Data-driven density derivative estimation, with applications to nonparametric clustering and bump hunting

José E. Chacón

Departamento de Matemáticas, Universidad de Extremadura, E-06006 Badajoz, Spain
e-mail: jechacon@unex.es

and

Tarn Duong

Theoretical and Applied Statistics Laboratory (LSTA), University Pierre and Marie Curie –
Paris 6, F-75005, Paris, France and Institute of Translational Neurosciences
(IHU-A-ICM), Pitié-Salpêtrière Hospital, F-75005, Paris, France
e-mail: tarn.duong@upmc.fr

Abstract: Important information concerning a multivariate data set, such as clusters and modal regions, is contained in the derivatives of the probability density function. Despite this importance, nonparametric estimation of higher order derivatives of the density functions have received only relatively scant attention. Kernel estimators of density functions are widely used as they exhibit excellent theoretical and practical properties, though their generalization to density derivatives has progressed more slowly due to the mathematical intractabilities encountered in the crucial problem of bandwidth (or smoothing parameter) selection. This paper presents the first fully automatic, data-based bandwidth selectors for multivariate kernel density derivative estimators. This is achieved by synthesizing recent advances in matrix analytic theory which allow mathematically and computationally tractable representations of higher order derivatives of multivariate vector valued functions. The theoretical asymptotic properties as well as the finite sample behaviour of the proposed selectors are studied. In addition, we explore in detail the applications of the new data-driven methods for two other statistical problems: clustering and bump hunting. The introduced techniques are combined with the mean shift algorithm to develop novel automatic, nonparametric clustering procedures which are shown to outperform mixture-model cluster analysis and other recent nonparametric approaches in practice. Furthermore, the advantage of the use of smoothing parameters designed for density derivative estimation for feature significance analysis for bump hunting is illustrated with a real data example.

1. Introduction

The estimation of density derivatives has full potential for applications. This has been noted even in the first seminal papers on density estimation, as Parzen (1962), which was also concerned with the estimation of the mode of a unimodal distribution, the value that makes zero the first density derivative. In the multivariate case, the pioneering work of Fukunaga and Hostetler (1975) showed how the estimation of the gradient vector can also be used for clustering and data filtering, leading to a substantial amount of literature on the subject, under the name of the *mean shift algorithm*. Looking further afield, Cheng (1995) made use of the mean shift idea for image analysis, and the highly-cited paper by Comaniciu and Meer (2002) showed how these techniques can be useful for low-level vision problems, discontinuity preserving smoothing and image segmentation. A further very popular use of the mean shift algorithm is for real-time object tracking, as described in Comaniciu, Ramesh and Meer (2003).

From the perspective of statistical data analysis, in the multidimensional context the estimation of the first and second derivatives of the density is crucial to identify significant features of the distribution, such as local extrema, valleys, ridges or saddle points. In this sense, Godtliebsen, Marron and Chaudhuri (2002) developed methods for determining and visualizing such features in dimension two, extending previous work on scale space ideas introduced in Chaudhuri and Marron (1999) for the univariate case (the SiZer approach), and the same authors also explored the application of this methodology to digital image analysis in Godtliebsen, Marron and Chaudhuri (2004). Duong *et al.* (2008) generalized these results for multivariate data in arbitrary dimensions and provided a novel visualization for three-dimensional data. These techniques have been widely applied recently in the field of flow cytometry; see Zeng *et al.* (2007); Naumann and Wand (2009) or Pratt *et al.* (2009).

Another relatively new problem that is closely related to gradient estimation is that of finding filaments in point clouds, which has applications in medical imaging, remote sensing, seismology and cosmology. This problem is rigorously stated and analyzed in Genovese *et al.* (2009). Filaments are one-dimensional curves embedded in a point process, and it can be shown that steepest ascent paths (i.e., the paths from each point following the gradient direction) concentrate around them, so gradient estimation appears as a useful tool for filament detection.

In this paper we focus on kernel estimators of multivariate density derivatives of arbitrary order, formally defined in Section 2 below. As for any kernel estimator, it is well known that the crucial factor that determines the performance of the estimator in practice is the choice of the bandwidth matrix. In the multivariate setting there are several levels of sophistication at the time of specifying the bandwidth matrix to be used in the kernel estimator (see Wand and Jones, 1995, Chapter 4). The most general bandwidth type consists of a symmetric positive definite matrix; it allows the kernel estimator to smooth in any direction whether coordinate or not. This general class of bandwidth matrices can be constrained to consider positive definite diagonal matrices, allowing for

different degrees of smoothing along each of the coordinate axis, or even further to consider a bandwidth matrix involving only a positive scalar multiple of the identity matrix, meaning that the same smoothing is applied to every coordinate direction. As noted in Wand and Jones (1993) in the density estimation context, the single-parameter class should not be used for unscaled data or, as stated by Comaniciu and Meer (2002) in terms of feature space analysis, to use this bandwidth class at least the validity of an Euclidean metric for the feature space should be previously checked.

The simpler parameterizations are in general more widely used than the unconstrained counterpart for two reasons: first, in practice they need less smoothing parameters to be tuned, and second, due to the difficulties encountered in the mathematical analysis of estimators with unconstrained bandwidths. However, Chacón, Duong and Wand (2011) provided a detailed error analysis of kernel density derivative estimators with unconstrained bandwidths and showed that the use of the simpler parameterizations can lead to a substantial loss in terms of efficiency, and that this problem becomes more and more important as the order of the derivative to be estimated increases.

Chacón, Duong and Wand (2011) also proposed an optimal bandwidth selector for the normal case, but they did not develop more sophisticated data-driven choices of the bandwidth matrix with applicability to more general densities, which is crucial to make density derivative estimation useful in practice. Along the same lines, Comaniciu and Meer (2002) argue that most existing bandwidth selection methods for the mean shift algorithm, all of them for the single-parameter class of bandwidths, are based on empirical arguments.

In the univariate case there exist some approaches to bandwidth selection for density derivative estimation: Härdle, Marron and Wand (1990) introduced a cross validation method and showed its optimality; Jones (1992) derived the relative rate of convergence of this method and also for a plug-in proposal; Wu (1997) studied two root n selectors in the Fourier domain, and more recently Dobrovidov and Rud'ko (2010) focused on the smoothed cross validation bandwidth selector for the density derivative. In the multivariate case, however, the issue of automatic bandwidth selection for density derivative estimation has remained largely unexplored. Given the smaller body of multivariate density estimation research as compared to their univariate cousins, it is not surprising that multivariate density derivative estimation suffers equally (if not more so) from a lack of solid results. To the best of our knowledge, in the literature the only published approaches to bandwidth selection for multivariate kernel estimation of density derivatives are the recent papers Horová and Vopatová (2011) and Horová, Koláček and Vopatová (2013), but both focus exclusively on the first derivative.

This paper proposes three new methods for unconstrained bandwidth matrix selection for the multivariate kernel density derivative estimator, and explores their applicability to other related statistical problems. In Section 2, we introduce the mathematical framework for the analysis of multivariate derivatives. In Section 3 we show that the relative rate of convergence of these unconstrained selectors is the same as for the classes of simpler bandwidth matrices, so that

from an asymptotic point of view our methods can be as successful as (and more flexible than) those needing less smoothing parameters. The finite-sample behaviour of the new bandwidths is investigated in Section 4, and their application to develop new data-driven nonparametric clustering methods via the mean shift algorithm is explored in Section 5, and for feature significance in Section 6. Finally, the proofs of the results are given in an appendix.

2. Kernel density derivative estimation

The problem of estimating the r-th derivative of a multivariate density is considered in this section. From a multivariate point of view, the r-th derivative of a function is understood as the set of all its partial derivatives of order r, rather than just one of them. Notice that, for instance, in a multivariate Taylor expansion of order r all of the partial derivatives of order r are needed to compute the r-th order term. Or, in another related example, all the second order partial derivatives are involved in the computation of the Hessian matrix.

All the r-th partial derivatives can be neatly organized into a single vector as follows: if f is a real d-variate density function and $\boldsymbol{x} = (x_1, \ldots, x_d)$, denote by $\mathsf{D} = \partial/\partial\boldsymbol{x} = (\partial/\partial x_1, \ldots, \partial/\partial x_d)$ the first derivative (gradient) operator. All the second order partial derivatives can be organized into the Hessian matrix $\mathsf{H}f = (\partial^2 f/\partial x_i \partial x_j)_{i,j=1}^d$, and the Hessian operator can be formally written as $\mathsf{H} = \mathsf{D}\mathsf{D}^\top$ if the usual convention $(\partial/\partial x_i)(\partial/\partial x_j) = \partial^2/(\partial x_i \partial x_j)$ is taken into account. For $r \geq 3$, however, it is not that clear how to organize the set containing all the d^r partial derivatives of order r. Here we adopt the unified approach used in Holmquist (1996a) or Kollo and von Rosen (2005, Section 1.4), where the r-th derivative of f is defined to be the vector $\mathsf{D}^{\otimes r}f = (\mathsf{D}f)^{\otimes r} = \partial^r f/\partial\boldsymbol{x}^{\otimes r} \in \mathbb{R}^{d^r}$. In the previous equation $\mathsf{D}^{\otimes r}$ denotes the r-th Kronecker power of the operator D; see, e.g., Magnus and Neudecker (1999) for the definition of the Kronecker product. Naturally, $\mathsf{D}^{\otimes 0}f = f$, $\mathsf{D}^{\otimes 1}f = \mathsf{D}f$ and, for example, $\mathsf{D}^{\otimes 2} = \text{vec}\,\mathsf{H}$, where vec denotes the operator which concatenates the columns of a matrix into a single vector.

Here we study the problem of estimating the r-th derivative $\mathsf{D}^{\otimes r}f$ from a sample $\mathbf{X}_1, \ldots, \mathbf{X}_n$ of independent and identically distributed random variables with common density f. The usual kernel estimator of f is defined as $\hat{f}_\mathbf{H}(\boldsymbol{x}) = n^{-1}\sum_{i=1}^n K_\mathbf{H}(\boldsymbol{x} - \mathbf{X}_i)$, where the kernel K is a spherically symmetric density function, the bandwidth $\mathbf{H}$ is a symmetric positive definite matrix and $K_\mathbf{H}(\boldsymbol{x}) = |\mathbf{H}|^{-1/2}K(\mathbf{H}^{-1/2}\boldsymbol{x})$. Thus, the most straightforward estimator of $\mathsf{D}^{\otimes r}f$ is just the r-th derivative of $\hat{f}_\mathbf{H}$, given by $\mathsf{D}^{\otimes r}\hat{f}_\mathbf{H}(\boldsymbol{x}) = n^{-1}\sum_{i=1}^n \mathsf{D}^{\otimes r}K_\mathbf{H}(\boldsymbol{x} - \mathbf{X}_i)$, where the roles of K and $\mathbf{H}$ can be separated for implementation purposes by noting that $\mathsf{D}^{\otimes r}K_\mathbf{H}(\boldsymbol{x}) = |\mathbf{H}|^{-1/2}(\mathbf{H}^{-1/2})^{\otimes r}\mathsf{D}^{\otimes r}K(\mathbf{H}^{-1/2}\boldsymbol{x})$, as shown in Chacón, Duong and Wand (2011), where for any matrix $\mathbf{A}$ it is understood that $\mathbf{A}^{\otimes 0} = 1 \in \mathbb{R}$ and $\mathbf{A}^{\otimes 1} = \mathbf{A}$. See, however, Jones (1994) for other possible estimators in the univariate context.

For the density estimation case ($r = 0$), Wand and Jones (1993) showed that the use of bandwidths belonging to the class $\mathcal{I} = \{h^2\mathbf{I}_d : h \in \mathbb{R}\}$, with $\mathbf{I}_d$ the

$d \times d$ identity matrix, or the class $\mathcal{D} = \{\mathrm{diag}(h_1^2, h_2^2, \ldots, h_d^2) : h_1, h_2, \ldots, h_d \in \mathbb{R}\}$, may lead to dramatically less efficient estimators than those based on bandwidth matrices drawn from $\mathcal{F}$, the space of all positive definite symmetric matrices. Moreover Chacón, Duong and Wand (2011) showed that the issue of efficiency loss is even more severe for $r \geq 1$. So the development of unconstrained bandwidth selectors for density derivative estimation, which is achieved in this paper, may also represent an important improvement in practice.

To measure the error committed by the kernel estimator for the sample at hand it is natural to consider the integrated squared error (ISE), defined as

$$\mathrm{ISE}_r(\mathbf{H}) = \int_{\mathbb{R}^d} \|\mathsf{D}^{\otimes r} \hat{f}_{\mathbf{H}}(\boldsymbol{x}) - \mathsf{D}^{\otimes r} f(\boldsymbol{x})\|^2 d\boldsymbol{x},$$

where $\|\cdot\|$ denotes the usual Euclidean norm. This quantity depends on the data, so it is common to consider the mean integrated squared error $\mathrm{MISE}_r(\mathbf{H}) = \mathbb{E}[\mathrm{ISE}_r(\mathbf{H})]$ as a non-stochastic measure of error, and its minimizer $\mathbf{H}_{\mathrm{MISE},r} = \mathrm{argmin}_{\mathbf{H} \in \mathcal{F}} \mathrm{MISE}_r(\mathbf{H})$ as the non-stochastic optimal bandwidth choice. A more detailed discussion of the advantages and disadvantages of the ISE and MISE viewpoints can be found in Jones (1991).

Standard calculations lead to the integrated variance plus integrated squared bias decomposition $\mathrm{MISE}_r(\mathbf{H}) = \mathrm{IV}_r(\mathbf{H}) + \mathrm{ISB}_r(\mathbf{H})$, where $\mathrm{IV}_r(\mathbf{H}) = \int_{\mathbb{R}^d} \mathrm{tr}\, \mathrm{Var}[\mathsf{D}^{\otimes r} \hat{f}_{\mathbf{H}}(\boldsymbol{x})] d\boldsymbol{x}$ and $\mathrm{ISB}_r(\mathbf{H}) = \int_{\mathbb{R}^d} \|\mathbb{E}[\mathsf{D}^{\otimes r} \hat{f}_{\mathbf{H}}(\boldsymbol{x})] - \mathsf{D}^{\otimes r} f(\boldsymbol{x})\|^2 d\boldsymbol{x}$. By expanding each of these two terms, Chacón, Duong and Wand (2011) showed that a more explicit form of the MISE is given by

$\mathrm{MISE}_r(\mathbf{H})$

$$\begin{aligned} &= \{n^{-1}|\mathbf{H}|^{-1/2}\,\mathrm{tr}\left((\mathbf{H}^{-1})^{\otimes r}\mathbf{R}(\mathsf{D}^{\otimes r}K)\right) - n^{-1}\,\mathrm{tr}\,\mathbf{R}^*(K_{\mathbf{H}} * K_{\mathbf{H}}, \mathsf{D}^{\otimes r}f)\} \\ &\quad + \{\mathrm{tr}\,\mathbf{R}^*(K_{\mathbf{H}} * K_{\mathbf{H}}, \mathsf{D}^{\otimes r}f) - 2\,\mathrm{tr}\,\mathbf{R}^*(K_{\mathbf{H}}, \mathsf{D}^{\otimes r}f) + \mathrm{tr}\,\mathbf{R}(\mathsf{D}^{\otimes r}f)\} \end{aligned}$$

$$(1)$$

where $\mathbf{R}(\mathbf{g}) = \int_{\mathbb{R}^d} \mathbf{g}(\boldsymbol{x})\mathbf{g}(\boldsymbol{x})^\top d\boldsymbol{x}$, and $\mathbf{R}^*(a, \mathbf{g}) = \int_{\mathbb{R}^d} (a * \mathbf{g})(\boldsymbol{x})\mathbf{g}(\boldsymbol{x})^\top d\boldsymbol{x}$ for a vector valued function $\mathbf{g}$ and a real valued function a, with $a * \mathbf{g}$ standing for a component-wise application of the convolution operator.

Moreover, writing $m_2(K)\mathbf{I}_d = \int_{\mathbb{R}^d} \boldsymbol{x}\boldsymbol{x}^\top K(\boldsymbol{x})d\boldsymbol{x}$, under some smoothness assumptions Chacón, Duong and Wand (2011) also showed that an asymptotic approximation of the MISE is given by

$$\begin{aligned} \mathrm{AMISE}_r(\mathbf{H}) &= n^{-1}|\mathbf{H}|^{-1/2}\,\mathrm{tr}\left((\mathbf{H}^{-1})^{\otimes r}\mathbf{R}(\mathsf{D}^{\otimes r}K)\right) \\ &\quad + \tfrac{m_2(K)^2}{4}\,\mathrm{tr}\left((\mathbf{I}_{d^r} \otimes \mathrm{vec}^\top \mathbf{H})\mathbf{R}(\mathsf{D}^{\otimes(r+2)}f)(\mathbf{I}_{d^r} \otimes \mathrm{vec}\,\mathbf{H})\right) \end{aligned}$$

$$(2)$$

and that the minimizer of this AMISE function, $\mathbf{H}_{\mathrm{AMISE},r} = \mathrm{argmin}_{\mathbf{H} \in \mathcal{F}} \mathrm{AMISE}_r(\mathbf{H})$, has entries of order $O(n^{-2/(d+2r+4)})$, leading to a minimum achievable AMISE of order $O(n^{-4/(d+2r+4)})$.

Although these expressions provide an insightful error analysis of multivariate kernel density derivative estimators, they are not directly implementable as software since they all involve the unknown density f. In the next section, we examine strategies to estimate these unknown quantities which lead to optimal data-based selectors for density derivative estimation.

