Test (2010) 19: 375–398
DOI 10.1007/s11749-009-0168-4

ORIGINAL PAPER

Multivariate plug-in bandwidth selection with unconstrained pilot bandwidth matrices

J.E. Chacón · T. Duong

Received: 21 April 2009 / Accepted: 3 September 2009 / Published online: 27 October 2009

Abstract Multivariate kernel density estimation is an important technique in exploratory data analysis. Its utility relies on its ease of interpretation, especially by graphical means. The crucial factor which determines the performance of kernel density estimation is the bandwidth matrix selection. Research in finding optimal bandwidth matrices began with restricted parameterizations of the bandwidth matrix which mimic univariate selectors. Progressively these restrictions were relaxed to develop more flexible selectors. In this paper, we propose the first plug-in bandwidth selector with the unconstrained parameterizations of both the final and pilot selectors. Up till now, the development of unconstrained pilot selectors was hindered by the traditional vectorization of higher-order derivatives which lead to increasingly intractable matrix algebraic expressions. We resolve this by introducing an alternative vectorization which gives elegant and tractable expressions. This allows us to quantify the asymptotic and finite sample properties of unconstrained pilot selectors. For target densities with intricate structure (such as multimodality), our unconstrained selectors show the most improvement over the existing plug-in selectors.

Keywords Asymptotic MISE · Multivariate kernel density estimation · Plug-in method · Pre-sphering · Unconstrained bandwidth selectors

Mathematics Subject Classification (2000) 62G07

J.E. Chacón (✉)
Departamento de Matemáticas, Universidad de Extremadura, Badajoz, Spain
e-mail: jechacon@unex.es

T. Duong
Computational Imaging and Modeling Group, Institut Pasteur, Paris, France
e-mail: tduong@pasteur.fr

1 Introduction

Data smoothing is an important tool in statistical exploratory data analysis. Kernel density estimation can be considered a fundamental setting for studying data smoothing. By this we mean that what we learn from kernel density estimation can be applied to other data smoothing contexts, e.g., regression, classification, goodness-of-fit testing, and bump-hunting; see the monographs of Silverman (1986), Wand and Jones (1995), and Simonoff (1996). Apart from this role as a learning ground for other smoothing problems, kernel density estimation is a useful technique in its own right.

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

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

where $\boldsymbol{x} = (x_1, x_2, \ldots, x_d)^T$ and $\mathbf{X}_i = (X_{i1}, X_{i2}, \ldots, X_{id})^T, i = 1, 2, \ldots, n$. Here $K(\boldsymbol{x})$ is the multivariate kernel, which we assume to be a spherically symmetric probability density function having a finite second-order moment, i.e., there exists $m_2(K) \in \mathbb{R}$ such that $m_2(K)\mathbf{I}_d = \int_{\mathbb{R}^d} \boldsymbol{x}\boldsymbol{x}^T K(\boldsymbol{x}) d\boldsymbol{x}$, where $\mathbf{I}_d$ is the $d \times d$ identity matrix. The parameter $\mathbf{H}$ is the bandwidth matrix, which is symmetric and positive definite; and $K_{\mathbf{H}}(\boldsymbol{x}) = |\mathbf{H}|^{-1/2} K(\mathbf{H}^{-1/2}\boldsymbol{x})$.

The crucial factor for kernel density estimation is to select an optimal value for the bandwidth matrix. In common with the majority of researchers in this field, we use the Mean Integrated Squared Error (MISE) as our optimality criterion:

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

For this criterion to make sense, we assume henceforth that both K and f are square integrable. The ideal MISE-optimal bandwidth selector is

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

where $\mathcal{F}$ is the set of all symmetric and positive definite $d \times d$ matrices. This ideal selector however is unattainable since the MISE does not have a tractable closed form in general. A common approach is to seek a tractable approximation, AMISE, which contains the leading terms in the asymptotic expansion of MISE. Details of this expansion are deferred to Sect. 2.2. All that we need at the moment is that the tractability of AMISE allows us to find the AMISE-optimal selector

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

which serves as our surrogate for $\mathbf{H}_{\mathrm{MISE}}$.

Plug-in bandwidth selectors are a major class of bandwidth selectors derived from this AMISE expansion. In the one-dimensional case, these selectors have a fast rate of asymptotic convergence and good finite-sample properties; see Park and Marron (1990) and Sheather and Jones (1991). The key to this performance is the selection of pilot bandwidths defined in terms of functionals of higher derivatives of the density function f. In more recent research, these plug-in selectors have been extended to the multidimensional case; see Wand and Jones (1994) and Duong and Hazelton (2003). In order to simplify the theoretical computations, these authors use univariate pilot bandwidths, which are not optimal for multivariate data. Thus they sacrifice flexibility for ease of computation. In this paper, we develop unconstrained pilot bandwidth matrices, thus opening the possibility for more flexible plug-in selectors.

In Sect. 2, we review the existing framework for plug-in selectors. We note that extending this framework is hampered by that higher-order (greater than two) derivatives of the density function exhibit mathematical difficulties. In Sect. 3, we develop an alternative formulation of these functionals, which leads to a more systematic approach to their definition and estimation. From this, we define unconstrained pilot bandwidth matrices. A multistage plug-in bandwidth selector is proposed in Sect. 4, which incorporates these unconstrained pilot matrices, and its rate of convergence to the AMISE-optimal selector is established. Section 5 contains a numerical simulation study which compares the performance of this new selector with existing plug-in selectors. We end with some concluding remarks.

2 Plug-in bandwidth selectors

2.1 Vectors of higher-order derivatives

Here we establish some notation to be used throughout the paper.

