D}_{\mathbf{H}}\bar{\Delta}_{\mathbf{H}}) * \bar{L}_{\mathbf{G}}(x-y)(\mathsf{D}_{\mathbf{H}}\bar{\Delta}_{\mathbf{H}}) * \bar{L}_{\mathbf{G}}(x-z)^{\top} f(x) f(y) f(z)\, dx dy dz \\
&= \frac{1}{4} m_2(K)^4 \big(\mathrm{vec}^{\top}\mathbf{H} \otimes \mathbf{I}_{d^2}\big) \iiint \mathsf{D}^{\otimes 4}\bar{L}_{\mathbf{G}}(x-y)\mathsf{D}^{\otimes 4}\bar{L}_{\mathbf{G}}(x-z)^{\top} f(x) f(y) f(z)\, dx dy dz \\
&\quad \times (\mathrm{vec}\,\mathbf{H} \otimes \mathbf{I}_{d^2})\{1 + o(1)\} \\
&= \frac{1}{4} m_2(K)^4 \big(\mathrm{vec}^{\top}\mathbf{H} \otimes \mathbf{I}_{d^2}\big) \iiint \bar{L}_{\mathbf{G}}(x-y)\bar{L}_{\mathbf{G}}(x-z) f(x)\mathsf{D}^{\otimes 4} f(y)\mathsf{D}^{\otimes 4} f(z)^{\top}\, dx dy dz \\
&\quad \times (\mathrm{vec}\,\mathbf{H} \otimes \mathbf{I}_{d^2})\{1 + o(1)\} \\
&= \frac{1}{4} m_2(K)^4 \big(\mathrm{vec}^{\top}\mathbf{H} \otimes \mathbf{I}_{d^2}\big) \int \mathsf{D}^{\otimes 4} f(x)\mathsf{D}^{\otimes 4} f(x)^{\top} f(x)\, dx\, (\mathrm{vec}\,\mathbf{H} \otimes \mathbf{I}_{d^2})\{1 + o(1)\}.
\end{aligned}$$

□

Duong & Hazelton (2005b) show that we can express MSE($\mathbf{G}$) in terms of the derivatives of SCV − MISE. We decompose this difference into (SCV − MISE2) + (MISE2 − MISE) and focus on the former, as it dominates the latter.

Proof of Theorem 2. From Duong & Hazelton (2005b), the leading term of the squared bias in MSE($\mathbf{G}$) is $\mathbb{E}\{\mathsf{D}_{\mathbf{H}}(\mathrm{SCV} - \mathrm{MISE2})(\mathbf{H})\}^{\top}\mathbb{E}\{\mathsf{D}_{\mathbf{H}}(\mathrm{SCV} - \mathrm{MISE2})(\mathbf{H})\}$. From Lemma 4.3, the difference $(\mathbb{E}\mathrm{SCV} - \mathrm{MISE2})(\mathbf{H}) = \frac{1}{4} m_2(K)^2 \boldsymbol{\omega}_4^{\top}(\mathrm{vec}\,\mathbf{H})^{\otimes 2}\{1 + o(1)\}$ so

$$\begin{aligned}
&\mathbb{E}\{\mathsf{D}_{\mathbf{H}}(\mathrm{SCV} - \mathrm{MISE2})(\mathbf{H})\}^{\top}\mathbb{E}\{\mathsf{D}_{\mathbf{H}}(\mathrm{SCV} - \mathrm{MISE2})(\mathbf{H})\} \\
&\qquad = \frac{1}{16} m_2(K)^4 \boldsymbol{\omega}_4^{\top}\Gamma_2\big(\mathrm{vec}\,\mathbf{H}\mathrm{vec}^{\top}\mathbf{H} \otimes \mathbf{I}_{d^2}\big)\Gamma_2^{\top}\boldsymbol{\omega}_4\{1 + o(1)\} \\
&\qquad = \frac{1}{4} m_2(K)^4 \boldsymbol{\omega}_4^{\top}\big(\mathrm{vec}\,\mathbf{H}\mathrm{vec}^{\top}\mathbf{H} \otimes \mathbf{I}_{d^2}\big)\boldsymbol{\omega}_4\{1 + o(1)\},
\end{aligned}$$

as it is not hard to check that $\Gamma_2^{\top}\boldsymbol{\omega}_4 = 2\boldsymbol{\omega}_4$. The leading term of the variance in MSE($\mathbf{G}$) is $\mathrm{var}\mathsf{D}_{\mathbf{H}}\mathrm{SCV}(\mathbf{H})$, which was already computed in Lemma 4.3. □