3. Bandwidth selection methods

In this section we propose three new methods to select the bandwidth matrix for kernel density derivative estimation from the data and study their asymptotic properties. These methods are inspired by the cross validation, plug-in and smoothed cross validation methodologies for the estimation of the density in the univariate case, hence their names.

3.1. Cross validation method

Cross validation (CV) techniques for bandwidth selection for univariate density estimation were introduced in Rudemo (1982) and Bowman (1984), and studied in detail in seminal papers like Hall (1983); Stone (1984) and Hall and Marron (1987). They can be motivated in terms of either ISE-oriented or MISE-oriented considerations.

In the case of multivariate density derivative estimation notice that, for a random variable $\mathbf{X}$ having density f and independent of $\mathbf{X}_1, \ldots, \mathbf{X}_n$, using integration by parts it is possible to write

$\mathrm{ISE}_r(\mathbf{H})$

$$= (-1)^r \operatorname{vec}^\top \mathbf{I}_{d^r} \left\{ n^{-2} \sum_{i,j=1}^n \mathsf{D}^{\otimes 2r} K_\mathbf{H} * K_\mathbf{H}(\mathbf{X}_i - \mathbf{X}_j) - 2\mathbb{E}\big[\mathsf{D}^{\otimes 2r} \hat{f}_\mathbf{H}(\mathbf{X})\big] \right\}$$

$$+ \operatorname{tr} \mathbf{R}(\mathsf{D}^{\otimes r} f)$$

provided that the kernel K is sufficiently smooth. The last term is irrelevant for minimizing concerns, and the two first terms can be unbiasedly estimated by

$\mathrm{CV}_r(\mathbf{H})$

$$= (-1)^r \operatorname{vec}^\top \mathbf{I}_{d^r} \left\{ n^{-2} \sum_{i,j=1}^n \mathsf{D}^{\otimes 2r} K_\mathbf{H} * K_\mathbf{H}(\mathbf{X}_i - \mathbf{X}_j) - 2n^{-1} \sum_{i=1}^n \mathsf{D}^{\otimes 2r} \hat{f}_{\mathbf{H},i}(\mathbf{X}_i) \right\}$$

$$= (-1)^r \operatorname{vec}^\top \mathbf{I}_{d^r} \left\{ n^{-2} \sum_{i,j=1}^n \mathsf{D}^{\otimes 2r} K_\mathbf{H} * K_\mathbf{H}(\mathbf{X}_i - \mathbf{X}_j) \right.$$

$$\left. - 2[n(n-1)]^{-1} \sum_{i \neq j} \mathsf{D}^{\otimes 2r} K_\mathbf{H}(\mathbf{X}_i - \mathbf{X}_j) \right\}$$

where $\mathsf{D}^{\otimes 2r} \hat{f}_{\mathbf{H},i}$ denotes the kernel estimator based on the sample with the i-th observation deleted.

From the MISE point of view notice that, for any smooth enough function L,

$$\operatorname{tr} \mathbf{R}^*(L_\mathbf{H}, \mathsf{D}^{\otimes r} f) = \int_{\mathbb{R}^d} (L_\mathbf{H} * \mathsf{D}^{\otimes r} f)(\boldsymbol{x})^\top \mathsf{D}^{\otimes r} f(\boldsymbol{x}) d\boldsymbol{x}$$

$$= (-1)^r \operatorname{vec}^\top \mathbf{I}_{d^r} \int_{\mathbb{R}^d} \mathsf{D}^{\otimes 2r} L_\mathbf{H} * f(\boldsymbol{x}) f(\boldsymbol{x}) d\boldsymbol{x}$$

$$= (-1)^r \operatorname{vec}^\top \mathbf{I}_{d^r} \mathbb{E}\big[\mathsf{D}^{\otimes 2r} L_\mathbf{H}(\mathbf{X}_1 - \mathbf{X}_2)\big],$$

so that $\operatorname{tr}\mathbf{R}^*(L_{\mathbf{H}}, \mathsf{D}^{\otimes r}f)$ can be unbiasedly estimated by

$$(-1)^r[n(n-1)]^{-1}\operatorname{vec}^\top \mathbf{I}_{d^r}\sum_{i\neq j}\mathsf{D}^{\otimes 2r}L_{\mathbf{H}}(\mathbf{X}_i - \mathbf{X}_j).$$

Therefore, in view of (1), $\operatorname{MISE}_r(\mathbf{H}) - \operatorname{tr}\mathbf{R}(\mathsf{D}^{\otimes r}f)$ can be unbiasedly estimated by

$$\begin{aligned}
&\operatorname{CV}_r(\mathbf{H})\\
&= n^{-1}|\mathbf{H}|^{-1/2}\operatorname{tr}\left((\mathbf{H}^{-1})^{\otimes r}\mathbf{R}(\mathsf{D}^{\otimes r}K)\right)\\
&\quad + (-1)^r[n(n-1)]^{-1}\operatorname{vec}^\top \mathbf{I}_{d^r}\sum_{i\neq j}^{n}\left\{(1-n^{-1})\mathsf{D}^{\otimes 2r}\bar{K}_{\mathbf{H}} - 2\mathsf{D}^{\otimes 2r}K_{\mathbf{H}}\right\}(\mathbf{X}_i - \mathbf{X}_j),
\end{aligned}$$

where $\bar{K} = K * K$. To check that these two CV expressions coincide, take into account that $\mathsf{D}^{\otimes 2r}K_{\mathbf{H}} * K_{\mathbf{H}} = \mathsf{D}^{\otimes 2r}(K_{\mathbf{H}} * K_{\mathbf{H}}) = \mathsf{D}^{\otimes 2r}(K * K)_{\mathbf{H}} = \mathsf{D}^{\otimes 2r}\bar{K}_{\mathbf{H}}$, so that using some properties of the Kronecker product and the vec operator (see, e.g., Magnus and Neudecker, 1999), the sum of the diagonal terms in the first $\operatorname{CV}_r(\mathbf{H})$ expression equals

$$\begin{aligned}
(-1)^r n^{-1}\operatorname{vec}^\top \mathbf{I}_{d^r}\mathsf{D}^{\otimes 2r}\bar{K}_{\mathbf{H}}(0) &= (-1)^r n^{-1}|\mathbf{H}|^{-1/2}\operatorname{vec}^\top \mathbf{I}_{d^r}(\mathbf{H}^{-1/2})^{\otimes 2r}\mathsf{D}^{\otimes 2r}\bar{K}(0)\\
&= (-1)^r n^{-1}|\mathbf{H}|^{-1/2}\operatorname{vec}^\top (\mathbf{H}^{-1})^{\otimes r}\mathsf{D}^{\otimes 2r}\bar{K}(0)\\
&= n^{-1}|\mathbf{H}|^{-1/2}\operatorname{vec}^\top (\mathbf{H}^{-1})^{\otimes r}\operatorname{vec}\mathbf{R}(\mathsf{D}^{\otimes r}K)\\
&= n^{-1}|\mathbf{H}|^{-1/2}\operatorname{tr}\left((\mathbf{H}^{-1})^{\otimes r}\mathbf{R}(\mathsf{D}^{\otimes r}K)\right)
\end{aligned}$$

where the third line follows by noting that

$$\operatorname{vec}\mathbf{R}(\mathsf{D}^{\otimes r}K) = (-1)^r\int_{\mathbb{R}^d}\mathsf{D}^{\otimes 2r}K(\boldsymbol{x})K(\boldsymbol{x})d\boldsymbol{x} = (-1)^r\mathsf{D}^{\otimes 2r}\bar{K}(0).$$

Surely the simplest formulation for CV (useful for implementation purposes) is

$$\begin{aligned}
\operatorname{CV}_r(\mathbf{H}) = (-1)^r|\mathbf{H}|^{-1/2}\operatorname{vec}^\top (\mathbf{H}^{-1})^{\otimes r}&\left\{n^{-2}\sum_{i,j=1}^{n}\mathsf{D}^{\otimes 2r}\bar{K}\left(\mathbf{H}^{-1/2}(\mathbf{X}_i - \mathbf{X}_j)\right)\right.\\
&\left. - 2[n(n-1)]^{-1}\sum_{i\neq j}\mathsf{D}^{\otimes 2r}K\left(\mathbf{H}^{-1/2}(\mathbf{X}_i - \mathbf{X}_j)\right)\right\}.
\end{aligned}$$

We denote by $\hat{\mathbf{H}}_{\mathrm{CV},r}$ the bandwidth matrix in $\mathcal{F}$ which minimizes $\operatorname{CV}_r(\mathbf{H})$.

3.2. Plug-in method

Plug-in (PI) bandwidth selection techniques are based on estimating the unknown quantities that appear in an asymptotic error formula and minimizing the resulting empirical criterion. Basic plug-in selectors for univariate density

estimation are described in Park and Marron (1990) and Sheather and Jones (1991). In the multivariate case, introducing the vector integrated density derivative functional, defined as

$$\boldsymbol{\psi}_{2r} = \int \mathsf{D}^{\otimes 2r} f(\boldsymbol{x}) f(\boldsymbol{x}) \, d\boldsymbol{x} = (-1)^r \operatorname{vec} \mathbf{R}(\mathsf{D}^{\otimes r} f),$$

allows us to rewrite the AMISE formula (2) for the r-th derivative as

$$\begin{aligned}
\mathrm{AMISE}_r(\mathbf{H}) = {} & n^{-1} |\mathbf{H}|^{-1/2} \operatorname{tr} \left((\mathbf{H}^{-1})^{\otimes r} \mathbf{R}(\mathsf{D}^{\otimes r} K) \right) \\
& + (-1)^r \tfrac{m_2(K)^2}{4} \boldsymbol{\psi}_{2r+4}^\top \left(\operatorname{vec} \mathbf{I}_{d^r} \otimes (\operatorname{vec} \mathbf{H})^{\otimes 2} \right).
\end{aligned}$$

Thus, the plug-in bandwidth selector $\hat{\mathbf{H}}_{\mathrm{PI},r}$ is defined to be the bandwidth in $\mathcal{F}$ minimizing the criterion

$$\begin{aligned}
\mathrm{PI}_r(\mathbf{H}) = {} & n^{-1} |\mathbf{H}|^{-1/2} \operatorname{tr} \left((\mathbf{H}^{-1})^{\otimes r} \mathbf{R}(\mathsf{D}^{\otimes r} K) \right) \\
& + (-1)^r \tfrac{m_2(K)^2}{4} \hat{\boldsymbol{\psi}}_{2r+4}^\top \left(\operatorname{vec} \mathbf{I}_{d^r} \otimes (\operatorname{vec} \mathbf{H})^{\otimes 2} \right),
\end{aligned}$$

where $\hat{\boldsymbol{\psi}}_{2r+4}$ is a suitable estimator of $\boldsymbol{\psi}_{2r+4}$.

Chacón and Duong (2010) analyzed the problem of estimating $\boldsymbol{\psi}_{2r}$ for an arbitrary r. Noting that $\boldsymbol{\psi}_{2r} = \mathbb{E}[\mathsf{D}^{\otimes 2r} f(\mathbf{X})]$, they proposed the kernel estimator

$$\hat{\boldsymbol{\psi}}_{2r}(\mathbf{G}) = n^{-1} \sum_{i=1}^n \mathsf{D}^{\otimes 2r} \hat{f}_{n\mathbf{G}}(\mathbf{X}_i) = n^{-2} \sum_{i,j=1}^n \mathsf{D}^{\otimes 2r} L_{\mathbf{G}}(\mathbf{X}_i - \mathbf{X}_j),$$

using a kernel L with pilot bandwidth $\mathbf{G}$, possibly different from K and $\mathbf{H}$. For the selection of the pilot bandwidth matrix $\mathbf{G}$, the same authors showed that the leading term of the mean squared error $\mathbb{E}\left[\|\hat{\boldsymbol{\psi}}_{2r}(\mathbf{G}) - \boldsymbol{\psi}_{2r}\|^2 \right]$ is given by the squared norm of the asymptotic bias vector

$$\boldsymbol{\omega}_{\mathrm{PI},2r}(\mathbf{G}) = n^{-1} |\mathbf{G}|^{-1/2} (\mathbf{G}^{-1/2})^{\otimes 2r} \mathsf{D}^{\otimes 2r} L(0) + \tfrac{m_2(L)}{2} (\operatorname{vec}^\top \mathbf{G} \otimes \mathbf{I}_{d^{2r}}) \boldsymbol{\psi}_{2r+2}, \tag{3}$$

so that the asymptotically optimal choice of the pilot bandwidth for the estimation of $\boldsymbol{\psi}_{2r}$ is $\mathbf{G}_{\mathrm{PI},2r} = \operatorname{argmin}_{\mathbf{G} \in \mathcal{F}} \|\boldsymbol{\omega}_{\mathrm{PI},2r}(\mathbf{G})\|^2$, which depends on $\boldsymbol{\psi}_{2r+2}$.

Hence, to select the pilot bandwidth $\mathbf{G}$ from the data we could substitute $\boldsymbol{\psi}_{2r+2}$ by another kernel estimator in (3) and minimize the squared norm of the resulting vector, but of course then the optimal bandwidth for the kernel estimator of ψ_{2r+2} depends on $\boldsymbol{\psi}_{2r+4}$, and so on. The usual strategy to overcome this problem is to use an m-stage algorithm as described in Chacón and Duong (2010), involving m successive kernel functional estimations with the initial bandwidth matrix chosen on the basis of a normal scale approximation. The resulting bandwidth obtained by minimizing $\mathrm{PI}_r(\mathbf{H})$ when $\boldsymbol{\psi}_{2r+4}$ is estimated using an m-stage algorithm is called an m-stage plug-in bandwidth selector for the r-th derivative.

3.3. Smoothed cross validation method

The smoothed cross validation (SCV) methodology for univariate density estimation was introduced in Hall, Marron and Park (1992), and a thorough study of this technique was made in Jones, Marron and Park (1991). However, it has not been until recently that its multivariate counterpart has been developed, in Duong and Hazelton (2005b) and Chacón and Duong (2011), and its use for univariate density derivative estimation has been explored (see Dobrovidov and Rud'ko, 2010).

A possible derivation of the SCV criterion for the problem of multivariate density derivative estimation is based on the approximation of the MISE obtained by replacing the exact integrated variance in equation (1) by its asymptotic approximation (the first term), while keeping the exact form for the integrated squared bias, so that $\text{MISE}_r(\mathbf{H}) \simeq \text{MISE2}_r(\mathbf{H})$ with

$$\text{MISE2}_r(\mathbf{H}) = n^{-1}|\mathbf{H}|^{-1/2}\operatorname{tr}\left((\mathbf{H}^{-1})^{\otimes r}\mathbf{R}(\mathsf{D}^{\otimes r}K)\right) + \operatorname{tr}\mathbf{R}^*(\bar{\Delta}_{\mathbf{H}}, \mathsf{D}^{\otimes r}f),$$

where $\Delta_{\mathbf{H}} = K_{\mathbf{H}} - K_0$ (here K_0 denotes the Dirac delta function) and $\bar{\Delta}_{\mathbf{H}} = \Delta_{\mathbf{H}} * \Delta_{\mathbf{H}} = \bar{K}_{\mathbf{H}} - 2K_{\mathbf{H}} + K_0$. The SCV criterion is obtained by replacing the unknown target $\mathsf{D}^{\otimes r}f$ in the MISE2 formula with a pilot estimator $\mathsf{D}^{\otimes r}\hat{f}_{\mathbf{G}}(\boldsymbol{x}) = n^{-1}\sum_{i=1}^{n}\mathsf{D}^{\otimes r}L_{\mathbf{G}}(\boldsymbol{x} - \mathbf{X}_i)$, leading to

$$\begin{aligned}
\text{SCV}_r(\mathbf{H}) &= n^{-1}|\mathbf{H}|^{-1/2}\operatorname{tr}\left((\mathbf{H}^{-1})^{\otimes r}\mathbf{R}(\mathsf{D}^{\otimes r}K)\right) \\
&\quad + (-1)^r n^{-2}(\operatorname{vec}^{\top}\mathbf{I}_{d^r})\sum_{i,j=1}^{n}\bar{\Delta}_{\mathbf{H}} * \mathsf{D}^{\otimes 2r}\bar{L}_{\mathbf{G}}(\mathbf{X}_i - \mathbf{X}_j),
\end{aligned}$$

where $\bar{L} = L * L$. When all the $\mathbf{X}_i$ are distinct and the diagonal terms $(i = j)$ are omitted in the previous sum it can be shown, using the properties of the Dirac delta function (see, e.g., Gel'fand and Shilov, 1964, Chapter I.2), that the SCV criterion coincides with the CV criterion for $\mathbf{G} = 0$.

The minimizer of $\text{SCV}_r(\mathbf{H})$ is defined to be $\hat{\mathbf{H}}_{\text{SCV},r}$. Its value depends on the pilot selector $\mathbf{G}$. Chacón and Duong (2011) showed that in the case $r = 0$ the leading term of the mean squared error $\mathbb{E}\|\operatorname{vec}(\hat{\mathbf{H}}_{\text{SCV},r} - \mathbf{H}_{\text{MISE},r})\|^2$ is given by the squared norm $\|\boldsymbol{\omega}_{\text{SCV},2r+4}(\mathbf{G})\|^2$ where $\boldsymbol{\omega}_{\text{SCV},2r+4}(\mathbf{G})$ is the same as the aforementioned $\boldsymbol{\omega}_{\text{PI},2r+4}(\mathbf{G})$ except that L is replaced by $\bar{L}$. Thus it is straightforward to define, analogously to the plug-in algorithm, the required optimal k-th stage pilot bandwidth of an m-stage SCV selector.