For any two matrices $\mathbf{A}$ and $\mathbf{B}$, we will write its Kronecker product as $\mathbf{A} \otimes \mathbf{B}$. This way, we will denote by

$$\mathbf{A}^{\otimes r} = \bigotimes_{i=1}^{r} \mathbf{A} = \overbrace{\mathbf{A} \otimes \cdots \otimes \mathbf{A}}^{r \text{ matrices}}$$

the rth Kronecker power of $\mathbf{A}$. If $\mathbf{A} \in \mathcal{M}_{m\times n}$ (i.e., $\mathbf{A}$ is a matrix of order $m \times n$), then $\mathbf{A}^{\otimes r} \in \mathcal{M}_{m^r\times n^r}$; therefore, we adopt the convention $\mathbf{A}^{\otimes 1} = \mathbf{A}$ and $\mathbf{A}^{\otimes 0} = 1 \in \mathbb{R}$.

Also, for any function $f : \mathbb{R}^d \to \mathbb{R}$, we denote its gradient vector by

$$\mathsf{D}f = (\partial f/\partial x_1, \dots, \partial f/\partial x_d)^T \in \mathbb{R}^d,$$

where $\mathbf{A}^T$ denotes the transpose of a matrix $\mathbf{A}$. For the higher-order derivatives of f, instead of arranging them into a matrix, as usual, we prefer to arrange them into a vector. Namely, we will write $\mathsf{D}^{\otimes r} f \in \mathbb{R}^{d^r}$ for the vector containing all the partial derivatives of order r. Our formal definition of this vectorized derivative, if we understand $(\partial/\partial x_i)(\partial/\partial x_j) = \partial^2/(\partial x_i \partial x_j)$, is $\mathsf{D}^{\otimes r} f = (\mathsf{D}f)^{\otimes r} = \partial^r f/(\partial \boldsymbol{x})^{\otimes r}$ with $\boldsymbol{x} = (x_1, \dots, x_d)^T$.

Notice that, with this vector notation, we can give a simple expression for the multivariate Taylor polynomial. Precisely, if all the elements in $\mathsf{D}^{\otimes p} f$ are continuous in a neighborhood of $\boldsymbol{x} \in \mathbb{R}^d$, then we can write

$$f(\boldsymbol{x} + \boldsymbol{h}) = \sum_{r=0}^{p} \frac{1}{r!} \big(\boldsymbol{h}^{\otimes r}\big)^T \mathsf{D}^{\otimes r} f(\boldsymbol{x}) + o\big(\|\boldsymbol{h}\|^p\big), \quad \boldsymbol{h} \in \mathbb{R}^d, \tag{2}$$

with $\|\cdot\|$ standing for the Euclidean norm.

In the following, we will also write $\operatorname{vec}\mathbf{A}$ and $\operatorname{vech}\mathbf{A}$ for the vector and vector half operators, respectively, applied to a symmetric matrix $\mathbf{A}$ (see Henderson and Searle 1979). For instance, if $\mathsf{H}f = \partial^2 f/(\partial \boldsymbol{x}\partial \boldsymbol{x}^T) \in \mathcal{M}_{d\times d}$ denotes the Hessian matrix of f, it follows that $\operatorname{vec}\mathsf{H}f = \mathsf{D}^{\otimes 2}f$. Besides, for a function $f\colon \mathbb{R}^d \to \mathbb{R}^p$, we introduce the notation $\mathbf{R}(f) = \int_{\mathbb{R}^d} f(\boldsymbol{x})f(\boldsymbol{x})^T d\boldsymbol{x} \in \mathcal{M}_{p\times p}$, which is a positive definite symmetric matrix such that $\operatorname{vec}\mathbf{R}(f) = \int_{\mathbb{R}^d} f(\boldsymbol{x})^{\otimes 2} d\boldsymbol{x} \in \mathbb{R}^{p^2}$. We will omit the bold font if $R(f) \in \mathbb{R}$ (i.e., when $p = 1$).

2.2 Alternative expressions for AMISE

We can decompose $\mathrm{MISE}(\mathbf{H}) = \mathrm{ISB}(\mathbf{H}) + \mathrm{IV}(\mathbf{H})$, where the integrated square bias (ISB) and integrated variance (IV) terms are given by $\mathrm{ISB}(\mathbf{H}) = \int_{\mathbb{R}^d}\{\mathbb{E}\hat{f}_{n\mathbf{H}}(\boldsymbol{x}) - f(\boldsymbol{x})\}^2 d\boldsymbol{x}$ and $\mathrm{IV}(\mathbf{H}) = \int_{\mathbb{R}^d} \operatorname{Var} \hat{f}_{n\mathbf{H}}(\boldsymbol{x})\, d\boldsymbol{x}$.

Plug-in methods for choosing the bandwidth matrix rely on an asymptotic form of the MISE, known as the AMISE. The usual expression of AMISE contains an approximation of the ISB as a quadratic form of $\operatorname{vech}\mathbf{H}$. Concretely, if all the elements in $\mathsf{D}^{\otimes 2}f$ are bounded, continuous, and square integrable and if $\operatorname{vech}\mathbf{H} \to 0$ and $n^{-1}|\mathbf{H}|^{-1/2} \to 0$ as $n \to \infty$, then Wand (1992) shows that $\mathrm{MISE}(\mathbf{H}) = \mathrm{AMISE}(\mathbf{H}) + o(n^{-1}|\mathbf{H}|^{-1/2} + \|\operatorname{vech}\mathbf{H}\|^2)$ with

$$\mathrm{AMISE}(\mathbf{H}) = n^{-1}|\mathbf{H}|^{-1/2}R(K) + \frac{m_2(K)^2}{4}\big(\operatorname{vech}^T\mathbf{H}\big)\boldsymbol{\Psi}_4(\operatorname{vech}\mathbf{H}), \tag{3}$$

where $\boldsymbol{\Psi}_4$ is the $\frac{1}{2}d(d+1) \times \frac{1}{2}d(d+1)$ matrix given by

$$\boldsymbol{\Psi}_4 = \int_{\mathbb{R}^d} \operatorname{vech}\big\{2\mathsf{H}f(\boldsymbol{x}) - \operatorname{dg}\mathsf{H}f(\boldsymbol{x})\big\}\operatorname{vech}^T\big\{2\mathsf{H}f(\boldsymbol{x}) - \operatorname{dg}\mathsf{H}f(\boldsymbol{x})\big\}\, d\boldsymbol{x}.$$

Here $\operatorname{dg}\mathsf{H}f$ denotes the diagonal matrix formed by replacing all off-diagonal entries of $\mathsf{H}f$ by zeroes.