Proof of Theorem 3. Combining theorem 2 from Chacón & Duong (2010) with Theorem 2 in this paper, we find that $\mathrm{MSE}(\hat{\mathbf{H}}_{\mathrm{SCV}}; \mathbf{G}) = O(n^{-4/(d+6)}\mathbf{J}_{d^2})(\mathrm{vec}\,\mathbf{H}_{\mathrm{MISE}})(\mathrm{vec}^\top \mathbf{H}_{\mathrm{MISE}})$. The result follows immediately from the approach in Duong & Hazelton (2005a). □

Proof of Corollary 1. Substituting $\mathbf{G} = g^2\mathbf{I}_d$ into the dominant squared bias term from Theorem 2, we obtain $\mathrm{AMSE}(g) = \frac{1}{4}m_2(K)^4(n^{-2}g^{-2d-8}A_1 + 2n^{-1}g^{-d-2}A_2 + g^4A_3)\{1 + o(1)\}$. Differentiating with respect to g and setting to zero leads to

$$(d+4)A_1n^{-1}g^{-2d-12} + A_2n^{-1}g^{-d-6} - 2A_3 = 0,$$

which is a quadratic in $n^{-1}g^{-d-6}$, whose solution is readily found. The discriminant of this quadratic is $(d+2)^2A_2^2 + 8(d+4)A_1A_3 > 0$, as A_1, $A_3 > 0$, so the solutions are real-valued. □

To show that the expression from Corollary 1 is essentially the same as the optimal pilot selector from Duong & Hazelton (2005b), we rely on the next lemma.

Lemma 3. *Let $\mathbf{A} \in \mathcal{M}_{d\times d}$ and $\boldsymbol{\Theta}_6 = \int (\mathsf{D}^2)^3 f(\boldsymbol{x}) f(\boldsymbol{x})\, d\boldsymbol{x}$. Then*

$$\begin{aligned}\mathrm{tr}(\mathbf{A}\boldsymbol{\Theta}_6) &= \mathrm{vec}^\top(\mathbf{A}\otimes\mathbf{I}_{d^2})\psi_6\\ \mathrm{tr}(\mathbf{A}\boldsymbol{\Theta}_6^2) &= \psi_6^\top\big(\mathbf{I}_d\otimes\mathbf{A}\otimes\mathrm{vec}\,\mathbf{I}_{d^2}\mathrm{vec}^\top\mathbf{I}_{d^2}\big)\psi_6.\end{aligned}$$

Proof. Let us denote by $\mathbf{K}_{m,n}$ the commutation matrix of order $mn\times mn$, whose properties we use next (see Magnus & Neudecker 1979). For a vector $\mathbf{a}\in\mathbb{R}^{d^2}$, we have that

$$\begin{aligned}\mathbf{a}^\top\mathrm{vec}\,(\mathsf{D}^2)^3 &= \mathbf{a}^\top(\mathsf{D}^2\otimes\mathsf{D}^2)\mathsf{D}^{\otimes 2} = \mathrm{vec}^\top(\mathsf{D}^2\otimes\mathsf{D}^2)\mathrm{vec}\,(\mathsf{D}^{\otimes 2}\mathbf{a}^\top)\\ &= (\mathsf{D}^{\otimes 4})^\top(\mathbf{I}_d\otimes\mathbf{K}_{dd}\otimes\mathbf{I}_d)(\mathbf{a}\otimes\mathbf{I}_{d^2})\mathsf{D}^{\otimes 2} = (\mathsf{D}^{\otimes 4})^\top(\mathbf{a}\otimes\mathbf{I}_{d^2})\mathsf{D}^{\otimes 2} \qquad (3)\\ &= \mathrm{vec}^\top\big(\mathbf{a}^\top\otimes\mathbf{I}_{d^2}\big)\mathsf{D}^{\otimes 6} = \big(\mathbf{a}^\top\otimes\mathrm{vec}^\top\mathbf{I}_{d^2}\big)\mathsf{D}^{\otimes 6}.\end{aligned}$$