3.4. Convergence results

Let $\hat{\mathbf{H}}_r = \operatorname{argmin}_{\mathbf{H}\in\mathcal{F}}\widehat{\text{MISE}}_r(\mathbf{H})$ be an arbitrary data-based bandwidth selector, built up on the basis of an estimated criterion $\widehat{\text{MISE}}_r(\mathbf{H})$. Following Duong and Hazelton (2005a), $\hat{\mathbf{H}}_r$ is said to converge to $\mathbf{H}_{\text{MISE},r}$ at relative rate $n^{-\alpha}$ if

$$\operatorname{vec}(\hat{\mathbf{H}}_r - \mathbf{H}_{\text{MISE},r}) = O_P(n^{-\alpha}\mathbf{J}_{d^2})\operatorname{vec}\mathbf{H}_{\text{MISE},r}$$

where O_P denotes element-wise order in probability and $\mathbf{J}_{d^2}$ is the $d^2 \times d^2$ matrix of ones. This order in probability statement can be difficult to derive directly. The next lemma provides a more tractable indirect method of calculating convergence rates.

Lemma 1. *Suppose that assumptions (A1)–(A3) given in the appendix hold. The discrepancy* $\mathrm{vec}(\hat{\mathbf{H}}_r - \mathbf{H}_{\mathrm{MISE},r})$ *is asymptotically equivalent to* $\mathsf{D}_{\mathbf{H}}[\widehat{\mathrm{MISE}}_r - \mathrm{MISE}_r](\mathbf{H}_{\mathrm{MISE},r})$, *where* $\mathsf{D}_{\mathbf{H}}$ *is shorthand for* $\partial/\partial \, \mathrm{vec}\, \mathbf{H}$. *Furthermore, the relative rate of convergence of* $\hat{\mathbf{H}}_r$ *is* $n^{-\alpha}$ *if*

$$
\mathbb{E}\{\mathsf{D}_{\mathbf{H}}[\widehat{\mathrm{MISE}}_r - \mathrm{MISE}_r](\mathbf{H}_{\mathrm{MISE},r})\}\mathbb{E}\{\mathsf{D}_{\mathbf{H}}[\widehat{\mathrm{MISE}}_r - \mathrm{MISE}_r](\mathbf{H}_{\mathrm{MISE},r})\}^\top
$$
$$
+ \mathrm{Var}\{\mathsf{D}_{\mathbf{H}}[\widehat{\mathrm{MISE}}_r - \mathrm{MISE}_r](\mathbf{H}_{\mathrm{MISE},r})\}
$$

can be expressed as $O(n^{-2\alpha}\mathbf{J}_{d^2}) \, \mathrm{vec}\, \mathbf{H}_{\mathrm{MISE},r} \, \mathrm{vec}^\top \mathbf{H}_{\mathrm{MISE},r}$.

The convergence rates of the three bandwidth selectors considered here are given in the following theorem, whose proof is deferred to the appendix.

Theorem 1. *Suppose that assumptions (A1)–(A5) given in the appendix hold. The relative rate of convergence to* $\mathbf{H}_{\mathrm{MISE},r}$ *is* $n^{-d/(2d+4r+8)}$ *for the cross validation selector* $\hat{\mathbf{H}}_{\mathrm{CV},r}$, *and* $n^{-2/(d+2r+6)}$ *for the plug-in selector* $\hat{\mathbf{H}}_{\mathrm{PI},r}$ *and the smoothed cross validation selector* $\hat{\mathbf{H}}_{\mathrm{SCV},r}$ *when* $d \geq 2$.

Jones (1992) computed the relative rate of convergence for the CV and PI selectors for the estimation of a single partial derivative, using a single-parameter bandwidth matrix (i.e., $\mathbf{H} \in \mathcal{I}$). The previous theorem shows that the unconstrained CV bandwidth attains the same rate as its constrained counterpart, yet with added flexibility that should be captured in the constant coefficient of the asymptotic expression, although the computation of an explicit form for this coefficient does not seem possible in general.

The convergence rate of the PI selector is $n^{-(2+\min\{2,d/2\})/(d+2r+6)}$ within the single-parameter bandwidth class $\mathcal{I}$, yielding a slightly faster convergence to the optimal constrained bandwidth. As explained in Chacón and Duong (2010, 2011) for the density case, this is due to the fact that the very special cancellation in the bias term which is achievable when using a single-parameter bandwidth is not possible in general for the unconstrained estimator. Nevertheless, the aforementioned papers showed that this slight loss in convergence rate terms is negligible in practice as compared with the fact that the targeted constrained optimal bandwidth is usually much less efficient than the unconstrained one (see also Section 4 below).

Theorem 1 also shows that the similarities noted in Chacón and Duong (2011) about the asymptotic properties of the PI and SCV methods for the density estimation problem persist for $r > 0$, since both selectors exhibit the same relative rate of convergence.

Jones, Marron and Sheather (1996, p. 406) exemplified how slow is the rate $n^{-1/10}$ of the CV selector for $d = 1$, $r = 0$ by noting that n has to be as large as $10^{10} = 10,000,000,000$ so that $n^{-1/10} = 0.1$. In the same spirit, to compare the

Comparison of the relative rate of convergence for the CV, PI and SCV selectors. For each combination of r, n and d in the table, the left entry in the corresponding cell shows $n^{-d/(2d+4r+8)}$ (CV selector) and the right entry $n^{-2/(d+2r+6)}$ (PI and SCV selectors) divided by $1000^{-1/6}$ (i.e. the rate for the CV selector with $n = 1000, d = 2, r = 0$)

		$d = 2$		$d = 3$		$d = 4$		$d = 5$	
		CV	PI/SCV	CV	PI/SCV	CV	PI/SCV	CV	PI/SCV
$r = 0$	$n = 10^3$	1.000	0.562	0.720	0.681	0.562	0.794	0.464	0.901
	$n = 10^4$	0.681	0.316	0.439	0.408	0.316	0.501	0.245	0.593
	$n = 10^5$	0.464	0.178	0.268	0.245	0.178	0.316	0.129	0.390
$r = 1$	$n = 10^3$	1.334	0.794	1.000	0.901	0.794	1.000	0.658	1.093
	$n = 10^4$	1.000	0.501	0.681	0.593	0.501	0.681	0.390	0.767
	$n = 10^5$	0.750	0.316	0.464	0.390	0.316	0.464	0.231	0.538
$r = 2$	$n = 10^3$	1.585	1.000	1.233	1.093	1.000	1.179	0.838	1.259
	$n = 10^4$	1.259	0.681	0.901	0.767	0.681	0.848	0.538	0.926
	$n = 10^5$	1.000	0.464	0.658	0.538	0.464	0.611	0.346	0.681

rates obtained in Theorem 1, Table 1 shows the values of $n^{-d/(2d+4r+8)}$ (CV) and $n^{-2/(d+2r+6)}$ (PI and SCV) divided by $1000^{-1/6}$, that is the rate for the CV selector $n = 1000, d = 2, r = 0$ which is used as a base case, for all the different combinations of $n = 10^3, 10^4, 10^5$, $d = 2, 3, 4, 5$ and $r = 0, 1, 2$. Ratios which are lower than 1 indicate the rate is faster than the base case, and ratios greater than 1 a slower rate. For $n = 10^3$, these ratios in Table 1 tend to be greater than 1, indicating that using this sample size will lead to a deteriorating convergence rate. On the other hand for the larger sample sizes, $n = 10^4, 10^5$, these ratios tend to be less than 1. This implies that convergence rates better than the CV selector for bivariate density estimation can be attained, even with higher dimensions and higher order derivatives, provided that sufficiently large (although still realistic) sample sizes are used. Of course this comparison only takes into account the asymptotic order of the convergence rates by ignoring the associated coefficients since explicit formulas for the latter are not available for $d \geq 2$. The finite sample behaviour of the bivariate case for moderate sample sizes is examined more closely in the next section.

4. Numerical study

4.1. Data-based algorithms

For most practical implementations the normal kernels are used, i.e. $K = L = \phi$. For $d \times d$ symmetric matrices $\mathbf{A}, \mathbf{B}$, and for $r, s \geq 0$, let

$$\eta_{2r,2s}(\boldsymbol{x}; \mathbf{A}, \mathbf{B}, \boldsymbol{\Sigma}) = [(\text{vec}^T \mathbf{A})^{\otimes r} \otimes (\text{vec}^T \mathbf{B})^{\otimes s}]\mathsf{D}^{\otimes 2r+2s}\phi_{\boldsymbol{\Sigma}}(\boldsymbol{x})$$

and write, for short, $\eta_{2r}(\boldsymbol{x}; \boldsymbol{\Sigma}) = \eta_{2r,0}(\boldsymbol{x}; \mathbf{I}_d, \mathbf{I}_d, \boldsymbol{\Sigma})$ and $\nu_r(\boldsymbol{\Sigma}) = (-1)^r \eta_{2r}(0; \boldsymbol{\Sigma})/\phi_{\boldsymbol{\Sigma}}(0)$.

Then the cross validation criterion can be rewritten as

$$\text{CV}_r(\mathbf{H})$$
$$= (-1)^r \left\{ n^{-2} \sum_{i,j=1}^{n} \eta_{2r}(\mathbf{X}_i - \mathbf{X}_j; 2\mathbf{H}) - 2[n(n-1)]^{-1} \sum_{i \neq j} \eta_{2r}(\mathbf{X}_i - \mathbf{X}_j; \mathbf{H}) \right\}.$$

Besides, the data-based m-stage selection algorithm for plug-in selectors is given by:

1. Initialize the m-th stage pilot selector to be the normal reference selector

$$\hat{\mathbf{G}}_{\mathrm{PI},2r+2m+2} = \left(\frac{2}{2r+2m+d+2}\right)^{2/(2r+2m+d+4)} 2\mathbf{S}n^{-2/(2r+2m+d+4)},$$

from Chacón, Duong and Wand (2011), where $\mathbf{S}$ is the sample variance of $\mathbf{X}_1,\ldots\mathbf{X}_n$.

2. For $k = m-1, m-2, \ldots, 1$, the optimal k-the stage pilot selector $\hat{\mathbf{G}}_{\mathrm{PI},2r+2k+2}$ is the minimizer of

$$\begin{aligned}
\|\hat{\boldsymbol{\omega}}_{\mathrm{PI},2r+2k+2}(\mathbf{G})\|^2 =\ & n^{-2}|\mathbf{G}|^{-1}(2\pi)^{-d}\mathrm{OF}(2r+2k+2)\nu_{r+k+1}(\mathbf{G}^{-2}) \\
& + (-1)^{r+k+1}(2\pi)^{-d/2}\mathrm{OF}(2r+2k+2)|\mathbf{G}|^{-1/2}n^{-3} \\
& \times \sum_{i,j=1}^{n} \eta_{2,2r+2k+2}(\mathbf{X}_i - \mathbf{X}_j; \mathbf{G}, \mathbf{G}^{-1}, \hat{\mathbf{G}}_{\mathrm{PI},2r+2k+4}) \\
& + \tfrac{1}{4}n^{-4}\Big[\sum_{i,j=1}^{n} \eta_{2,2r+2k+2}(\mathbf{X}_i - \mathbf{X}_j; \mathbf{G}, \mathbf{I}_d, \hat{\mathbf{G}}_{\mathrm{PI},2r+2k+4})\Big]^2,
\end{aligned}$$

where $\mathrm{OF}(2p) = (2p-1)(2p-3)\cdots 5 \cdot 3 \cdot 1$ for $p \in \mathbb{N}$. The numerical minimization over the class of positive-definite matrices is carried out as described in detail in Duong and Hazelton (2005b, Section 5.1).

3. The plug-in selector $\hat{\mathbf{H}}_{\mathrm{PI},r}$ is the minimizer of

$$\begin{aligned}
\mathrm{PI}_r(\mathbf{H}) =\ & n^{-1}|\mathbf{H}|^{-1/2}2^{-(d+r)}\pi^{-d/2}\nu_r(\mathbf{H}^{-1}) \\
& + (-1)^r(2n)^{-2}\sum_{i,j=1}^{n}\eta_{2,2r}(\mathbf{X}_i - \mathbf{X}_j; \mathbf{H}, \mathbf{I}_d, \hat{\mathbf{G}}_{\mathrm{PI},2r+4}).
\end{aligned}$$

The derivations of $\|\hat{\boldsymbol{\omega}}_{\mathrm{PI},2r+2k+2}(\mathbf{G})\|^2$ and $\mathrm{PI}_r(\mathbf{H})$ in the η functional form can be found in Chacón and Duong (2012). There it is also shown that, although it appears that these are less concise than the previous expressions, they facilitate efficient computation, both in terms of memory and execution time.

We observe that $\|\boldsymbol{\omega}_{\mathrm{SCV},2r+2k+2}\|^2$ is the same as $\|\boldsymbol{\omega}_{\mathrm{PI},2r+2k+2}\|^2$ except the three terms are multiplied by $2^{-d}, 2^{-d/2+1}$ and 4 respectively; since $\bar{\phi}(0) = \phi_{2\mathbf{I}}(0) = 2^{-d/2}\phi(0)$ and $m_2(\bar{\phi}) = 2m_2(\phi)$. Furthermore,

$$\begin{aligned}
\mathrm{SCV}_r(\mathbf{H}) =\ & n^{-1}|\mathbf{H}|^{-1/2}2^{-(d+r)}\pi^{-d/2}\nu_r(\mathbf{H}^{-1}) \\
& + (-1)^r n^{-2}\sum_{i,j=1}^{n}\big\{\eta_{2r}(\mathbf{X}_i - \mathbf{X}_j; 2\mathbf{H} + 2\mathbf{G}) \\
& - 2\eta_{2r}(\mathbf{X}_i - \mathbf{X}_j; \mathbf{H} + 2\mathbf{G}) + \eta_{2r}(\mathbf{X}_i - \mathbf{X}_j; 2\mathbf{G})\big\}.
\end{aligned}$$

So a data-based m-stage SCV selector is obtained from straightforward modifications of the PI selector algorithm above.
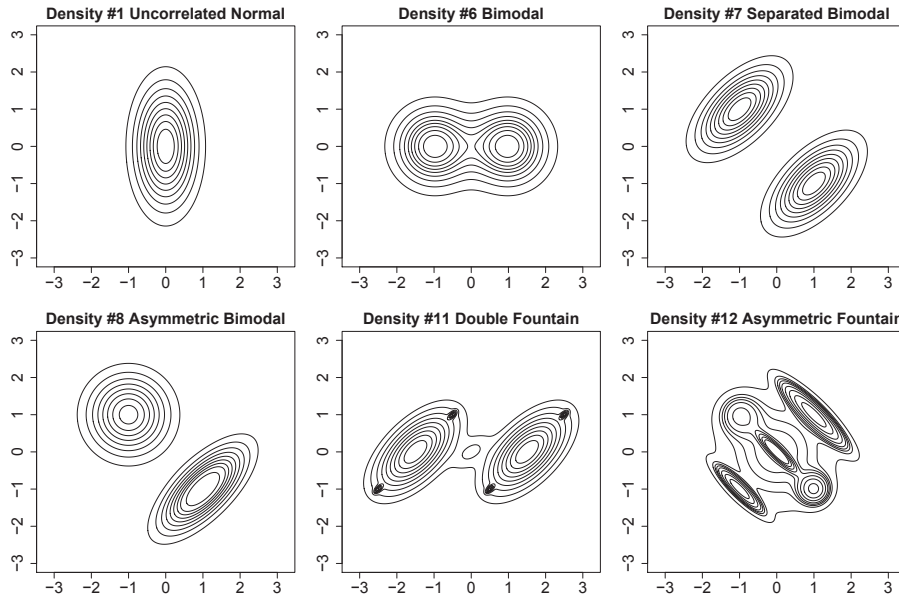
FIG 1. *Target bivariate normal mixture densities.*

4.2. Simulation study

The bandwidth selectors included in our simulation study were

- OR: oracle, i.e. the minimizer of the MISE for the target density
- NR: normal reference from Chacón, Duong and Wand (2011, Theorem 6), which is equal to $[4/(d + 2r + 2)]^{2/(d+2r+4)}\mathbf{S}n^{-2/(d+2r+4)}$
- CV: cross validation from Section 3.1
- PI: plug-in with 2-stage unconstrained pilots from Section 3.2
- SCV: smoothed cross validation with 2-stage unconstrained pilots from Section 3.3

We have developed efficient implementations of all these selectors and incorporated them into the existing R library `ks` (Duong, 2007). The target bivariate normal mixture densities that we considered are displayed in Figure 1. Their explicit definitions can be found in Chacón (2009).