Notice that $\boldsymbol{\Psi}_4 = \mathbf{D}_d^T\mathbf{R}(\mathsf{D}^{\otimes 2}f)\mathbf{D}_d$ with $\mathbf{D}_d$ standing for the duplication matrix of order d (see Magnus and Neudecker 1980). This way, we can rewrite

$$\mathrm{AMISE}(\mathbf{H}) = n^{-1}|\mathbf{H}|^{-1/2}R(K) + \frac{m_2(K)^2}{4}\big(\operatorname{vec}^T\mathbf{H}\big)\mathbf{R}\big(\mathsf{D}^{\otimes 2}f\big)(\operatorname{vec}\mathbf{H}). \tag{4}$$

Moreover, applying the well-known property that $\operatorname{vec}(\mathbf{ABC}) = (\mathbf{C}^T \otimes \mathbf{A})\operatorname{vec}\mathbf{B}$ to the second summand, we get

$$\mathrm{AMISE}(\mathbf{H}) = n^{-1}|\mathbf{H}|^{-1/2}R(K) + \frac{m_2(K)^2}{4}\big(\operatorname{vec}^T\mathbf{H}\big)^{\otimes 2}\operatorname{vec}\mathbf{R}\big(\mathsf{D}^{\otimes 2}f\big). \tag{5}$$

In this last formulation, the sense of a quadratic form for the ISB is somehow missing, but perhaps this form is the one that bears the strongest resemblance to the univariate case (see Rosenblatt 1956).

Expressions (4) and (5) are defined in terms of $\operatorname{vec}\mathbf{H}$. Since $\operatorname{vech}\mathbf{H}$ is the smallest vector which contains all the unique elements of $\mathbf{H}$, such expressions are not minimal in this sense. But when expressed in terms of $\operatorname{vec}\mathbf{H}$, the matrix analysis is more straightforward, as we will see in the following sections.

Given any of the AMISE approximations, we will change our goal of estimating $\mathbf{H}_{\text{MISE}}$ for estimating its surrogate $\mathbf{H}_{\text{AMISE}}$, which is defined as the minimizer of $\text{AMISE}(\mathbf{H})$ over $\mathcal{F}$. It is easy to show that $\mathbf{H}_{\text{AMISE}}$ is a reasonably good approximation to $\mathbf{H}_{\text{MISE}}$; see Duong and Hazelton (2005a, Remark 5).

Wand (1992) states that, unlike the univariate case, it is not possible to give a closed explicit formula for $\mathbf{H}_{\text{AMISE}}$ using expression (3). The same holds for expressions (4) and (5). Therefore, our strategy for estimating $\mathbf{H}_{\text{AMISE}}$ would be based on estimating the AMISE function in the first place and then choosing the plug-in bandwidth selector $\hat{\mathbf{H}}$ as the minimizer of the estimated criterion. As the only unknown term in the AMISE function is $\mathbf{R}(\mathsf{D}^{\otimes 2}f)$, this raises the new problem of estimating this parameter, which is indeed a matrix of order $d^2 \times d^2$. The different approaches to the estimation of $\mathbf{R}(\mathsf{D}^{\otimes 2}f)$ or $\boldsymbol{\Psi}_4$ give rise to different versions of the plug-in bandwidth matrix selector. We review some of the existing proposals for this problem, and introduce our new, improved one, in the next section.

3 Estimation of integrated density derivative matrices

The elements of $\operatorname{vec}\mathbf{R}(\mathsf{D}^{\otimes 2}f)$ or $\boldsymbol{\Psi}_4$ are integrals of products of density derivatives of f. Hall and Marron (1987), amongst others, have studied element-wise estimators of these functionals. These studies have focused on optimally estimating scalar functionals for univariate density estimation. For multivariate data, these functionals are arranged in a matrix which adds difficulties, so less attention has been directed towards them.

Wand and Jones (1994) applied existing optimal scalar functional estimators element-wise to construct the first estimator of the integrated density derivative matrix $\boldsymbol{\Psi}_4$. With this element-wise estimator, they did not need to reconsider the question of pilot selectors for matrix estimates. Writing the fourth-order integrated density derivatives as $\boldsymbol{\Psi}_4$ allows easy differentiation of the AMISE in (3) with respect to $\operatorname{vech}\mathbf{H}$.

The Duong and Hazelton (2003) selector modified the Wand and Jones (1994) estimator to a matrix-wise one. The former showed this to have better theoretical and numerical properties. The pilot selector developed by Duong and Hazelton (2003) for their matrix-wise estimator is parameterized as a positive scalar multiplied by the identity matrix. This constrained parameterization was chosen to obtain explicit expressions for the pilot selector by reducing it to univariate analysis. The aforementioned two pilot selectors appear to be the only ones currently available for plug-in selectors.

Our proposed selector is also based on a matrix-wise estimator of $\boldsymbol{\Psi}_4$ but with the added flexibility of pilot selectors parameterized as unconstrained matrices. We develop some new matrix analysis results to find explicit expressions for these unconstrained selectors. We go further, as we show in Sect. 3.2, by allowing some redundancy in rearranging $\boldsymbol{\Psi}_4$ into $\operatorname{vec}\mathbf{R}(\mathsf{D}^{\otimes 2}f)$ and by taking derivatives with respect to $\operatorname{vec}\mathbf{H}$. In contrast, the subsequent matrix analysis is simplified substantially so that we are able to construct unconstrained pilot selectors. This was not feasible with the previous two plug-in selectors.