We apply (3) with $\mathbf{a} = \mathrm{vec}\,\mathbf{A}$ to the trace of the product of a matrix $\mathbf{A}$ with the cube of the Hessian operator to obtain the first stated equality

$$\mathrm{tr}\{\mathbf{A}(\mathsf{D}^2)^3\} = (\mathrm{vec}^\top\mathbf{A})\mathrm{vec}\,(\mathsf{D}^2)^3 = \big(\mathrm{vec}^\top\mathbf{A}\otimes\mathrm{vec}^\top\mathbf{I}_{d^2}\big)\mathsf{D}^{\otimes 6} = \mathrm{vec}^\top(\mathbf{A}\otimes\mathbf{I}_{d^2})\mathsf{D}^{\otimes 6}.$$

To prove the second equality we use the first one, some more properties of the Kronecker product (Magnus & Neudecker 1999) and (3) with $\mathbf{a} = (\mathbf{I}_d\otimes\mathbf{A}^\top\otimes\mathrm{vec}^\top\mathbf{I}_{d^2})\psi_6$, to obtain

$$\begin{aligned}\mathrm{tr}(\mathbf{A}\boldsymbol{\Theta}_6^2) &= \psi_6^\top\mathrm{vec}\,\{(\mathbf{A}\boldsymbol{\Theta}_6)\otimes\mathbf{I}_{d^2}\}\\ &= \psi_6^\top(\mathbf{I}_d\otimes\mathbf{A}\otimes\mathrm{vec}\,\mathbf{I}_{d^2})\mathrm{vec}\,\Theta_6\\ &= \big[\big\{\psi_6^\top(\mathbf{I}_d\otimes\mathbf{A}\otimes\mathrm{vec}\,\mathbf{I}_{d^2})\big\}\otimes\mathrm{vec}^\top\mathbf{I}_{d^2}\big]\psi_6\\ &= \psi_6^\top(\mathbf{I}_d\otimes\mathbf{A}\otimes\mathrm{vec}\,\mathbf{I}_{d^2}\otimes\mathrm{vec}^\top\mathbf{I}_{d^2})\psi_6,\end{aligned}$$

as desired. □

Proof of Corollary 2. If L is a normal kernel then, using properties from section 3.3 in Chacón & Duong (2010),

$$\mathsf{D}^{\otimes 4}\bar{\phi}(0) = \mathsf{D}^{\otimes 4}\phi_{2\mathbf{I}_d}(0) = 2^{-(d+4)/2}\mathsf{D}^{\otimes 4}\phi(0) = 3 \times 2^{-d-2}\pi^{-d/2}\boldsymbol{\mathcal{S}}_{d,4}(\operatorname{vec}\mathbf{I}_d)^{\otimes 2}.$$

Substituting this into Corollary 1, the coefficients A_1 and A_2 can be rewritten. For A_1,

$$\begin{aligned}
\frac{16}{9}(4\pi)^d n^{4/(d+4)} A_1 &= \operatorname{tr}\big\{\big(\operatorname{vec}\mathbf{C}\operatorname{vec}^\top\mathbf{C}\otimes\mathbf{I}_{d^2}\big)\boldsymbol{\mathcal{S}}_{d,4}\big(\operatorname{vec}\mathbf{I}_d\operatorname{vec}^\top\mathbf{I}_d\big)^{\otimes 2}\big\}\\
&= \operatorname{tr}\big\{\big(\operatorname{vec}\mathbf{C}\operatorname{vec}^\top\mathbf{C}\operatorname{vec}\mathbf{I}_d\operatorname{vec}^\top\mathbf{I}_d\otimes\operatorname{vec}\mathbf{I}_d\operatorname{vec}^\top\mathbf{I}_d\big)\boldsymbol{\mathcal{S}}_{d,4}\big\}\\
&= (\operatorname{tr}\mathbf{C})\operatorname{tr}\big\{\big(\operatorname{vec}\mathbf{C}\operatorname{vec}^\top\mathbf{I}_d\otimes\operatorname{vec}\mathbf{I}_d\operatorname{vec}^\top\mathbf{I}_d\big)\boldsymbol{\mathcal{S}}_{d,4}\big\}\\
&= (\operatorname{tr}\mathbf{C})\operatorname{tr}\big\{(\operatorname{vec}\mathbf{C}\otimes\operatorname{vec}\mathbf{I}_d)\big(\operatorname{vec}^\top\mathbf{I}_d\otimes\operatorname{vec}^\top\mathbf{I}_d\big)\boldsymbol{\mathcal{S}}_{d,4}\big\}\\
&= (\operatorname{tr}\mathbf{C})\operatorname{tr}\big\{\operatorname{vec}(\mathbf{C}\otimes\mathbf{I}_d)(\mathbf{I}_d\otimes\mathbf{K}_{dd}\otimes\mathbf{I}_d)(\mathbf{I}_d\otimes\mathbf{K}_{dd}\otimes\mathbf{I}_d)\big(\operatorname{vec}^\top\mathbf{I}_{d^2}\big)\boldsymbol{\mathcal{S}}_{d,4}\big\}\\
&= (\operatorname{tr}\mathbf{C})\operatorname{tr}\big\{\operatorname{vec}(\mathbf{C}\otimes\mathbf{I}_d)\big(\operatorname{vec}^\top\mathbf{I}_{d^2}\big)\boldsymbol{\mathcal{S}}_{d,4}\big\},
\end{aligned}$$