For each selector and target density and for $r = 0, 1, 2$, we generated 100 samples of size $n = 1000$. The integrated squared error (ISE) between the resulting density estimate and the target density was computed as our measure of performance. The box plots of the log ISEs are shown in Figure 2. We also conducted the study for sample sizes $n = 400$ and $n = 4000$ but the conclusions extracted were much the same as for $n = 1000$ so we decided not to include these results here to avoid redundancies.
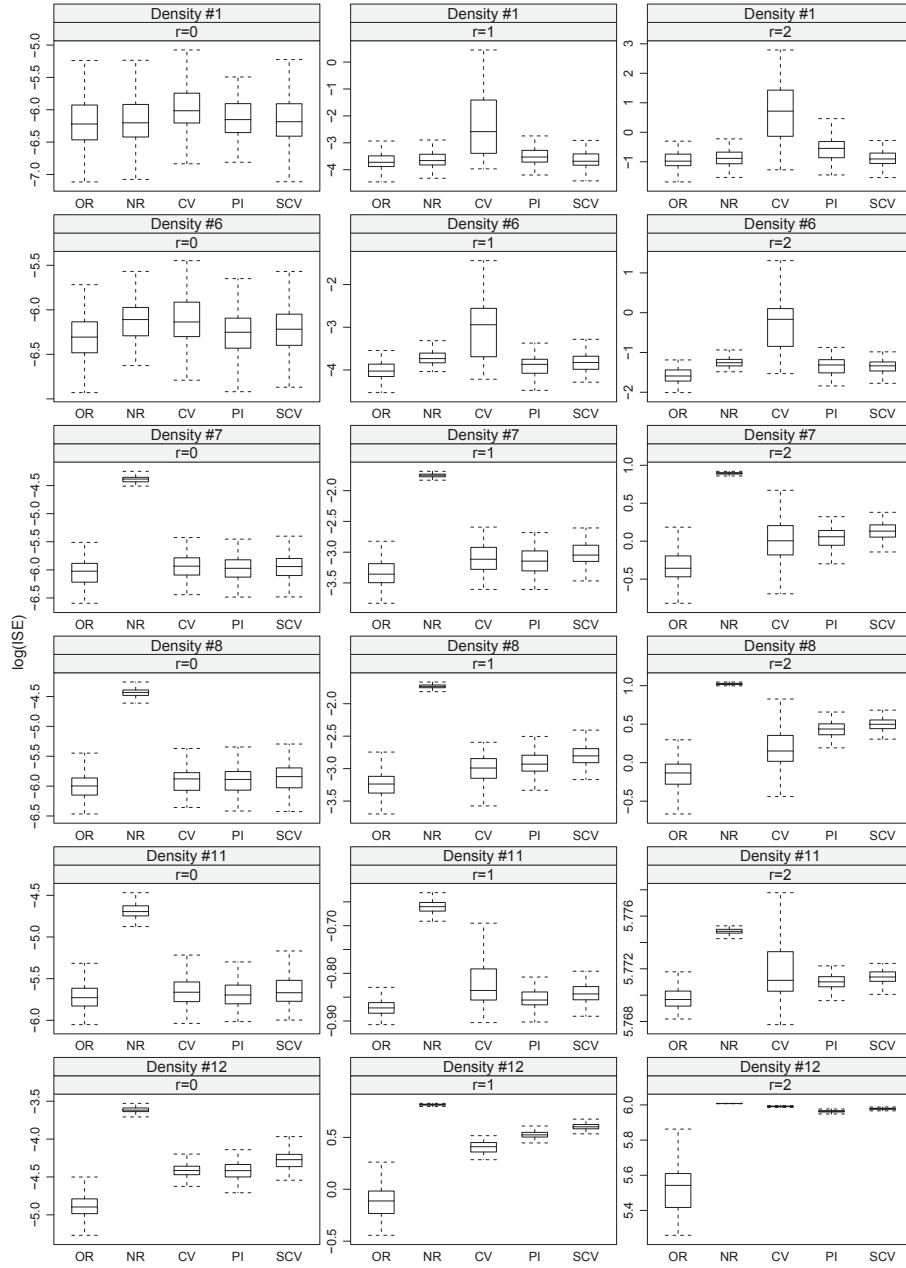
FIG 2. *Box plots of the logarithm of the ISEs for bandwidth selectors for $n = 1000, r = 0, 1, 2$ for the six bivariate target densities.*

By construction, the oracle selector (OR) is the best possible selector in terms of MISE given that the true target normal mixture density was used for its computation. As expected, it also has the uniformly lowest ISE. The normal reference selector (NR) was the only data-based selector previously available in the literature, and the results show that it is suitable only for density #1 since its ISEs for the other densities are uniformly higher than those of the other selectors. In line with other published simulation studies (Cao, Cuevas and González-Manteiga, 1994; Jones, Marron and Sheather, 1996), the CV selector displays larger variability in the ISEs than the PI and SCV selectors, though the former presents lower mean ISEs in some cases, e.g. density #12, $r = 0, 1$. We note also that the CV variability tends to increase with increasing r, whereas this is not observed for the two other hi-tech selectors. An anonymous referee drew our attention to the low variability of the introduced bandwidth selectors for this density #12, as compared with that of the oracle. This density has very complicated features, like modal regions of different shape and size, so this is the scenario where usually oversmoothing occurs, and we checked that this is indeed the case: the oracle tries hard to discover the true structure (hence its high variability), whereas all the data-driven bandwidths tend to consistently prefer a more conservative estimate, slightly oversmoothed. Given that the construction and theoretical properties of the PI and SCV selectors are similar, it is not surprising that their ISE performance is correspondingly similar for all the cases examined here. Either of these selectors would thus be our recommendation over the CV and NR selectors.

5. Applications to mean shift clustering

The so-called mean shift algorithm (Fukunaga and Hostetler, 1975) is an iterative procedure which, at every step, shifts the point obtained in the previous iteration in the direction of the density gradient, producing a convergent sequence that transports any initial value to a local maximum of the density along the steepest ascent path.

Specifically, the mean shift clustering algorithm can be described as follows: an initial point $\mathbf{Y}_0$ is transformed recursively to obtain a sequence defined by

$$\mathbf{Y}_{j+1} = \mathbf{Y}_j + \mathbf{A}\widehat{\mathsf{D}f}(\mathbf{Y}_j)/\hat{f}(\mathbf{Y}_j), \tag{4}$$

where $\hat{f}$ is an arbitrary density estimator, $\widehat{\mathsf{D}f}$ is an estimator of the density gradient, and $\mathbf{A}$ is a fixed $d \times d$ positive definite matrix, properly chosen to guarantee convergence of the sequence $(\mathbf{Y}_0, \mathbf{Y}_1, \ldots)$. This is easily recognized as a variant of the well-known gradient ascent algorithm employed to find the local maxima of a given function, but using the normalized density gradient (i.e., the density gradient divided by the density itself) instead of just the gradient in its definition. The advantages of using such a normalization are illustrated in Fukunaga and Hostetler (1975); Cheng (1995) and Comaniciu and Meer (2002); one of them is to accelerate the convergence of the resulting sequence for initial values of low density.

When kernel estimators are used in (4) the above procedure attains a particularly simple form. Assuming that the kernel K is a spherically symmetric function it follows that $K(\boldsymbol{x}) = \frac{1}{2}k(\|\boldsymbol{x}\|^2)$, where the function $k\colon \mathbb{R}_+ \to \mathbb{R}$ is known as the profile of K. Under the usual conditions that K is smooth and unimodal, its profile is decreasing so that $g(x) = -k'(x) \geq 0$. Therefore, noting that $\mathsf{D}K(\boldsymbol{x}) = -\boldsymbol{x}g(\|\boldsymbol{x}\|^2)$, the kernel density gradient estimator can be written as

$$\mathsf{D}\hat{f}_{\mathbf{H}}(\boldsymbol{x}) = n^{-1}|\mathbf{H}|^{-1/2}\mathbf{H}^{-1}\sum_{i=1}^{n}(\mathbf{X}_i - \boldsymbol{x})g\big((\boldsymbol{x} - \mathbf{X}_i)^{\top}\mathbf{H}^{-1}(\boldsymbol{x} - \mathbf{X}_i)\big)$$
$$= \mathbf{H}^{-1}\tilde{f}_{\mathbf{H}}(\boldsymbol{x})\mathbf{m}_{\mathbf{H}}(\boldsymbol{x}), \tag{5}$$

where $\tilde{f}_{\mathbf{H}}(\boldsymbol{x}) = n^{-1}|\mathbf{H}|^{-1/2}\sum_{i=1}^{n} g\big((\boldsymbol{x} - \mathbf{X}_i)^{\top}\mathbf{H}^{-1}(\boldsymbol{x} - \mathbf{X}_i)\big)$ can be understood as an unnormalized kernel estimator of f and the term

$$\mathbf{m}_{\mathbf{H}}(\boldsymbol{x}) = \frac{\sum_{i=1}^{n}\mathbf{X}_i g\big((\boldsymbol{x} - \mathbf{X}_i)^{\top}\mathbf{H}^{-1}(\boldsymbol{x} - \mathbf{X}_i)\big)}{\sum_{i=1}^{n} g\big((\boldsymbol{x} - \mathbf{X}_i)^{\top}\mathbf{H}^{-1}(\boldsymbol{x} - \mathbf{X}_i)\big)} - \boldsymbol{x}$$

is known as the *mean shift*. Thus, equation (5) can be re-arranged to note that $\mathbf{H}^{-1}\mathbf{m}_{\mathbf{H}}(\boldsymbol{x})$ provides a reasonable estimator of the normalized density gradient, and by taking $\mathbf{A} = \mathbf{H}$ in equation (4) it leads to the recursively defined sequence

$$\mathbf{Y}_{j+1} = \mathbf{Y}_j + \mathbf{m}_{\mathbf{H}}(\mathbf{Y}_j) = \frac{\sum_{i=1}^{n}\mathbf{X}_i g\big((\mathbf{Y}_j - \mathbf{X}_i)^{\top}\mathbf{H}^{-1}(\mathbf{Y}_j - \mathbf{X}_i)\big)}{\sum_{i=1}^{n} g\big((\mathbf{Y}_j - \mathbf{X}_i)^{\top}\mathbf{H}^{-1}(\mathbf{Y}_j - \mathbf{X}_i)\big)}. \tag{6}$$

When k is a convex and monotonically decreasing profile, and $\mathbf{H} = h^2\mathbf{I}_d$, Comaniciu and Meer (2002, Theorem 1) showed that the sequence $(\mathbf{Y}_0, \mathbf{Y}_1, \ldots)$ defined in this simple way converges to a local maximum of $\hat{f}_{\mathbf{H}}$, and their proof can be easily adapted to cover the case of an unconstrained $\mathbf{H}$ as well. The recursive formulation (6) was also motivated as an EM-type algorithm for mode finding in Li, Ray and Lindsay (2007), who proved its convergence under more general conditions.

Since the direction along which the data points are shifted, as well as the limit points of the sequences of successive locations (i.e., the solutions of $\mathsf{D}\hat{f}_{\mathbf{H}}(\boldsymbol{x}) = 0$), are directly related to the density gradient, our proposal is to take $\mathbf{H}$ in the mean shift algorithm as a bandwidth matrix selector for multivariate kernel density gradient estimation, using any of the methods introduced in Section 3. This choice is also supported by the results in Grund and Hall (1995) and Vieu (1996), where it was shown that the optimal bandwidth choice for estimating the mode of a density is closely related to the problem of density derivative estimation. Thus, the bandwidth choice is made with the goal of optimal identification of the density features in mind. This is in contrast with other proposals, as for example Comaniciu (2003), where a different criterion is taken into account to obtain an automatic variable-bandwidth selection algorithm.

When only a few iterations of the mean shift algorithm are performed, it is probably the case that convergence has not been reached yet. However, the procedure is still useful for other tasks. These include data filtering (Fukunaga and
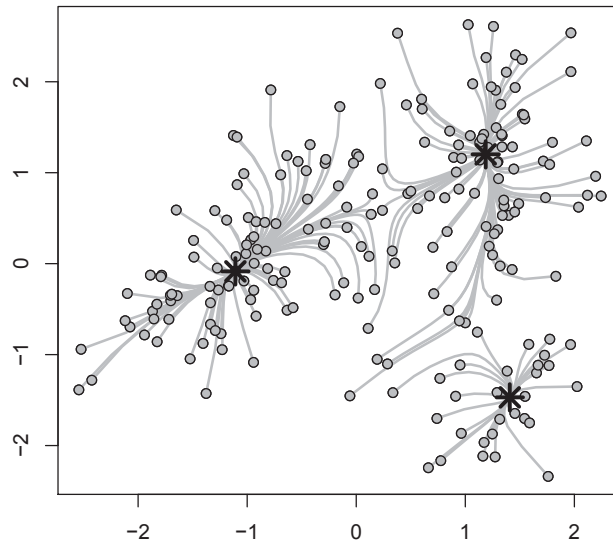
FIG 3. *Paths followed by the sample points as a result of the application of the mean shift algorithm. Sample of size $n = 210$ from a trimodal normal mixture density.*

Hostetler, 1975), which seeks to reduce the effect of noise in the determination of the geometric properties of a data set or in finding local principal curves, and also data sharpening (Choi and Hall, 1999; Hall and Minotte, 2002), which can be used to reduce the bias in kernel curve estimation and to adapt kernel estimators to pre-specified curve constraints.

The main statistical application of the mean shift procedure is for cluster analysis. For several additional applications in engineering, see Cheng (1995); Comaniciu and Meer (2002); Comaniciu, Ramesh and Meer (2003). When the mean shift algorithm is applied with any of the data points as starting value it induces a partition of the data in a natural way, by assigning the same cluster to all the data points that converge to the same local maximum. This is called *modal clustering* in Li, Ray and Lindsay (2007). Notice that this methodology does not require the number of clusters to be specified in advance, and that it allows clusters of arbitrary shape to be discovered. Moreover, since the mean shift algorithm can be applied with any starting point, it does not produce only a partition of the data, but a partition of the whole space.

To illustrate the use of mean shift clustering Figure 3 shows the result of applying the mean shift algorithm to a sample of size $n = 210$ from a trimodal normal mixture (the one labeled Trimodal III in Wand and Jones (1993)). The black bold stars show the location of the three modes found and the paths in grey starting from every data point depict their ascent towards their associated density mode.

5.1. Simulation results

As pointed out above, any of the bandwidth selection methods for kernel density gradient estimation introduced in Section 3 leads automatically to a new non-parametric clustering procedure via the mean shift algorithm. To explore the finite sample properties of these new proposals, their performance is compared here to other related existing methods.

Given the enormous amount of literature on clustering techniques, it would be impossible to include all the different clustering procedures in this comparison, so a brief selection of techniques similar to the one introduced here have been considered:

- CLUES algorithm (CLUstEring based on local Shrinking), proposed in Wang, Qiu and Zamar (2007), is an iterative algorithm closely related to the mean shift algorithm, but in which the shift is performed at each iteration by computing the coordinate-wise median of the K-nearest neighbors of the previous iteration point.
- PDFC algorithm (PDF Clustering), proposed in Azzalini and Torelli (2007), is also based on a kernel density estimate. Its high density regions are computed and the connected components of this regions are identified as sample clusters. The bandwidth used in the kernel estimate is just a diagonal normal scale rule-of-thumb for the density (not the density gradient), multiplied by a subjectively chosen shrinkage factor $3/4$ to correct for oversmoothing. We are aware of the existence of other clustering methods based on high density regions, as for instance Cuevas, Febrero and Fraiman (2001) or Rinaldo and Wasserman (2010), but decided to include in this admittedly limited study only the PDFC algorithm due to its simplicity.
- MCLUST algorithm (Mixture model CLUSTering), as surveyed in Fraley and Raftery (2002), is included in the comparison since it can be recognized as the parametric golden standard.

These three methodologies are compared with mean shift clustering using unconstrained bandwidth matrices for density gradient estimation obtained with: 1) the normal scale rule derived in Chacón, Duong and Wand (2011) (labeled NR), 2) the cross-validation bandwidth (labeled CV), 3) the plug-in bandwidth (labeled PI), and 4) the smoothed cross-validation bandwidth (labeled SCV).

The comparison is made along five test clustering problems, generated by five bivariate mixture densities that have been chosen to investigate the performance of the methods in a typical parametric setup (two normal mixture densities) and in situations with non-ellipsoidal cluster shapes, having also different scales. Specifically, the five mixture densities in the study are:

1. Trimodal III density from Wand and Jones (1993).
2. Quadrimodal density from Wand and Jones (1993).
3. 4-crescent model. This model is intended to mimic the distribution explored in Figure 7 of Comaniciu (2003). Since an explicit expression of the
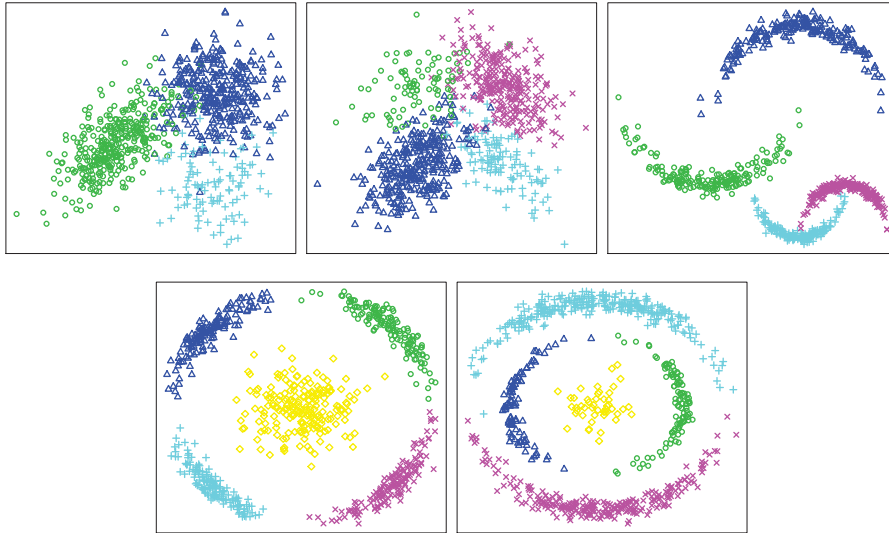
FIG 4. *Samples of size $n = 800$ from each of the models considered in the clustering simulation study. Different cluster membership is indicated with different plotting characters and colours.*

density function is not given there, our model has been generated as a suitable modification of Experiment 4 in Fukunaga (1990, p. 546). Namely, a bivariate random vector $\mathbf{X}$ is defined to have a crescent distribution with center $\mathbf{O} \in \mathbb{R}^2$, radius $r > 0$ and convexity indicator $\kappa \in \{0, 1\}$, denoted $C(\mathbf{O}, r, \kappa)$ if $\mathbf{X} = \mathbf{O} + (r\cos\Theta, (-1)^\kappa r\sin\Theta)^\top + \mathbf{U}$, where Θ is normally distributed with mean $\pi/2$ and variance $(\pi/6)^2$ and $\mathbf{U}$ is a bivariate centred normal vector with variance matrix $(r/20)^2 \mathbf{I}_2$. Then, the 4-crescent model is the equally weighted 4-component mixture density with components $C((-1, 1)^\top, 1, 1)$, $C((0, 0.5)^\top, 1, 0)$, $C((0, 0)^\top, 0.5, 1)$ and $C((0.5, -0.5)^\top, 0.5, 0)$.