3.1 The existing proposals

3.1.1 The method of Wand and Jones (1994)

Wand and Jones (1994) notice that every element in the matrix $\boldsymbol{\Psi}_4$ can be written as the integrated density derivative functional

$$\psi_{\mathbf{r}} = \int_{\mathbb{R}^d} f^{(\mathbf{r})}(\boldsymbol{x}) f(\boldsymbol{x})\, d\boldsymbol{x}$$

for some multiindex $\mathbf{r} = (r_1, \dots, r_d) \in \mathbb{N}_0^d$ such that $|\mathbf{r}| = \sum_{i=1}^d r_i = 4$, where we are denoting

$$f^{(\mathbf{r})} = \frac{\partial^{|\mathbf{r}|} f}{\partial x_1^{r_1} \cdots \partial x_d^{r_d}}.$$

Therefore, they proposed an element-wise estimate of $\boldsymbol{\Psi}_4$, where every element $\psi_{\mathbf{r}}$ in $\boldsymbol{\Psi}_4$ is estimated by

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

Here the pilot bandwidth $\mathbf{G}$ and the kernel L may differ from $\mathbf{H}$ and K, respectively.

As every $\psi_{\mathbf{r}}$ is a real number, Wand and Jones (1994) measure the error of the estimate $\hat{\psi}_{\mathbf{r}}(\mathbf{G})$ through the mean-squared error function, $\mathrm{MSE}_{\mathbf{r}}(\mathbf{G}) = \mathbb{E}[\{\hat{\psi}_{\mathbf{r}}(\mathbf{G}) - \psi_{\mathbf{r}}\}^2]$. Wand (1992) provides an asymptotic approximation of this MSE function, the AMSE, but shows that explicit minimization of $\mathrm{AMSE}_{\mathbf{r}}(\mathbf{G})$ is possible only if we restrict the pilot bandwidth matrix to be of the form $\mathbf{G} = g^2\mathbf{I}_d$ with $g > 0$. Moreover, Wand and Jones (1994) state that for general d and unconstrained $\mathbf{G}$, it is difficult even to provide succinct expressions for the estimator $\hat{\psi}_{\mathbf{r}}(\mathbf{G})$ itself, so they focus on the bivariate case.

Besides, Wand and Jones (1994) highlight the fact that, if we want to use an unconstrained bandwidth matrix $\mathbf{H}$ in the estimation of f, then the number of functionals $\psi_{\mathbf{r}}$ that we must estimate (therefore, the number of pilot bandwidths g that we must choose) increases drastically with the dimension, and, thus, they also restrict themselves to the case of a diagonal $\mathbf{H}$. However, in some situations, the use of a diagonal $\mathbf{H}$ could cause a considerable loss in efficiency; see Wand and Jones (1993) and Chacón (2009).

3.1.2 The method of Duong and Hazelton (2003)

Duong and Hazelton (2003) propose a plug-in method for selecting an unconstrained bandwidth matrix $\mathbf{H}$ which is similar to the one described above. However, they remark an important drawback of the previous approach: if the elements of the matrix

$\boldsymbol{\Psi}_4$ are estimated separately, using kernel estimators with different pilot bandwidths, then the resulting matrix estimate may not be positive definite, leading to an AMISE estimate not having a finite global minimum, so that no plug-in bandwidth estimate is obtained in those cases.

To overcome this difficulty, they recommend choosing the same pilot bandwidth $\mathbf{G}$ for all the $\hat{\psi}_{\mathbf{r}}(\mathbf{G})$ estimates with $|\mathbf{r}| = 4$. This is equivalent to estimating $\boldsymbol{\Psi}_4 = \mathbf{D}_d^T \mathbf{R}(\mathsf{D}^{\otimes 2} f)\mathbf{D}_d$ with the matrix estimate $\mathbf{D}_d^T \mathbf{R}(\mathsf{D}^{\otimes 2} \hat{f}_{n\mathbf{G}})\mathbf{D}_d$. This way, in addition to the fact that only one pilot bandwidth needs to be chosen, the use of a plug-in matrix estimator guarantees that the resulting estimate is always positive definite.

They define their optimal pilot bandwidth as the one minimizing the sum of the MSEs corresponding to all the $\psi_{\mathbf{r}}$ elements in $\boldsymbol{\Psi}_4$; that is, the asymptotically optimal $\mathbf{G}$ is the one minimizing $\mathrm{SAMSE}(\mathbf{G}) = \sum_{|\mathbf{r}|=4} \mathrm{AMSE}_{\mathbf{r}}(\mathbf{G})$. Nevertheless, in Duong and Hazelton (2003), a procedure for using an unconstrained $\mathbf{G}$ is not provided either, and the restriction $\mathbf{G} = g^2\mathbf{I}_d$ is also imposed on the pilot bandwidth matrix.

3.1.3 Problems with scaling procedures

Both methods, the one proposed by Wand and Jones (1994) and the one by Duong and Hazelton (2003) share the undesirable feature that the pilot bandwidth matrix is restricted to be of the form $\mathbf{G} = g^2\mathbf{I}_d$ for $g > 0$. This bandwidth matrix parameterization should not be used blindly for multivariate unscaled data, as shown by Wand and Jones (1993). Therefore, Duong and Hazelton (2003) propose to pre-scale or pre-sphere the data before applying the plug-in selection method, and then to transform back to obtain the plug-in bandwidth matrix for the original data (see also Duong and Hazelton 2005b).

The goal of these transformations is to get the transformed data to have the same dispersion in all the coordinate directions. However, even if that is the case, the bandwidth matrix with a single smoothing parameter could be a bad idea in some situations. To illustrate this, let us consider two densities that we are going to use later in Sect. 5, namely, the Correlated Normal and the Separated Bimodal (test densities #2 and #7, respectively). In Fig. 1, we depict the scatterplots of a sample of size $n = 5000$ coming from the aforementioned densities, together with the corresponding scatterplots for the sphered data. We see that data coming from the first distribution certainly seem to have a spherically symmetric distribution when the sphering transformation is applied. However, for the second distribution, although we get the variance matrix of the transformed data to be the identity, these data do not seem to follow a spherically symmetric distribution at all, because their bimodality is retained after the transformation. So, there is still much to lose if we use the single-smoothing-parameter parameterization even for the sphered data.