and for A_2,

$$\begin{aligned}
\frac{4}{3}(4\pi)^{d/2} n^{4/(d+4)} A_2 &= \operatorname{tr}\big\{\big(\operatorname{vec}\mathbf{C}\operatorname{vec}^\top\mathbf{C}\otimes\mathbf{I}_{d^2}\big)\boldsymbol{\mathcal{S}}_{d,4}(\operatorname{vec}\mathbf{I}_d)^{\otimes 2}\psi_6^\top(\operatorname{vec}\mathbf{I}_d\otimes\mathbf{I}_{d^4})\big\}\\
&= \operatorname{tr}\big\{\boldsymbol{\mathcal{S}}_{d,4}\big(\operatorname{vec}\mathbf{C}\operatorname{vec}^\top\mathbf{C}\operatorname{vec}\mathbf{I}_d\otimes\operatorname{vec}\mathbf{I}_d\big)\psi_6^\top(\operatorname{vec}\mathbf{I}_d\otimes\mathbf{I}_{d^4})\big\}\\
&= (\operatorname{tr}\mathbf{C})\operatorname{tr}\big\{\boldsymbol{\mathcal{S}}_{d,4}(\operatorname{vec}\mathbf{C}\otimes\operatorname{vec}\mathbf{I}_d)\psi_6^\top(\operatorname{vec}\mathbf{I}_d\otimes\mathbf{I}_{d^4})\big\}.
\end{aligned}$$

Rewriting these expressions for A_1 and A_2, and that for A_3, as scalar products of vectorized matrices gives the result.

The equivalent coefficients from Duong & Hazelton (2005b) are

$$\begin{aligned}
A_1' &= \frac{1}{16}(4\pi)^{-d} n^{-4/(d+4)}(\operatorname{tr}\mathbf{C})\{4+(d+4)\operatorname{tr}\mathbf{C}\},\\
A_2' &= \frac{1}{16}(4\pi)^{-d/2} n^{-4/(d+4)}\operatorname{tr}\big[\{2\mathbf{C}^2+(\operatorname{tr}\mathbf{C})\mathbf{C}\}\boldsymbol{\Theta}_6\big],\\
A_3' &= \frac{1}{4} n^{-4/(d+4)}\operatorname{tr}\big(\mathbf{C}^2\boldsymbol{\Theta}_6^2\big).
\end{aligned}$$

Applying the identities from the previous lemma to convert $\operatorname{tr}\big\{[2\mathbf{C}^2+(\operatorname{tr}\mathbf{C})\mathbf{C}]\boldsymbol{\Theta}_6\big\}$ and $\operatorname{tr}(\mathbf{C}^2\boldsymbol{\Theta}_6^2)$ to expressions involving ψ_6 establishes the result. □

References

ABDOUS, B. (1999). *L2* version of the double kernel method. *Statistics* **32**, 249–266.

BERLINET, A. & DEVROYE, L. (1994). A comparison of kernel density estimates. *Publ. l'Inst. Statist. l'Univ. Paris* **38**, 3–59.

CAO, R. (1993). Bootstrapping the mean integrated squared error. *J. Multivariate Anal.* **45**, 137–160.

CAO, R., CUEVAS, A. & GONZÁLEZ-MANTEIGA, W. (1994). A comparative study of several smoothing methods in density estimation. *Comput. Statist. Data Anal.* **17**, 153–176.

CHACÓN, J.E. (2009). Data-driven choice of the smoothing parametrization for kernel density estimators. *Canad. J. Statist.* **37**, 249–265.

CHACÓN, J.E. & DUONG, T. (2010). Multivariate plug-in bandwidth selection with unconstrained pilot bandwidth matrices. *Test* **19**, 375–398.