4. Broken ring model. This model aims to reproduce the sampling scheme shown in Figure 3 in Wang, Qiu and Zamar (2007). Precisely, a bivariate random vector $\mathbf{X}$ is defined to have a standard half-crescent distribution with mean angle θ, denoted $HC(\theta)$ if $\mathbf{X} = (\cos\Theta, \sin\Theta)^\top + \mathbf{U}$, where Θ is normally distributed with mean θ and variance $(\pi/12)^2$ and $\mathbf{U}$ is a bivariate centred normal vector with variance matrix $(1/20)^2 \mathbf{I}_2$. Then, the broken ring model is the 5-component mixture density having a centred normal component with variance $(1/5)^2 \mathbf{I}_2$ and weight $1/4$, and four standard half-crescent components with equal weights $3/16$ and mean angles $\pi/4$, $3\pi/4$, $5\pi/4$ and $7\pi/4$, respectively.

5. Eye model. This model is a variation of the former. It is also a 5-component mixture density with a centred normal component with variance $(1/5)^2 \mathbf{I}_2$ as before, but with a weight $1/20$. The other 4 components are centred crescent distributions (i.e., $\mathbf{O} = (0, 0)^\top$), two of them with radius 1 and

TABLE 2
Average adjusted Rand index (ARI) for 100 simulation runs of size $n = 500$ of each distribution

	NR	CV	PI	SCV	CLUES	PDFC	MCLUST
Trimodal III	0.700	0.694	0.752	0.546	0.583	0.754	0.715
Quadrimodal	0.518	0.617	0.630	0.505	0.519	0.641	0.790
4-crescent	0.569	0.920	0.913	0.932	0.805	0.834	0.481
Broken ring	0.983	0.918	0.983	0.986	0.984	0.975	0.811
Eye	0.606	0.742	0.765	0.585	0.548	0.544	0.420

the two possible convexity indicators, respectively, having weight 1/8 each; and the other two with radius 1.5 and also the two possible convexity indicators, but with weight 7/20 each, and rotated 90 degrees.

A clearer picture of all these models is provided by Figure 4, which shows samples of size $n = 800$ for each of them.

In common with Azzalini and Torelli (2007); Wang, Qiu and Zamar (2007) and many others, the performance of each clustering method is measured through the adjusted Rand index (ARI), which was introduced by Hubert and Arabie (1985) as a corrected-for-chance version of the proportion of agreements between two partitions of a given data set. This index is the overall preferred accuracy measure in the simulation study of Milligan and Cooper (1986). An ARI value of 1 indicates that all estimated memberships are the same as the true memberships, whereas a value close to 0 indicates that the estimated cluster assignation does not differ much from random assignment. For the comparison, 100 samples of size 500 were drawn from each of the five test models, the data were clustered according to the seven methods in the study (four mean shift procedures plus CLUES, PDFC and MCLUST) and the ARI was computed to measure the performance of each method for each of these data sets. Table 2 presents the average ARI values obtained.

In view of Table 2, none of the methods compared is uniformly the best. In the group of the mean shift procedures, the use of the PI bandwidth seems to exhibit the best overall performance. The CV choice can be rated second best, with similar or even slightly (but not significantly) better average ARI in some cases. The SCV bandwidth shows an unexpectedly inferior performance for the normal mixture models, but it has an acceptable behaviour for the models with non-standard cluster shapes. Finally, the normal scale rule NR is clearly inferior in four out of the five models, but it performs surprisingly well for the broken ring model; since it is the least intensive method in computational terms, it could be useful at least to provide a quick initial analysis, especially in higher dimensions.

The comparison with the parametric method MCLUST followed the expected guidelines: for the normal mixture models MCLUST showed good results, especially for the difficult quadrimodal density, but it seems unable to adapt itself to the non-standard cluster shape situations. On the contrary, CLUES is not very powerful for a standard setup with ellipsoidal clusters, but seems to performs reasonably well for non-standard problems. Finally, PDFC shows remarkable

results in the simulation study, in spite of the ad hoc choice of the bandwidth in which it is based, and its performance is comparable to that of the best mean shift procedure, with the only exception of the eye model. Surely a more careful study of the bandwidth selection problem would improve the quality of the PDFC method further.

5.2. Real data examples

The mean shift algorithm in conjunction with the new proposed bandwidth selection rules was also applied to some real data sets. It is well-known that the kernel density estimator tends to produce spurious bumps (i.e., unimportant modes caused by a single observation) in the tails of the distribution, and that this problem seems enhanced in higher dimensions, due to the empty space phenomenon and the curse of dimensionality (see, for instance, Simonoff, 1996, Chapter 4). For real data sets, this may result in a number of data points forming singleton clusters after applying the mean shift algorithm.

Furthermore, in some applications the researcher may be interested in forming more homogeneous groups so that, say, insignificant groups of size less than $\alpha\%$ of the biggest group are not allowed in the outcome of the clustering algorithm. This goal can be achieved as follows: apply the mean shift algorithm to the whole data set and identify all the data points forming groups of size less than $\alpha\%$ of the biggest group, then leave those singular data points out of the estimation process in the mean shift algorithm and re-compute the data-based bandwidth and the density and density gradient estimators in (4) using only non-singular data points. Since the mean shift algorithm produces a partition of the whole space, these left-out data points can be naturally assigned to any of the corresponding newly obtained clusters. If this new assignment again contains insignificant clusters then iterate the process until the eventual partition satisfies the desired requirements. This correction is similar (although a little different) to the stage called "merging clusters based on the coverage rate" in Li, Ray and Lindsay (2007), and will be referred henceforth as *correction for insignificant groups*.

5.2.1. E.coli data

The *E. coli* data set is provided by the UCI machine learning database repository (Frank and Asuncion, 2010). The original data were contributed by Kenta Nakai at the Institute of Molecular and Cellular Biology of Osaka University. The data represent seven features calculated from the amino acid sequences of $n = 336$ E.coli proteins, classified in eight classes according to their localization sites, labeled imL (2 observations), omL (5), imS (2), om (20), pp (52), imU (35), im (77), cp (143). A more detailed description of this data set can be found in Horton and Nakai (1996). Since two of the original seven features are binary variables, only the remaining five continuous variables ($d = 5$), scaled to have unit variance, were retained for the cluster analysis.

The number of groups identified by the mean shift procedure with correction for insignificant groups (using $\alpha = 5\%$ as a default) was 5 for PI and SCV bandwidths, which is the natural choice if the insignificant clusters imL, omL and imS are merged into bigger groups. The mean shift algorithm found 6 groups using the NR bandwidth and 7 with the CV bandwidth. Since in this example the true cluster membership is available from the original data, it is also possible to compare the performance of the methods using the ARI. The ARIs for these configurations were 0.63 (NR bandwidth), 0.671 (CV), 0.667 (PI) and 0.559 (SCV). In contrast, CLUES and PDFC indicated a severely underestimated number of groups in the data, namely 3 and 2, respectively, and whereas CLUES obtains a remarkably high ARI anyway (0.697), the performance of PDFC is poor for this data set in ARI terms (0.386). MCLUST also gives a reasonable answer, with 6 groups and an ARI of 0.642.

5.2.2. Olive oil data

These data were introduced in Forina *et al.* (1983), and consist of eight chemical measurements on $n = 572$ olive oil samples from three regions of Italy. The three regions R1, R2 and R3 are further divided into nine areas, with areas A1 (25 observations), A2 (56), A3 (206) and A4 (36) in region R1 (totalling 323 observations); areas A5 (65) and A6 (33) in region R2 (totalling 98); and areas A7 (50), A8 (50) and A9 (51) in region R3 (totalling 151). Detailed cluster analyses of this data set are given in Stuetzle (2003) and Azzalini and Torelli (2007). Taking into account the compositional nature of these data, they were transformed following the guidelines in the latter reference, first dealing with the effect of rounding zeroes when the chemical measurement was below the instrument sensitivity level and then applying the additive log-ratio transform to place the data in a 7-dimensional Euclidean space (see Pawlowsky-Glahn and Buccianti, 2011, for a recent monograph on compositional data). Then, cluster analysis was carried out over the first five principal components of the scaled Euclidean variables.

The results of the analysis indicated that whereas some methods seemed to target the partition of the data into major regions, others tried hard to discover the sub-structure of areas. This was clearly recognized when the ARIs of the groupings were computed either with respect to one classification or the other. Naturally, if a method produced a grouping which was accurate with respect to major regions, it had lower ARI with respect to the division into areas.

CLUES, PDFC and the mean shift algorithm using the NR bandwidth clearly favoured grouping the data into major categories. The PDFC method obtained a remarkable ARI of 0.841 by clustering the data into 3 groups, whereas CLUES only found 2 groups resulting in an ARI of 0.680. Using the NR bandwidth the mean shift algorithm achieved an ARI of 0.920 with respect to the true grouping into major regions; it correctly identified all the data points in regions R1 and R2, although region R3 appeared divided into several subregions.

In contrast, MCLUST and the mean shift algorithm combined with all the more sophisticated bandwidth selectors tended to produce groupings closer to
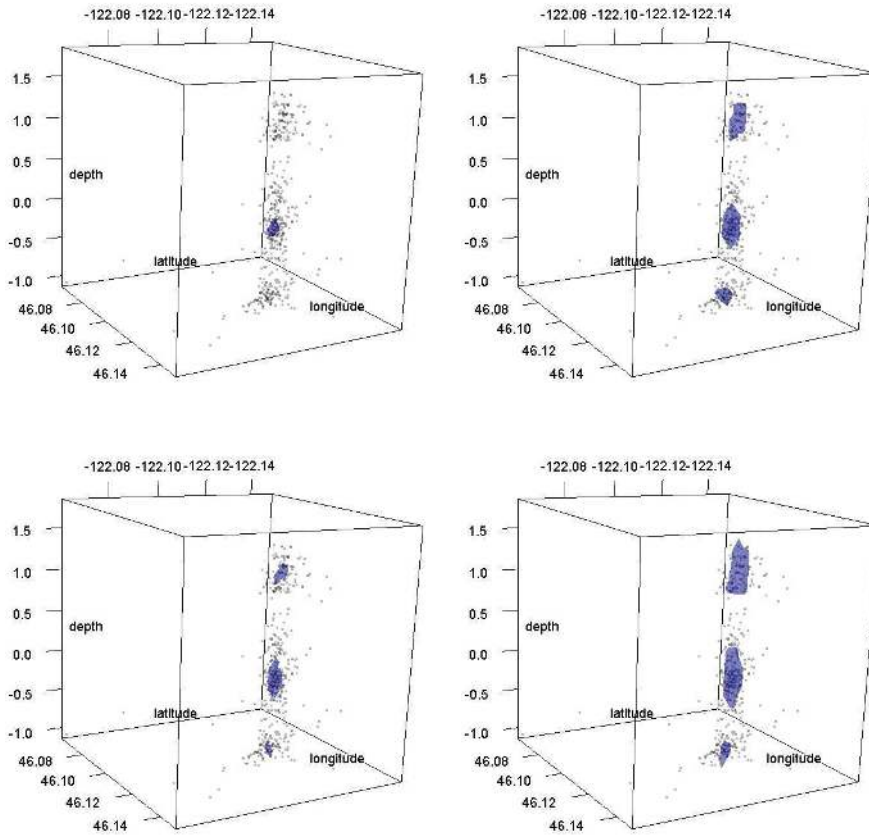
FIG 5. *Significant negative curvature regions (in blue). (Upper left) Plug-in selector $r = 0$. (Upper right) Plug-in selector $r = 2$. (Lower left) SCV selector $r = 0$. (Lower right) SCV selector $r = 2$. The significant curvature regions or modal regions are more clearly delimited from the surrounding scatter point cloud with the selectors corresponding to second derivative.*

the assignment into smaller areas. MCLUST showed the existence of 8 groups and achieved an ARI of 0.739 with respect to the true distribution into areas. The mean shift analyses with the CV, PI and SCV bandwidths all found 7 groups, leading to ARIs of 0.741 (CV bandwidth), 0.791 (PI) and 0.782 (SCV).

6. Applications to bump-hunting with feature significance

It is not always easy to interpret visually estimates of multivariate derivatives. To assist us, we use the significant negative curvature regions of Duong *et al.* (2008), defined as the set containing the values of $\boldsymbol{x} \in \mathbb{R}^d$ such that the null hypothesis that the Hessian $\mathsf{H}f(\boldsymbol{x})$ is positive definite is significantly rejected. The

appropriate kernel test statistic, null distribution and adjustment for multiple testing is outlined in Duong *et al.* (2008) and implemented in the `feature` library in R. Significant negative curvature regions corresponds to a modal region in the density function, and hence a local maxima in data density. These authors focused on the scale space approach of smoothing and so did not develop optimal bandwidth selectors for their density derivative estimates.

Here, we compare the significant curvature regions obtained using a usual $r = 0$ bandwidth selector to those with an $r = 2$ optimal bandwidth in Figure 5 on the earthquake data from Scott (1992). The recorded measurements are the latitude and longitude (in degrees) and depth (in km) of epicenters of 510 earthquakes. Here, negative latitude indicates west of the International Date Line, and negative depth indicates distances below the Earth's surface. The depth is transformed using –log(–depth). For these transformed data, we use PI selectors $\mathbf{H}_{\mathrm{PI},0}$ and $\mathbf{H}_{\mathrm{PI},2}$ and SCV selectors $\mathbf{H}_{\mathrm{SCV},0}$ and $\mathbf{H}_{\mathrm{SCV},2}$.

As expected from asymptotic theory, bandwidths for Hessian estimation are larger in magnitude than bandwidths for density estimation. Moreover only the central modal region is present using $\mathbf{H}_{\mathrm{PI},0}$, whereas with $\mathbf{H}_{\mathrm{PI},2}$, the three local modal regions are more clearly delimited from the surrounding space, confirming the three modes obtained with subjective bandwidth selection by Scott (1992).

Acknowledgements

Appendix A: Proofs

Henceforth the following assumptions are made:

(A1) K is a symmetric d-variate density such that $\int \boldsymbol{xx}^\top K(\boldsymbol{x})\,d\boldsymbol{x} = m_2(K)\mathbf{I}_d$ and all its partial derivatives up to order $2r + 1$ are bounded, continuous and square integrable.

(A2) f is a density function with all its partial derivatives up to order $2r + 6$ bounded, continuous and square integrable.

(A3) $\mathbf{H} = \mathbf{H}_n$ is a sequence of bandwidth matrices such that all entries of $n^{-1}|\mathbf{H}|^{-1/2}(\mathbf{H}^{-1})^{\otimes r}$ and $\mathbf{H}$ tend to zero as $n \to \infty$.

These do not form a minimal set of assumptions, but they serve as useful starting point for the results that we subsequently develop. Besides, in this section integrals without any integration limits are assumed to be integrated over the appropriate Euclidean space. We also assume that suitable regularity conditions are satisfied so that the exchange of term-by-term integration and differentiation of Taylor expansions are well-defined.