3.2 Unconstrained pilot bandwidth matrices

In this section, we provide a method for estimating the AMISE function which does not impose any constraint on the pilot bandwidth matrix and therefore does not need any pre-transformation of the data for its application.

Motivated by the analysis in Sect. 2.2, here we are going to study the problem of estimating the matrix $\mathbf{R}(\mathsf{D}^{\otimes s} f) \in \mathcal{M}_{d^s \times d^s}$ for a general $s \in \mathbb{N}$. We are going to

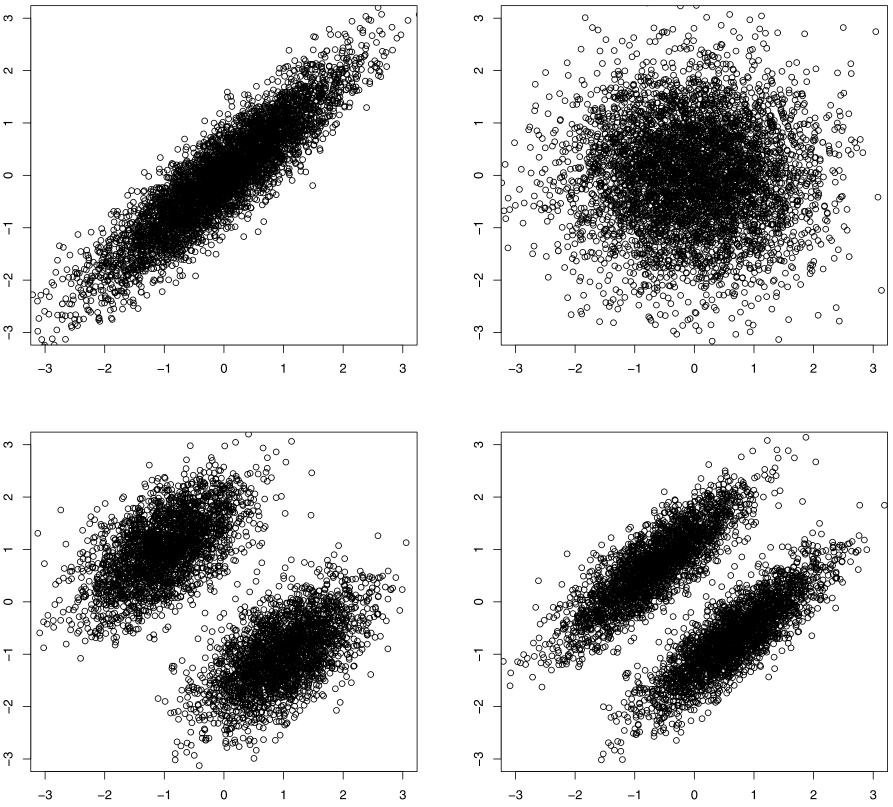

Fig. 1 Scatterplots for samples of size $n = 5000$ from the Correlated Normal and Separated Bimodal distributions (*left column*, *top to bottom*) and the corresponding sphered data (*right column*)

proceed as in Sect. 3.5 of Wand and Jones (1995), because there the same problem is studied in the univariate case. In this sense, our estimator is a multivariate generalization of the one studied in Jones and Sheather (1991) (see also Hall and Marron 1987). However, notice that this generalization to the multivariate case for general d and unconstrained bandwidth matrices is far from trivial, as Wand and Jones (1994) themselves recognize.

If we define

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

then using integration by parts, it is easy to show, under sufficient smoothness assumptions on f, that $\operatorname{vec} \mathbf{R}(\mathsf{D}^{\otimes s} f) = (-1)^s \boldsymbol{\psi}_{2s}$. Therefore, the problem of estimating the matrix $\mathbf{R}(\mathsf{D}^{\otimes s} f)$ is equivalent to that of estimating the vector $\boldsymbol{\psi}_r$ for r even.

The fact that $\boldsymbol{\psi}_r = \mathbb{E}\mathsf{D}^{\otimes r} f(\mathbf{X})$ for any random variable $\mathbf{X}$ with density f motivates the estimator

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

The mean square error of the estimator $\hat{\boldsymbol{\psi}}_r(\mathbf{G})$ is defined as

$$\mathrm{MSE}(\mathbf{G}) = \mathbb{E}\big\{\|\hat{\boldsymbol{\psi}}_r(\mathbf{G}) - \boldsymbol{\psi}_r\|^2\big\}.$$

We prefer to measure the error using this real function, in contrast with some vector-valued (or even matrix-valued) definitions of error, because our aim is to approach the notion of distance between the estimator and the parameter, in order to select the best possible $\mathbf{G}$. However, it can be easily seen that this function can be decomposed as $\mathrm{MSE}(\mathbf{G}) = \mathrm{B}^2(\mathbf{G}) + \mathrm{V}(\mathbf{G})$, where $\mathrm{B}(\mathbf{G}) = \|\mathbb{E}\hat{\boldsymbol{\psi}}_r(\mathbf{G}) - \boldsymbol{\psi}_r\|$ is the norm of the bias vector, and $\mathrm{V}(\mathbf{G}) = \operatorname{tr}\operatorname{Var}\hat{\boldsymbol{\psi}}_r(\mathbf{G})$ is the trace of the variance matrix of the estimator.

The asymptotic approximation to this MSE function is given in the next theorem.