CHACÓN, J.E., DUONG, T. & WAND, M.P. (2011). Asymptotics for general multivariate kernel density derivatives. *Statistica Sinica*. **21,** 807–840..

DEVROYE, L. (1996). Random variate generation in one line of code. In *1996 Winter Simulation Conference Proceedings*, eds. J.M. Charnes, D.J. Morrice, D.T. Brunner and J.J. Swain, pp. 265–272. San Diego, CA: ACM.

DUONG, T. (2007). ks: Kernel density estimation and kernel discriminant analysis for multivariate data in R. *J. Statist. Softw.* **21(7)**, 1–16.

DUONG, T. & HAZELTON, M.L. (2003). Plug-in bandwidth matrices for bivariate kernel density estimation. *J. Nonparametr. Stat.* **15**, 17–30.

DUONG, T. & HAZELTON, M.L. (2005a). Convergence rates for unconstrained bandwidth matrix selectors in multivariate kernel density estimation. *J. Multivariate Anal.* **93**, 417–433.

DUONG, T. & HAZELTON, M.L. (2005b). Cross-validation bandwidth matrices for multivariate kernel density estimation. *Scand. J. Statist.* **32**, 485–506.

HALL, P. & MARRON, J.S. (1987). Estimation of integrated squared density derivatives. *Statist. Probab. Lett.* **6**, 109–115.

HALL, P. & MARRON, J.S. (1991). Lower bounds for bandwidth selection in density estimation. *Probab. Theory Related Fields* **90**, 149–163.

HALL, P., MARRON, J.S. & PARK, B.U. (1992). Smoothed cross validation. *Probab. Theory Rel. Fields* **92**, 1–20.

HOLMQUIST, B. (1985). The direct product permuting matrices. *Linear Multilinear Algebra* **17**, 117–141.

HOLMQUIST, B. (1988). Moments and cumulants of the multivariate normal distribution. *Stochastic Anal.* **6**, 273–278.

HOLMQUIST, B. (1996). The d-variate vector Hermite polynomial of order k. *Linear Algebra Appl.* **237/238**, 155–190.

JAMMALAMADAKA, S.R., RAO, T.S. & TERDIK, G. (2006). Higher order cumulants of random vectors and applications to statistical inference and time series. *Sankhyā* **68**, 326–356.

JONES, M.C. (1998). On some kernel density estimation bandwidth selectors related to the double kernel method. *Sankhyā* **60**, 249–264.

JONES, M.C. & SHEATHER, S.J. (1991). Using non-stochastic terms to advantage in kernel-based estimation of integrated squared density derivatives. *Statist. Probab. Lett.* **11**, 511–514.

JONES, M.C., MARRON, J.S. & PARK, B.U. (1991). A simple root n bandwidth selector. *Annals Statist.* **19**, 1919–1932.

MAGNUS, J.R. & NEUDECKER, H. (1979). The commutation matrix: some properties and applications. *Annals Statist.* **7**, 381–394.

MAGNUS, J.R. & NEUDECKER, H. (1999). *Matrix Differential Calculus with Applications in Statistics and Econometrics*, rev. edn. Chichester: John Wiley & Sons.

SAIN, S.R., BAGGERLY, K.A. & SCOTT, D.W. (1994). Cross-validation of multivariate densities. *J. Amer. Statist. Assoc.* **89**, 807–817.

SCHAUER, K., DUONG, T., BLEAKLEY, K., BARDIN, S., BORNENS, M. & GOUD, B. (2010). Probabilistic density maps to study global endomembrane organization. *Nat. Methods* **7**, 560–568.

SCHOTT, J.R. (2003). Kronecker product permutation matrices and their application to moment matrices of the normal distribution. *J. Multivariate Anal.* **87**, 177–190.

SCOTT, D.W. (1992). *Multivariate Density Estimation: Theory, Practice, and Visualization*. New York: John Wiley & Sons.

SIMONOFF, J.S. (1996). *Smoothing Methods in Statistics*. Berlin: Springer-Verlag.

WAND, M.P. (1992). Error analysis for general multivariate kernel estimators. *J. Nonparametr. Stat.* **2**, 2–15.

WAND, M.P. & JONES, M.C. (1993). Comparison of smoothing parametrizations in bivariate kernel density estimation. *J. Amer. Statist. Assoc.* **88**, 520–528.

WAND, M.P. & JONES, M.C. (1994). Multivariate plug-in bandwidth selection. *Comput. Statist.* **9**, 97–117.