Proof of Lemma 1. Reasoning as in Lemma 1 in Duong and Hazelton (2005a), it follows that $\mathrm{vec}(\hat{\mathbf{H}}_r - \mathbf{H}_{\mathrm{MISE},r})$ is asymptotically equivalent to

$$-[\mathsf{D}_{\mathbf{H}}^2 \mathrm{MISE}_r(\mathbf{H}_{\mathrm{MISE},r})]^{-1}\mathsf{D}_{\mathbf{H}}[\widehat{\mathrm{MISE}}_r - \mathrm{MISE}_r](\mathbf{H}_{\mathrm{MISE},r}),$$

where $\mathsf{D}_{\mathbf{H}}^2 = \partial^2/(\partial\,\mathrm{vec}\,\mathbf{H}\partial\,\mathrm{vec}^\top\mathbf{H})$ denotes the Hessian operator corresponding to $\mathsf{D}_{\mathbf{H}}$. Therefore, it suffices to show that $\mathsf{D}_{\mathbf{H}}^2 \mathrm{MISE}_r(\mathbf{H}_{\mathrm{MISE},r}) = O(\mathbf{J}_{d^2})$. But for any $\mathbf{H}$ with entries of order $O(n^{-2/(d+2r+4)})$, as $\mathbf{H}_{\mathrm{MISE},r}$, the results in Chacón, Duong and Wand (2011) imply that $\mathrm{MISE}_r(\mathbf{H})$ is equivalent to $\mathrm{AMISE}_r(\mathbf{H})$ and, moreover, the smoothness assumptions ensure that $\mathsf{D}_{\mathbf{H}}^2 \mathrm{MISE}_r(\mathbf{H})$ is of the same order as $\mathsf{D}_{\mathbf{H}}^2 \mathrm{AMISE}_r(\mathbf{H})$. And it is not hard to show that for the asymptotic integrated squared bias term we have $\mathsf{D}_{\mathbf{H}}^2\big\{\boldsymbol{\psi}_{2r+4}^\top\big(\mathrm{vec}\,\mathbf{I}_{d^r} \otimes (\mathrm{vec}\,\mathbf{H})^{\otimes 2}\big)\big\} = O(\mathbf{J}_{d^2})$ and similarly for the asymptotic integrated variance, thus finishing the proof. $\qquad\square$

A.1. Convergence rate for the CV bandwidth

Lemma 1 shows that $\mathrm{vec}(\hat{\mathbf{H}}_{\mathrm{CV},r} - \mathbf{H}_{\mathrm{MISE},r})$ is asymptotically equivalent to $\mathsf{D}_{\mathbf{H}}[\mathrm{CV}_r - \mathrm{MISE}_r](\mathbf{H}_{\mathrm{MISE},r})$. Since $\mathbb{E}[\mathrm{CV}_r(\mathbf{H})] = \mathrm{MISE}_r(\mathbf{H}) - \mathrm{tr}\,\mathbf{R}(\mathsf{D}^{\otimes r}f)$ for all $\mathbf{H}$, it follows that the order of $\mathrm{vec}(\hat{\mathbf{H}}_{\mathrm{CV},r} - \mathbf{H}_{\mathrm{MISE},r})$ is given by the (root) order of

$$\mathrm{Var}\big\{\mathsf{D}_{\mathbf{H}}[\mathrm{CV}_r - \mathrm{MISE}_r](\mathbf{H}_{\mathrm{MISE},r})\big\}$$

$$\sim \mathrm{Var}\left\{[n(n-1)]^{-1}\sum_{i\neq j}^n \mathsf{D}_{\mathbf{H}}\big[\,\mathrm{vec}^\top(\mathbf{H}^{-1})^{\otimes r}(\mathsf{D}^{\otimes 2r}\tilde{K})_{\mathbf{H}}(\mathbf{X}_i - \mathbf{X}_j)\big]\Big|_{\mathbf{H}=\mathbf{H}_{\mathrm{MISE},r}}\right\},$$

where $\tilde{K} = K*K - 2K$. So denoting $\boldsymbol{\varphi}_{\mathbf{H}}(\boldsymbol{x}) = \mathsf{D}_{\mathbf{H}}\big[\,\mathrm{vec}^\top(\mathbf{H}^{-1})^{\otimes r}(\mathsf{D}^{\otimes 2r}\tilde{K})_{\mathbf{H}}(\boldsymbol{x})\big]$, by standard U-statistics theory the previous variance is of the same order as $4n^{-1}(\boldsymbol{\Xi}_1 - \boldsymbol{\Xi}_0) + 2n^{-2}\boldsymbol{\Xi}_2$, where

$$\boldsymbol{\Xi}_1 = \mathbb{E}[\boldsymbol{\varphi}_{\mathbf{H}}(\mathbf{X}_1 - \mathbf{X}_2)\boldsymbol{\varphi}_{\mathbf{H}}(\mathbf{X}_1 - \mathbf{X}_3)^\top]$$
$$\boldsymbol{\Xi}_2 = \mathbb{E}[\boldsymbol{\varphi}_{\mathbf{H}}(\mathbf{X}_1 - \mathbf{X}_2)\boldsymbol{\varphi}_{\mathbf{H}}(\mathbf{X}_1 - \mathbf{X}_2)^\top]$$
$$\boldsymbol{\Xi}_0 = \mathbb{E}[\boldsymbol{\varphi}_{\mathbf{H}}(\mathbf{X}_1 - \mathbf{X}_2)]\mathbb{E}[\boldsymbol{\varphi}_{\mathbf{H}}(\mathbf{X}_1 - \mathbf{X}_2)]^\top$$

with $\mathbf{H}$ of the order of $\mathbf{H}_{\mathrm{MISE},r}$, namely having all its entries of order $O(n^{-2/(d+2r+4)})$. The following lemma provides an explicit expression for the function $\boldsymbol{\varphi}_{\mathbf{H}}(\boldsymbol{x})$ that will be helpful to evaluate $\boldsymbol{\Xi}_p, p = 0, 1, 2$.

Lemma 2. *The function $\boldsymbol{\varphi}_{\mathbf{H}}(\boldsymbol{x})$ can be explicitly expressed as $\boldsymbol{\varphi}_{\mathbf{H}}(\boldsymbol{x}) = \mathbf{A}(\mathsf{D}^{\otimes 2r}\tilde{K})_{\mathbf{H}}(\boldsymbol{x}) + \mathbf{B}\boldsymbol{\rho}_{\mathbf{H}}(\boldsymbol{x})$ where the function $\boldsymbol{\rho}\colon \mathbb{R}^d \to \mathbb{R}^{d^{2r+2}}$ is given by $\boldsymbol{\rho}(\boldsymbol{x}) = (\mathbf{I}_{d^{2r}} \otimes \boldsymbol{x} \otimes \mathbf{I}_d)\mathsf{D}^{\otimes(2r+1)}\tilde{K}(\boldsymbol{x})$ and the matrices $\mathbf{A} \equiv \mathbf{A}(\mathbf{H}) \in \mathcal{M}_{d^2 \times d^{2r}}$, $\mathbf{B} \equiv \mathbf{B}(\mathbf{H}) \in \mathcal{M}_{d^2 \times d^{2r+2}}$ are defined as*

$$\mathbf{A} = -\tfrac{1}{2}(\mathrm{vec}^\top\mathbf{H}^{\otimes -r} \otimes \mathrm{vec}\,\mathbf{H}^{-1}) - r(\mathrm{vec}^\top\mathbf{H}^{\otimes -(r-1)} \otimes \mathbf{H}^{\otimes -2})$$
$$\mathbf{B} = -[\mathrm{vec}^\top\mathbf{H}^{\otimes -r} \otimes (\mathbf{H}^{1/2} \otimes \mathbf{H}^{1/2} + \mathbf{I}_d \otimes \mathbf{H})^{-1}]$$

where we understand that $\mathbf{H}^{\otimes -r} = (\mathbf{H}^{-1})^{\otimes r} = (\mathbf{H}^{\otimes r})^{-1}$.

Proof. Since $\operatorname{vec}^\top \mathbf{H}^{\otimes -r}(\mathsf{D}^{\otimes 2r}\tilde{K})_{\mathbf{H}}(\boldsymbol{x}) = |\mathbf{H}|^{-1/2}\operatorname{vec}^\top \mathbf{H}^{\otimes -r}\mathsf{D}^{\otimes 2r}\tilde{K}(\mathbf{H}^{-1/2}\boldsymbol{x})$, its differential is decomposed into three terms

$$d\big(\operatorname{vec}^\top \mathbf{H}^{\otimes -r}(\mathsf{D}^{\otimes 2r}\tilde{K})_{\mathbf{H}}(\boldsymbol{x})\big) = d(|\mathbf{H}|^{-1/2})\operatorname{vec}^\top \mathbf{H}^{\otimes -r}\mathsf{D}^{\otimes 2r}\tilde{K}(\mathbf{H}^{-1/2}\boldsymbol{x})$$
$$+ |\mathbf{H}|^{-1/2}(d\operatorname{vec}\mathbf{H}^{\otimes -r})^\top \mathsf{D}^{\otimes 2r}\tilde{K}(\mathbf{H}^{-1/2}\boldsymbol{x})$$
$$+ |\mathbf{H}|^{-1/2}\operatorname{vec}^\top \mathbf{H}^{\otimes -r}d\big(\mathsf{D}^{\otimes 2r}\tilde{K}(\mathbf{H}^{-1/2}\boldsymbol{x})\big).$$

From Chacón and Duong (2010), the differentials involved in the first two terms can be expressed as

$$d(|\mathbf{H}|^{-1/2}) = -\tfrac{1}{2}|\mathbf{H}|^{-1/2}(\operatorname{vec}^\top \mathbf{H}^{-1})d\operatorname{vec}\mathbf{H} \qquad \text{and}$$
$$d(\operatorname{vec}\mathbf{H}^{\otimes -r}) = -\boldsymbol{\Gamma}_r[\operatorname{vec}\mathbf{H}^{\otimes -(r-1)}\otimes \mathbf{H}^{\otimes -2}]d\operatorname{vec}\mathbf{H},$$

where $\boldsymbol{\Gamma}_r$ is a matrix such that $\boldsymbol{\Gamma}_r^\top\mathsf{D}^{\otimes 2r} = r\mathsf{D}^{\otimes 2r}$. For the third term,

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

since $[\mathsf{D}(\mathsf{D}^\top)^{\otimes 2r}]\operatorname{vec}\mathbf{H}^{\otimes -r} = \operatorname{vec}\big(\mathbf{I}_d[\mathsf{D}(\mathsf{D}^\top)^{\otimes 2r}]\operatorname{vec}\mathbf{H}^{\otimes -r}\big) = (\operatorname{vec}^\top \mathbf{H}^{\otimes -r}\otimes \mathbf{I}_d)\mathsf{D}^{\otimes 2r+1}$. Finally, using $d\operatorname{vec}\mathbf{H}^{-1/2} = -(\mathbf{H}^{1/2}\otimes\mathbf{H}+\mathbf{H}\otimes\mathbf{H}^{1/2})^{-1}d\operatorname{vec}\mathbf{H}$ from Chacón and Duong (2010), it follows that $d(\mathbf{H}^{-1/2}\boldsymbol{x}) = (\boldsymbol{x}^\top\otimes\mathbf{I}_d)d\operatorname{vec}\mathbf{H}^{-1/2} = -(\boldsymbol{x}^\top\mathbf{H}^{-1/2}\otimes\mathbf{I}_d)(\mathbf{I}_d\otimes\mathbf{H}+\mathbf{H}^{1/2}\otimes\mathbf{H}^{1/2})^{-1}d\operatorname{vec}\mathbf{H}$. Thus the derivative reads

$$\mathsf{D}_{\mathbf{H}}\big(\operatorname{vec}^\top \mathbf{H}^{\otimes -r}(\mathsf{D}^{\otimes 2r}\tilde{K})_{\mathbf{H}}(\boldsymbol{x})\big)$$
$$= -\tfrac{1}{2}|\mathbf{H}|^{-1/2}(\operatorname{vec}^\top \mathbf{H}^{\otimes -r}\otimes\operatorname{vec}\mathbf{H}^{-1})\mathsf{D}^{\otimes 2r}\tilde{K}(\mathbf{H}^{-1/2}\boldsymbol{x})$$
$$- r|\mathbf{H}|^{-1/2}(\operatorname{vec}^\top \mathbf{H}^{\otimes -(r-1)}\otimes\mathbf{H}^{\otimes -2})\mathsf{D}^{\otimes 2r}\tilde{K}(\mathbf{H}^{-1/2}\boldsymbol{x})$$
$$- |\mathbf{H}|^{-1/2}(\mathbf{H}^{1/2}\otimes\mathbf{H}^{1/2}+\mathbf{I}_d\otimes\mathbf{H})^{-1}(\mathbf{H}^{-1/2}\boldsymbol{x}\otimes\mathbf{I}_d)$$
$$\times(\operatorname{vec}^\top \mathbf{H}^{\otimes -r}\otimes\mathbf{I}_d)\mathsf{D}^{\otimes 2r+1}\tilde{K}(\mathbf{H}^{-1/2}\boldsymbol{x}).$$

The central factors of the third term on the right hand side can be rewritten as

$$(\mathbf{H}^{1/2}\otimes\mathbf{H}^{1/2}+\mathbf{I}_d\otimes\mathbf{H})^{-1}(\mathbf{H}^{-1/2}\boldsymbol{x}\otimes\mathbf{I}_d)(\operatorname{vec}^\top \mathbf{H}^{\otimes -r}\otimes\mathbf{I}_d)$$
$$= (\mathbf{H}^{1/2}\otimes\mathbf{H}^{1/2}+\mathbf{I}_d\otimes\mathbf{H})^{-1}(\operatorname{vec}^\top \mathbf{H}^{\otimes -r}\otimes\mathbf{H}^{-1/2}\boldsymbol{x}\otimes\mathbf{I}_d)$$
$$= \operatorname{vec}^\top \mathbf{H}^{\otimes -r}\otimes[(\mathbf{H}^{1/2}\otimes\mathbf{H}^{1/2}+\mathbf{I}_d\otimes\mathbf{H})^{-1}(\mathbf{H}^{-1/2}\boldsymbol{x}\otimes\mathbf{I}_d)]$$
$$= [\operatorname{vec}^\top \mathbf{H}^{\otimes -r}\otimes(\mathbf{H}^{1/2}\otimes\mathbf{H}^{1/2}+\mathbf{I}_d\otimes\mathbf{H})^{-1}](\mathbf{I}_{d^{2r}}\otimes\mathbf{H}^{-1/2}\boldsymbol{x}\otimes\mathbf{I}_d),$$

as desired. $\qquad\qquad\square$

We now return to the task of finding the asymptotic order of $\boldsymbol{\Xi}_0$, $\boldsymbol{\Xi}_1$ and $\boldsymbol{\Xi}_2$. For that, some preliminary notation is needed. For any real function a we denote its vector moment of order p as $\boldsymbol{\mu}_p(a) = \int_{\mathbb{R}^d}\boldsymbol{x}^{\otimes p}a(\boldsymbol{x})d\boldsymbol{x}$. For instance, Chacón and Duong (2011) showed that $\mu_0(\tilde{K}) = -1$, $\boldsymbol{\mu}_1(\tilde{K}) = \boldsymbol{\mu}_2(\tilde{K}) = \boldsymbol{\mu}_3(\tilde{K}) = 0$ and

$\boldsymbol{\mu}_4(\tilde{K}) = 6m_2(K)^2\boldsymbol{\mathcal{S}}_{d,4}(\text{vec}\,\mathbf{I}_d)^{\otimes 2}$, where $\boldsymbol{\mathcal{S}}_{d,r}$ denotes the symmetrizer matrix of order r (see Holmquist, 1985), defined as the (only) matrix such that premultiplying a Kronecker product of any r vectors in $\mathbb{R}^d$ by $\boldsymbol{\mathcal{S}}_{d,r}$ results in the average of all possible permutations of the r-fold product. We also introduce here the notation $\mathbf{K}_{m,n}$ for the commutation matrix of order $mn \times mn$ (Magnus and Neudecker, 1979).

So taking this into account, for the calculation of the asymptotic order of $\boldsymbol{\Xi}_0$, a fourth order Taylor expansion of $\mathsf{D}^{\otimes 2r}f(\boldsymbol{x} - \mathbf{H}^{1/2}\boldsymbol{z})$, in the form of Kollo and von Rosen (2005, Theorem 1.4.8) or Chacón, Duong and Wand (2011), gives

$$(\mathsf{D}^{\otimes 2r}\tilde{K})_{\mathbf{H}} * f(\boldsymbol{x})$$

$$= \int \mathsf{D}^{\otimes 2r}\tilde{K}(\boldsymbol{z})f(\boldsymbol{x} - \mathbf{H}^{1/2}\boldsymbol{z})d\boldsymbol{z}$$

$$= (\mathbf{H}^{1/2})^{\otimes 2r}\int \tilde{K}(\boldsymbol{z})\mathsf{D}^{\otimes 2r}f(\boldsymbol{x} - \mathbf{H}^{1/2}\boldsymbol{z})d\boldsymbol{z}$$

$$\sim (\mathbf{H}^{1/2})^{\otimes 2r}\sum_{p=0}^{4}\frac{(-1)^p}{p!}\int \tilde{K}(\boldsymbol{z})\left[\mathbf{I}_{d^{2r}} \otimes (\boldsymbol{z}^\top\mathbf{H}^{1/2})^{\otimes p}\right]\mathsf{D}^{\otimes 2r+p}f(\boldsymbol{x})d\boldsymbol{z}$$

$$= (\mathbf{H}^{1/2})^{\otimes 2r}\sum_{p=0}^{4}\frac{(-1)^p}{p!}\left[\mathbf{I}_{d^{2r}} \otimes (\boldsymbol{\mu}_p(\tilde{K})^\top(\mathbf{H}^{1/2})^{\otimes p})\right]\mathsf{D}^{\otimes 2r+p}f(\boldsymbol{x})$$

$$= -(\mathbf{H}^{1/2})^{\otimes 2r}\mathsf{D}^{\otimes 2r}f(\boldsymbol{x}) + \tfrac{1}{4}m_2(K)^2(\mathbf{H}^{1/2})^{\otimes 2r}\left[\mathbf{I}_{d^{2r}} \otimes ((\text{vec}^\top\mathbf{H})^{\otimes 2}\boldsymbol{\mathcal{S}}_{d,4})\right]$$

$$\times \mathsf{D}^{\otimes 2r+4}f(\boldsymbol{x})$$

$$= -(\mathbf{H}^{1/2})^{\otimes 2r}\mathsf{D}^{\otimes 2r}f(\boldsymbol{x}) + \tfrac{1}{4}m_2(K)^2\left[(\mathbf{H}^{1/2})^{\otimes 2r} \otimes (\text{vec}^\top\mathbf{H})^{\otimes 2}\right]\mathsf{D}^{\otimes 2r+4}f(\boldsymbol{x}),$$

Therefore, since $\text{vec}^\top\mathbf{H}^{\otimes -r}(\mathbf{H}^{1/2})^{\otimes 2r} = \text{vec}^\top\mathbf{I}_{d^r}$ and $\mathsf{D}_{\mathbf{H}}(\text{vec}\,\mathbf{H})^{\otimes 2} = (\mathbf{I}_{d^2} \otimes \text{vec}^\top\mathbf{H})(\mathbf{I}_{d^4} + \mathbf{K}_{d^2,d^2})$, we obtain

$$\boldsymbol{\varphi}_{\mathbf{H}} * f(\boldsymbol{x})$$

$$= \mathsf{D}_{\mathbf{H}}\left[\text{vec}^\top\mathbf{H}^{\otimes -r}(\mathsf{D}^{\otimes 2r}\tilde{K})_{\mathbf{H}} * f(\boldsymbol{x})\right]$$

$$\sim \mathsf{D}_{\mathbf{H}}\left\{\tfrac{1}{4}m_2(K)^2\left[\text{vec}^\top\mathbf{I}_{d^r} \otimes (\text{vec}^\top\mathbf{H})^{\otimes 2}\right]\mathsf{D}^{\otimes 2r+4}f(\boldsymbol{x})\right\}$$

$$= \tfrac{1}{4}m_2(K)^2\left(\text{vec}^\top\mathbf{I}_{d^r} \otimes \mathbf{I}_{d^2} \otimes \text{vec}^\top\mathbf{H}\right)[\mathbf{I}_{d^{2r}} \otimes (\mathbf{I}_{d^2} + \mathbf{K}_{d^2,d^2})]\mathsf{D}^{\otimes 2r+4}f(\boldsymbol{x})$$

$$= \tfrac{1}{2}m_2(K)^2\left(\text{vec}^\top\mathbf{I}_{d^r} \otimes \mathbf{I}_{d^2} \otimes \text{vec}^\top\mathbf{H}\right)\mathsf{D}^{\otimes 2r+4}f(\boldsymbol{x}).$$

Using this,

$$\mathbb{E}[\boldsymbol{\varphi}_{\mathbf{H}}(\mathbf{X}_1 - \mathbf{X}_2)] = \int \boldsymbol{\varphi}_{\mathbf{H}} * f(\boldsymbol{x})f(\boldsymbol{x})d\boldsymbol{x}$$

$$\sim \tfrac{1}{2}m_2(K)^2\left(\text{vec}^\top\mathbf{I}_{d^r} \otimes \mathbf{I}_{d^2} \otimes \text{vec}^\top\mathbf{H}\right)\boldsymbol{\psi}_{2r+4}$$

and $\boldsymbol{\Xi}_0 = O(\mathbf{J}_{d^2})\,\text{vec}\,\mathbf{H}\,\text{vec}^\top\mathbf{H}$.