Theorem 1 *Suppose that*:

(L) *L is a symmetric d-variate density such that $\int_{\mathbb{R}^d} zz^T L(z)dz = m_2(L)\mathbf{I}_d$ with each element in $\mathsf{D}^{\otimes j}L$ bounded, continuous, and square integrable for $0 \le j \le r$.*

(D) *All the elements in $\mathsf{D}^{\otimes j} f$ are bounded, continuous, and square integrable for $0 \le j \le r+2$.*

(G) *The bandwidth sequence $\mathbf{G} = \mathbf{G}_n$ is such that* $\operatorname{vec}\mathbf{G} \to 0$.

Then, the MSE function is asymptotically equivalent to $\mathrm{AMSE}(\mathbf{G}) = \mathrm{AB}^2(\mathbf{G}) + \mathrm{AV}(\mathbf{G})$ *with*

$$\mathrm{AB}^2(\mathbf{G}) = \left\| n^{-1}|\mathbf{G}|^{-1/2}\big(\mathbf{G}^{-1/2}\big)^{\otimes r}\mathsf{D}^{\otimes r}L(0) + \frac{m_2(L)}{2}\big(\operatorname{vec}^T\mathbf{G}\otimes\mathbf{I}_{d^r}\big)\boldsymbol{\psi}_{r+2}\right\|^2,$$

$$\mathrm{AV}(\mathbf{G}) = 4n^{-1}\operatorname{tr}\operatorname{Var}\mathsf{D}^{\otimes r} f(\mathbf{X}) + 2n^{-2}\psi_0|\mathbf{G}|^{-1/2}\operatorname{tr}\Big(\big(\mathbf{G}^{-1}\big)^{\otimes r}\mathbf{R}\big(\mathsf{D}^{\otimes r}L\big)\Big).$$

The function $\mathrm{AB}^2(\mathbf{G})$ appearing in the previous display consists of the square norm of the sum of two vectors. Similarly to the situation in the univariate case (see Jones and Sheather 1991), the square norm of the first of these two vectors is easily seen to be of order $O(n^{-2}|\mathbf{G}|^{-1}\operatorname{tr}^r\mathbf{G}^{-1})$, while the term depending on $\mathbf{G}$ in $\mathrm{AV}(\mathbf{G})$ is of smaller order, namely $O(n^{-2}|\mathbf{G}|^{-1/2}\operatorname{tr}^r\mathbf{G}^{-1})$. Therefore, the pilot bandwidth matrix $\mathbf{G}$ can be chosen on the basis of bias alone. Hence, we will consider the (asymptotically) optimal pilot bandwidth to be $\mathbf{G}_{\mathrm{AMSE},r} = \operatorname{argmin}_{\mathbf{G}\in\mathcal{F}}\mathrm{AB}^2(\mathbf{G})$. The next theorem describes the order of this pilot bandwidth and its performance in terms of MSE.

Theorem 2 *Assume hypotheses* (L), (F), *and* (G) *from Theorem 1. Then, the pilot bandwidth matrix $\mathbf{G}_{\mathrm{AMSE},r}$ is of order $n^{-2/(r+d+2)}$. For $d \ge 2$, the MSE obtained when this bandwidth matrix is used in $\hat{\boldsymbol{\psi}}_r(\mathbf{G})$ is of order $n^{-\min\{r+d+2,4\}/(r+d+2)}$.*

It is to be remarked that the results of Theorems 1 and 2 coincide with those given by Jones and Sheather (1991) in the univariate case. The only exception to this generalization is the MSE rate in Theorem 2. Whereas, for $d = 1$, this rate is

$O(n^{-\min\{r+3,5\}/(r+3)})$, we see that, for $d \geq 2$, the rate is slightly slower than expected. This phenomenon is due to the fact that in the univariate case the optimal pilot bandwidth annihilates the dominant term of the squared bias, while, for $d \geq 2$, only square bias minimization is possible in general.

3.3 Exact normal calculations

Of course, the estimator $\hat{\boldsymbol{\psi}}_r(\mathbf{G})$ would be of little use in practice if we could not provide an explicit expression for it. To seek for that goal, we first notice that, for an arbitrary kernel L, it is not hard to show that

$$\mathsf{D}^{\otimes r} L_{\mathbf{G}}(\boldsymbol{x}) = \big(\mathbf{G}^{-1/2}\big)^{\otimes r}\big(\mathsf{D}^{\otimes r} L\big)_{\mathbf{G}}(\boldsymbol{x}) = |\mathbf{G}|^{-1/2}\big(\mathbf{G}^{-1/2}\big)^{\otimes r}\mathsf{D}^{\otimes r} L\big(\mathbf{G}^{-1/2}\boldsymbol{x}\big), \quad (6)$$

so that we can easily compute our estimator if we just know an explicit form for the vector function $\mathsf{D}^{\otimes r} L$.

From now on we are going to use Gaussian kernels; that is, we set $K = L = \phi$, where $\phi(\boldsymbol{x}) = (2\pi)^{-d/2}\exp(-\frac{1}{2}\boldsymbol{x}^T\boldsymbol{x})$ for $\boldsymbol{x} \in \mathbb{R}^d$. Then, we can take advantage of the results in the excellent paper by Holmquist (1996a) to write $\mathsf{D}^{\otimes r}\phi(\boldsymbol{x}) = (-1)^r\phi(\boldsymbol{x})\boldsymbol{\mathcal{H}}_r(\boldsymbol{x})$, where $\boldsymbol{\mathcal{H}}_r(\boldsymbol{x})$ denotes the rth vector Hermite polynomial, given explicitly by

$$\boldsymbol{\mathcal{H}}_r(\boldsymbol{x}) = \sum_{j=0}^{[r/2]}(-1)^j \mathrm{OF}(2j)\binom{r}{2j}\boldsymbol{\mathcal{S}}_{d,r}\big(\boldsymbol{x}^{\otimes(r-2j)} \otimes (\operatorname{vec}\mathbf{I}_d)^{\otimes j}\big).$$