Similarly,

$$\boldsymbol{\Xi}_1 = \int \boldsymbol{\varphi}_{\mathbf{H}} * f(\boldsymbol{x})\boldsymbol{\varphi}_{\mathbf{H}} * f(\boldsymbol{x})^\top f(\boldsymbol{x})d\boldsymbol{x}$$

$$\sim \tfrac{1}{4} m_2(K)^4 \big(\text{vec}^\top \mathbf{I}_{d^r} \otimes \mathbf{I}_{d^2} \otimes \text{vec}^\top \mathbf{H} \big)$$

$$\times \left\{ \int \mathsf{D}^{\otimes 2r+4} f(\boldsymbol{x}) \mathsf{D}^{\otimes 2r+4} f(\boldsymbol{x})^\top f(\boldsymbol{x}) d\boldsymbol{x} \right\} \big(\text{vec}\, \mathbf{I}_{d^r} \otimes \mathbf{I}_{d^2} \otimes \text{vec}\, \mathbf{H} \big)$$

$$= O(\mathbf{J}_{d^2})\, \text{vec}\, \mathbf{H}\, \text{vec}^\top \mathbf{H}.$$

Finally, note that $\boldsymbol{\Xi}_2 = |\mathbf{H}|^{-1/2} \mathbb{E}[(\boldsymbol{\varphi}\boldsymbol{\varphi}^\top)_{\mathbf{H}}(\mathbf{X}_1 - \mathbf{X}_2)]$, where $\boldsymbol{\varphi}(\boldsymbol{x}) = \mathbf{A}(\mathsf{D}^{\otimes 2r}\tilde{K})(\boldsymbol{x}) + \mathbf{B}\boldsymbol{\rho}(\boldsymbol{x})$, which also depends on $\mathbf{H}$ through $\mathbf{A}$ and $\mathbf{B}$. Besides,

$$\mathbb{E}[(\boldsymbol{\varphi}\boldsymbol{\varphi}^\top)_{\mathbf{H}}(\mathbf{X}_1 - \mathbf{X}_2)] = \iint (\boldsymbol{\varphi}\boldsymbol{\varphi}^\top)(\boldsymbol{z}) f(\boldsymbol{y}) f(\boldsymbol{y} + \mathbf{H}^{1/2}\boldsymbol{z}) d\boldsymbol{y} d\boldsymbol{z}$$

$$\sim R(f) \int \boldsymbol{\varphi}(\boldsymbol{z}) \boldsymbol{\varphi}(\boldsymbol{z})^\top d\boldsymbol{z}$$

which, in view of Lemma 2, leads to $\boldsymbol{\Xi}_2 = O(\mathbf{J}_{d^2}|\mathbf{H}|^{-1/2}) \text{vec}\, \mathbf{H}^{\otimes -(r+1)} \text{vec}^\top \mathbf{H}^{\otimes -(r+1)}$.

Putting all these together, since every element of $\mathbf{H}_{\text{MISE},r}$ is $O(n^{-2/(d+2r+4)})$,

$$4n^{-1}(\boldsymbol{\Xi}_1 - \boldsymbol{\Xi}_0) + 2n^{-2}\boldsymbol{\Xi}_2 \sim O(\mathbf{J}_{d^2} n^{-d/(d+2r+4)})\, \text{vec}\, \mathbf{H}_{\text{MISE},r}\, \text{vec}^\top \mathbf{H}_{\text{MISE},r}$$

and therefore $\text{vec}(\hat{\mathbf{H}}_{\text{CV},r} - \mathbf{H}_{\text{MISE},r}) = O(\mathbf{J}_{d^2} n^{-d/(2d+4r+8)})\, \text{vec}\, \mathbf{H}_{\text{MISE},r}$.

A.2. Convergence rate for the PI bandwidth

Henceforth, in addition to (A1)–(A3) the following assumptions on the pilot kernel L and the pilot bandwidth $\mathbf{G}$ are made:

(A4) L is a symmetric d-variate density such that $\int \boldsymbol{x}\boldsymbol{x}^\top L(\boldsymbol{x})\, d\boldsymbol{x} = m_2(L)\mathbf{I}_d$ and all its partial derivatives up to order $2r + 4$ are bounded, continuous and square integrable.

(A5) $\mathbf{G} = \mathbf{G}_n$ is a sequence of bandwidth matrices such that all entries of $n^{-1}|\mathbf{G}|^{-1/2}(\mathbf{G}^{-1})^{\otimes r+2}$ and $\mathbf{G}$ tend to zero as $n \to \infty$.

To make use of Lemma 1 once more, notice that the difference between the MISE and its estimate is

$$\text{PI}_r(\mathbf{H}) - \text{MISE}_r(\mathbf{H}) \sim (-1)^r \tfrac{m_2(K)^2}{4} (\hat{\boldsymbol{\psi}}_{2r+4}(\mathbf{G}) - \boldsymbol{\psi}_{2r+4})^\top \big(\text{vec}\, \mathbf{I}_{d^r} \otimes (\text{vec}\, \mathbf{H})^{\otimes 2} \big)$$

so taking into account $\mathsf{D}_{\mathbf{H}}(\text{vec}\, \mathbf{H})^{\otimes 2} = (\mathbf{I}_{d^2} \otimes \text{vec}^\top \mathbf{H})(\mathbf{I}_{d^4} + \mathbf{K}_{d^2,d^2})$ again, we come to

$$\mathsf{D}_{\mathbf{H}}[\text{PI}_r(\mathbf{H}) - \text{MISE}_r(\mathbf{H})] \sim (-1)^r \tfrac{m_2(K)^2}{2} (\text{vec}^\top \mathbf{I}_{d^r} \otimes \mathbf{I}_{d^2} \otimes \text{vec}^\top \mathbf{H})$$

$$\times (\hat{\boldsymbol{\psi}}_{2r+4}(\mathbf{G}) - \boldsymbol{\psi}_{2r+4}),$$

so that the performance of $\mathbf{H}_{\text{PI},r}$ is determined by the performance of $\hat{\boldsymbol{\psi}}_{2r+4}(\mathbf{G})$ as an estimator of $\boldsymbol{\psi}_{2r+4}$.

From Theorem 2 in Chacón and Duong (2010) the optimal pilot bandwidth $\mathbf{G}$ for the estimator $\hat{\boldsymbol{\psi}}_{2r+4}(\mathbf{G})$ is of order $n^{-2/(d+2r+6)}$, leading to $\mathbb{E}\big[\|\hat{\boldsymbol{\psi}}_{2r+4}(\mathbf{G}) - \boldsymbol{\psi}_{2r+4}\|^2\big] = O(n^{-4/(d+2r+6)})$, and then $\mathsf{D}_{\mathbf{H}}[\text{PI}_r(\mathbf{H}) - \text{MISE}_r(\mathbf{H})] = O_P(n^{-2/(d+2r+6)}\mathbf{J}_{d^2})\, \text{vec}\, \mathbf{H}$. So finally we arrive to $\text{vec}(\hat{\mathbf{H}}_{\text{PI},r} - \mathbf{H}_{\text{MISE},r}) = O_P(n^{-2/(d+2r+6)}\mathbf{J}_{d^2})\, \text{vec}\, \mathbf{H}_{\text{MISE}}$ by applying Lemma 1.

A.3. Convergence rate for the SCV bandwidth

As in Chacón and Duong (2011), it can be shown that the function MISE_r can be replaced for $\mathrm{MISE2}_r$ everywhere in the asymptotic analysis, since the difference between their respective minimizers is of relative order faster than $n^{-1/2}$, which is the fastest attainable rate in bandwidth selection (Hall and Marron, 1991).

So to apply Lemma 1 it is also possible consider $\mathrm{MISE2}_r$ instead of MISE_r, hence we focus on analyzing the difference $\mathrm{SCV}_r(\mathbf{H}) - \mathrm{MISE2}_r(\mathbf{H})$ at $\mathbf{H}$ of the same order as $\mathbf{H}_{\mathrm{MISE},r}$. To begin with, note that using a fourth order Taylor expansion of $\mathsf{D}^{\otimes 2r}\bar{L}(\mathbf{G}^{-1/2}\boldsymbol{x} - \mathbf{G}^{-1/2}\mathbf{H}^{1/2}\boldsymbol{z})$ results in

$$\bar{\Delta}_{\mathbf{H}} * \mathsf{D}^{\otimes 2r}\bar{L}_{\mathbf{G}}(\boldsymbol{x})$$

$$= \int \bar{\Delta}_{\mathbf{H}}(\boldsymbol{z})\mathsf{D}^{\otimes 2r}\bar{L}_{\mathbf{G}}(\boldsymbol{x} - \boldsymbol{z})\,d\boldsymbol{z}$$

$$= |\mathbf{G}|^{-1/2}(\mathbf{G}^{-1/2})^{\otimes 2r}\int \bar{\Delta}(\boldsymbol{z})\mathsf{D}^{\otimes 2r}\bar{L}(\mathbf{G}^{-1/2}\boldsymbol{x} - \mathbf{G}^{-1/2}\mathbf{H}^{1/2}\boldsymbol{z})\,d\boldsymbol{z}$$

$$\sim |\mathbf{G}|^{-1/2}(\mathbf{G}^{-1/2})^{\otimes 2r}$$

$$\times \sum_{p=0}^{4}\frac{(-1)^p}{p!}\int \bar{\Delta}(\boldsymbol{z})[\mathbf{I}_{d^{2r}} \otimes (\boldsymbol{z}^{\top}\mathbf{H}^{1/2}\mathbf{G}^{-1/2})^{\otimes p}]\mathsf{D}^{\otimes 2r+p}\bar{L}(\mathbf{G}^{-1/2}\boldsymbol{x})\,d\boldsymbol{z}$$

$$= \tfrac{1}{4}m_2(K)^2|\mathbf{G}|^{-1/2}(\mathbf{G}^{-1/2})^{\otimes 2r}[\mathbf{I}_{d^{2r}} \otimes (\mathrm{vec}^{\top}\mathbf{H})^{\otimes 2}(\mathbf{G}^{-1/2})^{\otimes 4}]\mathsf{D}^{\otimes 2r+4}\bar{L}(\mathbf{G}^{-1/2}\boldsymbol{x})$$

$$= \tfrac{1}{4}m_2(K)^2|\mathbf{G}|^{-1/2}[\mathbf{I}_{d^{2r}} \otimes (\mathrm{vec}^{\top}\mathbf{H})^{\otimes 2}](\mathbf{G}^{-1/2})^{\otimes(2r+4)}\mathsf{D}^{\otimes 2r+4}\bar{L}(\mathbf{G}^{-1/2}\boldsymbol{x})$$

$$= \tfrac{1}{4}m_2(K)^2[\mathbf{I}_{d^{2r}} \otimes (\mathrm{vec}^{\top}\mathbf{H})^{\otimes 2}]\mathsf{D}^{\otimes 2r+4}\bar{L}_{\mathbf{G}}(\boldsymbol{x}),$$

where we have made use of the fact that $\mu_0(\bar{\Delta}) = \boldsymbol{\mu}_1(\bar{\Delta}) = \boldsymbol{\mu}_2(\bar{\Delta}) = \boldsymbol{\mu}_3(\bar{\Delta}) = 0$ and $\boldsymbol{\mu}_4(\bar{\Delta}) = 6m_2(K)^2\boldsymbol{\mathcal{S}}_{d,4}(\mathrm{vec}\,\mathbf{I}_d)^{\otimes 2}$, and that the entries of $\mathbf{G}^{-1}\mathbf{H}$ tend to zero as a consequence of (A3) and (A5).

This asymptotic approximation is then used to expand the terms in

$$\mathbb{E}[\mathrm{SCV}_r(\mathbf{H}) - \mathrm{MISE2}_r(\mathbf{H})] = (-1)^r\,\mathrm{vec}^{\top}\,\mathbf{I}_{d^r}\left\{n^{-1}\bar{\Delta}_{\mathbf{H}} * \mathsf{D}^{\otimes 2r}\bar{L}_{\mathbf{G}}(0)\right.$$

$$\left. + (1 - n^{-1})\mathbb{E}\big[(\bar{\Delta}_{\mathbf{H}} * \mathsf{D}^{\otimes 2r}\bar{L}_{\mathbf{G}})(\mathbf{X}_1 - \mathbf{X}_2)\big] - \int \bar{\Delta}_{\mathbf{H}} * \mathsf{D}^{\otimes 2r}f(\boldsymbol{x})f(\boldsymbol{x})\,d\boldsymbol{x}\right\}.$$

Precisely, for the first term we have

$$\bar{\Delta}_{\mathbf{H}} * \mathsf{D}^{\otimes 2r}\bar{L}_{\mathbf{G}}(0) \sim \tfrac{1}{4}m_2(K)^2|\mathbf{G}|^{-1/2}[\mathbf{I}_{d^{2r}} \otimes (\mathrm{vec}^{\top}\mathbf{H})^{\otimes 2}](\mathbf{G}^{-1/2})^{\otimes 2r+4}$$
$$\times \mathsf{D}^{\otimes 2r+4}\bar{L}(0),$$

and for the second term

$$\mathbb{E}\big[(\bar{\Delta}_{\mathbf{H}} * \mathsf{D}^{\otimes 2r}\bar{L}_{\mathbf{G}})(\mathbf{X}_1 - \mathbf{X}_2)\big]$$

$$\sim \tfrac{1}{4}m_2(K)^2[\mathbf{I}_{d^{2r}} \otimes (\mathrm{vec}^{\top}\mathbf{H})^{\otimes 2}]\iint \mathsf{D}^{\otimes 2r+4}\bar{L}_{\mathbf{G}}(\boldsymbol{x} - \boldsymbol{y})f(\boldsymbol{x})f(\boldsymbol{y})\,d\boldsymbol{x}d\boldsymbol{y}$$

$$= \tfrac{1}{4}m_2(K)^2[\mathbf{I}_{d^{2r}} \otimes (\mathrm{vec}^{\top}\mathbf{H})^{\otimes 2}]\iint \bar{L}_{\mathbf{G}}(\boldsymbol{x} - \boldsymbol{y})\mathsf{D}^{\otimes 2r+4}f(\boldsymbol{x})f(\boldsymbol{y})\,d\boldsymbol{x}d\boldsymbol{y}$$

$$\sim \tfrac{1}{4}m_2(K)^2[\mathbf{I}_{d^{2r}} \otimes (\mathrm{vec}^\top \mathbf{H})^{\otimes 2}] \iint \bar{L}(\boldsymbol{w}) \sum_{p=0}^{2} \frac{(-1)^p}{p!} [\mathbf{I}_{d^{2r+4}} \otimes (\boldsymbol{w}^\top \mathbf{G}^{1/2})^{\otimes p}]$$

$$\times \, \mathsf{D}^{\otimes 2r+4+p} f(\boldsymbol{y}) f(\boldsymbol{y}) \, d\boldsymbol{w} d\boldsymbol{y}$$

$$= \tfrac{1}{4}m_2(K)^2[\mathbf{I}_{d^{2r}} \otimes (\mathrm{vec}^\top \mathbf{H})^{\otimes 2}] \sum_{p=0}^{2} \frac{(-1)^p}{p!} [\mathbf{I}_{d^{2r+4}} \otimes \{\boldsymbol{\mu}_p(\bar{L})^\top (\mathbf{G}^{1/2})^{\otimes p}\}] \boldsymbol{\psi}_{2r+4+p}$$

$$= \tfrac{1}{4}m_2(K)^2[\mathbf{I}_{d^{2r}} \otimes (\mathrm{vec}^\top \mathbf{H})^{\otimes 2}] \boldsymbol{\psi}_{2r+4} + \tfrac{1}{4}m_2(K)^2 m_2(L)$$

$$\times \, [\mathbf{I}_{d^{2r}} \otimes (\mathrm{vec}^\top \mathbf{H})^{\otimes 2} \otimes \mathrm{vec}^\top \mathbf{G}] \boldsymbol{\psi}_{2r+6},$$

since $\boldsymbol{\mu}_0(\bar{L}) = 1, \boldsymbol{\mu}_1(\bar{L}) = 0$ and $\boldsymbol{\mu}_2(\bar{L}) = 2\boldsymbol{\mu}_2(L) = 2m_2(L)\,\mathrm{vec}\,\mathbf{I}_d$. Finally, noting that $\mathsf{D}^{\otimes 2r}\tilde{K}_{\mathbf{H}} = (\mathbf{H}^{-1/2})^{\otimes 2r}(\mathsf{D}^{\otimes 2r}\tilde{K})_{\mathbf{H}}$ and making use of the previously obtained expansion for $(\mathsf{D}^{\otimes 2r}\tilde{K})_{\mathbf{H}} * f$, the third term is

$$\int \bar{\Delta}_{\mathbf{H}} * \mathsf{D}^{\otimes 2r} f(\boldsymbol{x}) f(\boldsymbol{x}) \, d\boldsymbol{x} = \int \mathsf{D}^{\otimes 2r}\tilde{K}_{\mathbf{H}} * f(\boldsymbol{x}) f(\boldsymbol{x}) \, d\boldsymbol{x} + \boldsymbol{\psi}_{2r}$$

$$\sim \tfrac{1}{4}m_2(K)^2 [\mathbf{I}_{d^{2r}} \otimes (\mathrm{vec}^\top \mathbf{H})^{\otimes 2}] \boldsymbol{\psi}_{2r+4}.$$

Thus,

$$\mathbb{E}[\mathrm{SCV}_r(\mathbf{H}) - \mathrm{MISE2}_r(\mathbf{H})]$$

$$\sim \tfrac{1}{4}m_2(K)^2 n^{-1}|\mathbf{G}|^{-1/2}[\mathrm{vec}^\top \mathbf{I}_{d^r} \otimes (\mathrm{vec}^\top \mathbf{H})^{\otimes 2}](\mathbf{G}^{-1/2})^{\otimes 2r+4}\mathsf{D}^{\otimes 2r+4}\bar{L}(0)$$

$$+ \tfrac{1}{4}m_2(K)^2 m_2(L)[\mathrm{vec}^\top \mathbf{I}_{d^r} \otimes (\mathrm{vec}^\top \mathbf{H})^{\otimes 2} \otimes \mathrm{vec}^\top \mathbf{G}] \boldsymbol{\psi}_{2r+6}$$

Calculations in Section 3 give $\mathbf{G}$ is order $n^{-2/(2r+d+6)}$, as for the plug-in selector, so substituting to this into the derivative of the previous equation yields

$$\mathbb{E}\{\mathsf{D}_{\mathbf{H}}[\mathrm{SCV}_r(\mathbf{H}) - \mathrm{MISE2}_r(\mathbf{H})]\} = O([n^{-1}|\mathbf{G}|^{-1/2}(\mathrm{tr}\,\mathbf{G})^{-r-2} + \mathrm{tr}\,\mathbf{G}]\mathbf{J}_{d^2})\,\mathrm{vec}\,\mathbf{H}$$

$$= O(n^{-2/(2r+d+6)}\mathbf{J}_{d^2})\,\mathrm{vec}\,\mathbf{H}.$$

Lemma 1 shows that $\mathrm{vec}(\hat{\mathbf{H}}_{\mathrm{SCV},r} - \mathbf{H}_{\mathrm{MISE},r})$ is asymptotically equivalent to $\mathsf{D}_{\mathbf{H}}[\mathrm{SCV}_r - \mathrm{MISE2}_r](\mathbf{H}_{\mathrm{MISE},r})$. Since it was stated in Section 3 that $\mathbb{E}\big[\|\,\mathrm{vec}(\hat{\mathbf{H}}_{\mathrm{SCV},r} - \mathbf{H}_{\mathrm{MISE},r})\|^2\big]$ is dominated by its squared bias term, then $\mathrm{vec}(\hat{\mathbf{H}}_{\mathrm{SCV},r} - \mathbf{H}_{\mathrm{MISE},r}) = O_P(n^{-2/(2r+d+6)}\mathbf{J}_{d^2})\,\mathrm{vec}\,\mathbf{H}_{\mathrm{MISE},r}.$

References

AZZALINI, A. AND TORELLI, N. (2007) Clustering via nonparametric density estimation. *Stat. Comput.*, **17**, 71–80. MR2370969

BOWMAN, A.W. (1984) An alternative method of cross-validation for the smoothing of density estimates. *Biometrika*, **71**, 353–360. MR0767163

CAO, R., CUEVAS, A. AND 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. MR2531830

CHACÓN, J.E. AND DUONG, T. (2010) Multivariate plug-in bandwidth selection with unconstrained pilot bandwidth matrices. *Test*, **19**, 375–398. MR2677734

CHACÓN, J.E. AND DUONG, T. (2011) Unconstrained pilot selectors for smoothed cross validation. *Aust. New Zealand J. Statist.*, **53**, 331–351. MR2897374

CHACÓN, J.E. AND DUONG, T. (2012) Efficient recursive algorithms for functionals based on higher order derivatives of the multivariate Gaussian density. In preparation.

CHACÓN, J.E., DUONG, T. AND WAND, M.P. (2011) Asymptotics for general multivariate kernel density derivative estimators. *Statistica Sinica*, **21**, 807–840. MR2829857

CHAUDHURI, P. AND MARRON, J.S. (1999) SiZer for exploration of structure in curves. *J. Amer. Statist. Assoc.*, **94**, 807–823. MR1723347

CHENG, Y. (1995) Mean shift, mode seeking, and clustering. *IEEE T. Pattern Anal.*, **17**, 790–799.

CHOI, E. AND HALL, P. (1999) Data sharpening as a prelude to density estimation. *Biometrika*, **86**, 941–947. MR1741990

COMANICIU, D. (2003) An algorithm for data-driven bandwidth selection. *IEEE T. Pattern Anal.*, **25**, 281–288.

COMANICIU, D. AND MEER, P. (2002) Mean shift: A robust approach toward feature space analysis. *IEEE Trans. Pattern Anal.*, **24**, 603–619.

COMANICIU, D., RAMESH, V. AND MEER, P. (2003) Kernel-based object tracking. *IEEE Trans. Pattern Anal.*, **25**, 564–577.

CUEVAS, A., FEBRERO, M. AND FRAIMAN, R. (2001) Cluster analysis: a further approach based on density estimation. *Comput. Statist. Data Anal.*, **36**, 441–459. MR1855727

DOBROVIDOV, A.V. AND RUD'KO, I.M. (2010) Bandwidth selection in nonparametric estimator of density derivative by smoothed cross-validation method. *Autom. Remote Control*, **71**, 209–224. MR2654498

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., COWLING, A., KOCH, I. AND WAND, M.P. (2008) Feature significance for multivariate kernel density estimation. *Comput. Stat. Data Anal.*, **52**, 4225–4242. MR2432459

DUONG, T. AND HAZELTON, M.L. (2003) Plug-in bandwidth matrices for bivariate kernel density estimation. *J. Nonparametr. Stat.*, **15**, 17–30. MR1958957

DUONG, T. AND HAZELTON, M.L. (2005a) Convergence rates for unconstrained bandwidth matrix selectors in multivariate kernel density estimation. *J. Multivariate Anal.*, **93**, 417–433. MR2162646

DUONG, T. AND HAZELTON, M.L. (2005b) Cross-validation bandwidth matrices for multivariate kernel density estimation. *Scand. J. Statist.*, **32**, 485–506. MR2204631

FORINA M., ARMANINO C., LANTERI S. AND TISCORNIA E. (1983) Classification of olive oils from their fatty acid composition. In: H. Martens and H.J. Russwurm (Eds.), *Food Research and Data Analysis*, Applied Science Publishers, London, pp. 189–214.

FRALEY, C. AND RAFTERY, A.E. (2002) Model-based clustering, discriminant analysis, and density estimation. *J. Amer. Statist. Assoc.*, **97**, 611–631. MR1951635

FRANK, A. AND ASUNCION, A. (2010) *UCI Machine Learning Repository* [http://archive.ics.uci.edu/ml]. University of California, Irvine, School of Information and Computer Science.

Fukunaga, K. (1990) *Introduction to Statistical Pattern Recognition*, 2nd Ed. Academic Press, Boston. MR1075415

FUKUNAGA, K. AND HOSTETLER, L.D. (1975) The estimation of the gradient of a density function, with applications in pattern recognition. *IEEE T. Inform. Theory*, **21**, 32–40. MR0388638

Gel'fand, I.M. and Shilov, G.E. (1966) *Generalized Functions, Volume 1: Properties and Operations*. Academic Press, New York. MR0435831

GENOVESE, C.R, PERONE-PACIFICO, M., VERDINELLY, I. AND WASSERMAN, L. (2009) On the path density of a gradient field. *Ann. Statist.*, **37**, 3236–3271. MR2549559

GODTLIEBSEN, F., MARRON, J.S. AND CHAUDHURI, P. (2002) Significance in scale space for bivariate density estimation. *J. Comput. Graph. Statist.*, **11**, 1–21. MR1937281

GODTLIEBSEN, F., MARRON, J.S. AND CHAUDHURI, P. (2004) Statistical significance of features in digital images. *Image Vision Comput.*, **22**, 1093–1104.

GRUND, B. AND HALL, P. (1995) On the minimisation of the L^p error in mode estimation. *Ann Statist.*, **23**, 2264–2284. MR1389874

HALL, P. (1983) Large sample optimality of least squares cross-validation in density estimation. *Ann. Statist.*, **11**, 1156–1174. MR0720261

HALL, P. AND MARRON, J.S. (1987) Extent to which least-squares cross-validation minimises integrated square error in nonparametric density estimation. *Probab. Theory Rel. Fields*, **74**, 567–581. MR0876256

HALL, P. AND MARRON, J.S. (1991) Lower bounds for bandwidth selection in density estimation. *Probab. Theory Rel. Fields*, **90**, 149–163. MR1128068

HALL, P., MARRON, J.S. AND PARK, B.U. (1992) Smoothed cross validation. *Probab. Theory Rel. Fields*, **92**, 1–20. MR1156447

HALL, P. AND MINOTTE, M.C. (2002) High order data sharpening for density estimation. *J. R. Stat. Soc. Ser. B Stat. Methodol.*, **64**, 141–157. MR1883130

HÄRDLE, W., MARRON, J.S. AND WAND, M.P. (1990) Bandwidth choice for density derivatives. *J. R. Stat. Soc. Ser. B Stat. Methodol.*, **52**, 223–232. MR1049312

HOLMQUIST, B. (1985) The direct product permuting matrices. *Linear Multilinear Algebra*, **17**, 117–141. MR0860181

HOLMQUIST, B. (1996a) The d-variate vector Hermite polynomial of order k. *Linear Algebra Appl.*, **237/238**, 155–190. MR1382672

HOLMQUIST, B. (1996b) Expectations of products of quadratic forms in normal variables. *Stochastic Anal. Appl.*, **14**, 149–164. MR1382579

HOROVÁ, I., KOLÁČEK, J. AND VOPATOVÁ, K. (2013) Full bandwidth matrix selectors for gradient kernel density estimate. *Comput. Statist. Data Anal.*, **57**, 364–376.

HOROVÁ, I. AND VOPATOVÁ, K. (2011) Kernel density gradient estimate. In *Recent Advances in Functional Data Analysis and Related Topics* (ed F. Ferraty), pp. 177–182, Physica Verlag, Heidelberg. MR2815579

HORTON, P. AND NAKAI, K. (1996) A probabilistic classification system for predicting the cellular localization sites of proteins. Proceedings of *Intelligent Systems in Molecular Biology (ISMB-96)*, 109–115.

HUBERT, L. AND ARABIE, P. (1985) Comparing partitions. *J. Classification*, **2**, 193–218.

JONES, M.C. (1991) The roles of ISE and MISE in density estimation. *Statist. Probab. Lett.*, **12**, 51–56. MR1116757

JONES, M.C. (1992) Potential for automatic bandwidth choice in variations on kernel density estimation. *Statist. Probab. Lett.*, **13**, 351–356.

JONES, M.C. (1994) On kernel density derivative estimation. *Comm. Statist. Theory Methods*, **23**, 2133–2139. MR1293175

JONES, M.C., MARRON, J.S. AND PARK, B.U. (1991) A simple root n bandwidth selector. *Ann. Statist.*, **19**, 1919–1932. MR1135156

JONES, M.C., MARRON, J.S., AND SHEATHER, S.J. (1996) A brief survey of bandwidth selection for density estimation. *J. Amer. Statist. Assoc.*, **91**, 401–407. MR1394097

MAGNUS, J.R. AND NEUDECKER, H. (1979) The commutation matrix: some properties and applications. *Ann. Statist.*, **7**, 381–394. MR0520247

KOLLO, T. AND VON ROSEN, D. (2005) *Advanced Multivariate Statistics with Matrices.* Springer, Dordrecht. MR2162145

LI, J., RAY, S. AND LINDSAY, B.G. (2007) A nonparametric statistical approach to clustering via mode identification. *Journal of Machine Learning Research*, **8**, 1687–1723. MR2332445

MAGNUS, J.R. AND NEUDECKER, H. (1999) *Matrix Differential Calculus with Applications in Statistics and Econometrics: Revised Edition.* John Wiley & Sons, Chichester. MR1698873

MATHAI, A.M. AND PROVOST, S.B. (1992) *Quadratic Forms in Random Variables: Theory and Applications.* Marcel Dekker, New York. MR1192786

MILLIGAN, G.W. AND COOPER, M.C. (1986) A study of the comparability of external criteria for hierarchical cluster analysis. *Multivariate Behav. Res.*, **21**, 441–458.

NAUMANN, U. AND WAND, M.P. (2009) Automation in high-content flow cytometry screening. *Cytometry A*, **75A**, 789–797.

PARK, B.U. AND MARRON, J.S. (1990) Comparison of data-driven bandwidth selectors. *J. Amer. Statist. Assoc.*, **85**, 66–72.

PARZEN, E. (1962) On estimation of a probability density function and mode. *Ann. Math. Statist.*, **33**, 1065–1076. MR0143282

Pawlowsky-Glahn, V. and Buccianti, A. (2011) *Compositional Data Analysis: Theory and Applications.* John Wiley & Sons, Chichester. MR2920574

Pratt, J.P., Zeng, Q.T., Ravnic, D., Huss, H., Rawn, J. and Mentzer, S.J. (2009) Hierarchical clustering of monoclonal antibody reactivity patterns in nonhuman species. *Cytometry A*, **75A**, 734–742.

Rinaldo, A. and Wasserman, L. (2010) Generalized density clustering. *Ann. Statist.*, **38**, 2678–2722. MR2722453

Rudemo, M. (1982) Empirical choice of histograms and kernel density estimators. *Scand. J. Statist.*, **9**, 65–78. MR0668683

Schott, J.R. (2003) Kronecker product permutation matrices and their application to moment matrices of the normal distribution. *J. Multivariate Anal.*, **87**, 177–190. MR2007267

Sheather, S.J. and Jones, M.C. (1991) A reliable data-based bandwidth selection method for kernel density estimation. *J. R. Stat. Soc. Ser. B Stat. Methodol.*, **53**, 683–690. MR1125725

Scott, D.W. (1992) *Multivariate Density Estimation: Theory, Practice, and Visualization.* John Wiley & Sons, New York. MR1191168

Simonoff, J.S. (1996) *Smoothing Methods in Statistics.* Springer-Verlag, Berlin. MR1391963

Stone, C.J. (1984) An asymptotically optimal window selection rule for kernel density estimates. *Ann. Statist.*, **12**, 1285–1297. MR0760688

Stuetzle, W. (2003) Estimating the cluster tree of a density by analyzing the minimal spanning tree of a sample. *J. Classification*, **20**, 25–47. MR1983120

Vieu, P. (1996) A note on density mode estimation. *Statist. Probab. Lett.*, **26**, 297–307. MR1393913

Wand, M.P. and Jones, M.C. (1993) Comparison of smoothing parameterizations in bivariate kernel density estimation. *J. Amer. Statist. Assoc.*, **88**, 520–528. MR1224377

Wand, M.P. and Jones, M.C. (1995). *Kernel smoothing*, Chapman & Hall. MR1319818

Wang, X., Qiu, W. and Zamar, R.H. (2007) CLUES: A non-parametric clustering method based on local shrinking. *Comput. Statist. Data Anal.*, **52**, 286–298. MR2409982

Wu, T.-J. (1997) Root n bandwidth selectors for kernel estimation of density derivatives. *J. Amer. Statist. Assoc.*, **92**, 536–547. MR1467847

Zeng, Q.T., Pratt, J.P., Pak, J., Ravnic, D., Huss, H. and Mentzer, S.J. (2007) Feature-guided clustering of multi-dimensional flow cytometry datasets. *Journal of Biomedical Informatics*, **40**, 325–331.