Here, $[a]$ denotes the integer part of a real number a, and, for an even number m, $\mathrm{OF}(m) = (m-1)(m-3)\cdots 3\cdot 1$ denotes its odd factorial, and $\boldsymbol{\mathcal{S}}_{d,r} \in \mathcal{M}_{d^r\times d^r}$ stands for the d-variate symmetrizer matrix of order r, considered for the first time in Holmquist (1985) and studied in recent papers as Schott (2003) (under the name of Kronecker product permutation matrix) and Meijer (2005) (where it is called r-way symmetrization matrix). Essentially, this matrix is defined by the property that, for arbitrary vectors $\boldsymbol{v}_1, \ldots, \boldsymbol{v}_r \in \mathbb{R}^d$,

$$\boldsymbol{\mathcal{S}}_{d,r}\left(\bigotimes_{i=1}^{r}\boldsymbol{v}_i\right) = \frac{1}{r!}\sum_{\sigma}\left(\bigotimes_{i=1}^{r}\boldsymbol{v}_{\sigma(i)}\right),$$

where the sum is extended over all possible permutations σ of r elements. Explicit formulas for this symmetrizer matrix can be found in Holmquist (1996a), Schott (2003), and Meijer (2005).

As it happens for the univariate case, if we are to describe a multistage plug-in method for choosing the bandwidth matrix (see Sect. 4.1 below), at the initial stage, a normal reference estimate of $\boldsymbol{\psi}_r$ for some even r is needed; see Wand and Jones (1995, pp. 72–74). This means that it is necessary to calculate $\boldsymbol{\psi}_r^{\mathrm{NR}}$, the vector $\boldsymbol{\psi}_r$ in the case where $f = \phi_{\boldsymbol{\Sigma}}$. We show in the appendix that such a vector can be written as

$$\boldsymbol{\psi}_r^{\mathrm{NR}} = \frac{(-1)^{r/2} r!}{2^{r+d}(r/2)!\pi^{d/2}}|\boldsymbol{\Sigma}|^{-1/2}\boldsymbol{\mathcal{S}}_{d,r}\big(\operatorname{vec}\boldsymbol{\Sigma}^{-1}\big)^{\otimes(r/2)}. \quad (7)$$

After some algebraic manipulation, this formula in fact coincides with the one provided by Wand and Jones (1995, (3.7)), for the univariate case.

Moreover, for $f = \phi_{\mathbf{\Sigma}}$ and $L = \phi$, it is possible to prove (see Appendix) that the optimal pilot bandwidth matrix admits the explicit expression

$$\mathbf{G}^{\mathrm{NR}}_{\mathrm{AMSE},r} = \left(\frac{2}{r+d}\right)^{2/(r+d+2)} 2\mathbf{\Sigma} n^{-2/(r+d+2)}. \tag{8}$$

This pilot bandwidth coincides with the corresponding one in the normal case for the method of Jones and Sheather (1991) for $d = 1$.

4 A new multistage plug-in bandwidth selector

Here we will use the results of the previous section to propose a new multistage plug-in method for selecting the bandwidth matrix. This new method improves the existing ones in the sense that the choice of the pilot bandwidth is made with no restrictions, over the whole class of symmetric positive definite matrices. In this sense, our method is a generalization of the method by Sheather and Jones (1991) to the multivariate case.

4.1 The method

All the plug-in methods are based on selecting the bandwidth matrix to numerically minimize $\mathrm{PI}(\mathbf{H})$, the AMISE formula with $\mathbf{R}(\mathsf{D}^{\otimes 2} f)$ replaced by a suitable estimator $\widehat{\mathbf{R}(\mathsf{D}^{\otimes 2} f)}$. As $\operatorname{vec}\mathbf{R}(\mathsf{D}^{\otimes 2} f) = \boldsymbol{\psi}_4$, we propose to estimate $\mathbf{R}(\mathsf{D}^{\otimes 2} f)$ using an ℓ-stage estimation method for $\boldsymbol{\psi}_4$, which can be described as follows:

1. Compute $\hat{\boldsymbol{\psi}}^{\mathrm{NR}}_{4+2\ell}$, the value of $\boldsymbol{\psi}^{\mathrm{NR}}_{4+2\ell}$ with $\mathbf{\Sigma}$ replaced by $\mathbf{S}$, the covariance matrix of the data. Plug this estimate into the formula of $\mathrm{AB}^2(\mathbf{G})$ corresponding to $r = 2 + 2\ell$ and numerically minimize to obtain $\hat{\mathbf{G}}_{2+2\ell}$, an estimate of $\mathbf{G}_{\mathrm{AMSE},2+2\ell}$.
2. For $j = 2 + 2\ell, 2\ell, \ldots, 6$:
   (a) Use $\hat{\mathbf{G}}_j$ to compute $\hat{\boldsymbol{\psi}}_j(\hat{\mathbf{G}}_j)$.
   (b) Plug $\hat{\boldsymbol{\psi}}_j(\hat{\mathbf{G}}_j)$ in the formula of $\mathrm{AB}^2(\mathbf{G})$ corresponding to $r = j - 2$ and numerically minimize to obtain $\hat{\mathbf{G}}_{j-2}$, an estimate of $\mathbf{G}_{\mathrm{AMSE},j-2}$.
3. Employ $\hat{\mathbf{G}}_4$ to compute $\hat{\boldsymbol{\psi}}_{4,\ell} = \hat{\boldsymbol{\psi}}_4(\hat{\mathbf{G}}_4)$.

Finally, using the Gaussian kernel, the proposed ℓ-stage plug-in bandwidth selector is

$$\hat{\mathbf{H}}_{\mathrm{PI},\ell} = \operatorname*{argmin}_{\mathbf{H}\in\mathcal{F}} \left\{ n^{-1}|\mathbf{H}|^{-1/2}(4\pi)^{-d/2} + \frac{1}{4}\left(\operatorname{vec}^T \mathbf{H}\right)^{\otimes 2} \hat{\boldsymbol{\psi}}_{4,\ell} \right\}.$$

Based on the recommendations made for the univariate case, we will mainly consider here the two-stage plug-in selector, $\hat{\mathbf{H}}_{\mathrm{PI}} = \hat{\mathbf{H}}_{\mathrm{PI},2}$.

4.2 Asymptotics

Next, we provide the relative rate of convergence of our selector $\hat{\mathbf{H}}_{\mathrm{PI}}$ to the asymptotically optimal $\mathbf{H}_{\mathrm{AMISE}}$. As in Duong and Hazelton (2005a), this rate may be defined to be $n^{-\alpha}$ if $\operatorname{vec}(\hat{\mathbf{H}}_{\mathrm{PI}} - \mathbf{H}_{\mathrm{AMISE}}) = O_P(\mathbf{J}_d n^{-\alpha}) \operatorname{vec} \mathbf{H}_{\mathrm{AMISE}}$, where $\mathbf{J}_d$ denotes the $d \times d$ matrix of ones, and the O_P notation is meant to be applied element-wise.

Theorem 3 *Assume hypotheses* (L) *and* (F) *from Theorem* 1 *for* $r = 4$. *Then*, *for* $d \geq 2$, *the relative rate of convergence of* $\hat{\mathbf{H}}_{\mathrm{PI}}$ *to* $\mathbf{H}_{\mathrm{AMISE}}$ *is* $n^{-2/(d+6)}$.

The relative rate obtained in the previous result is the same as for the SAMSE plug-in selector of Duong and Hazelton (2003), which is slightly slower than the rate $n^{-4/(d+12)}$ attained by the full plug-in selector of Wand and Jones (1994). In this sense, we think that this slower rate reflects the price to be paid for being sure that the estimator of $\mathbf{R}(\mathsf{D}^{\otimes 2} f)$ is positive definite.

In fact, the reason why the rate of the element-wise estimator of $\mathbf{R}(\mathsf{D}^{\otimes 2} f)$ cannot be recovered is that, as stated in Sect. 3.2, using a single pilot bandwidth $\mathbf{G}$, for $d \geq 2$, only square bias minimization is possible in general, in contrast to the univariate case, where the optimal choice of the pilot bandwidth leads to annihilation of the dominant part of the bias, and so to better rates.

5 Simulations

In this section, we undertake a numerical simulation study to compare the finite sample performance of the following selectors:

- The Wand and Jones (1994) plug-in selector with individual pilot selectors parameterized by $\mathbf{G} = g^2 \mathbf{I}_d$, labeled WJ.
- The Duong and Hazelton (2003) plug-in selector with a single selector parameterized by $\mathbf{G} = g^2 \mathbf{I}_d$, labeled DH.
- Our proposed plug-in selector with unconstrained pilot selectors, labeled CD.

We consider samples of size $n = 100$ and $n = 1000$ for 500 simulation runs. For each simulation, we compute the Integrated Squared Error (ISE) between the resulting kernel density estimate and the target density. All these selectors are implemented in the `R` library `ks` (Duong 2009).

5.1 Bivariate study

The six bivariate target densities are some of those appearing in Chacón (2009) and cover a wide range of density shapes (we keep here their names and numbers). Their contour plots are depicted in Fig. 2. Target density #1 is a single normal density, and so it can be considered a base case. Densities #6, #7, #8, #11, and #12 are multimodal with varying degrees of intricate structure.

In Figs. 3 and 4 we show the box-plots of the distributions of the ISEs corresponding to each method for each target density and $n = 100$ and $n = 1000$, respectively,

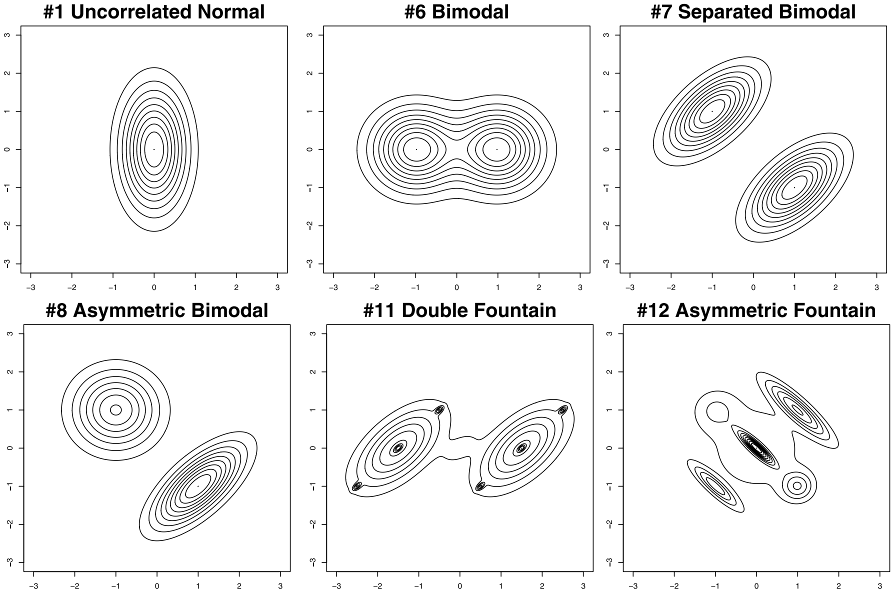

Fig. 2 Contour plots for the 6 target densities

based on the 500 simulation runs. As expected, the differences between the three plug-in methods become clearer for the larger sample size. Two clear conclusions can be drawn:

1. For densities #1 and #6 our selector performs as well as the other two. In fact, all the three selectors have an entirely similar behavior. This is surprising because, for these densities, the use of pretransformations is highly advisable, leading to a simpler method, so a priori we expected the new method to be outperformed by the other two. Nevertheless, although our proposal is more general (because of the use of an unconstrained pilot), it does not lose power against the other plug-in methods even if we have a situation where a pilot bandwidth matrix with a single smoothing parameter is appropriate.
2. However, if the density is such that the single-parameter parameterization of the pilot bandwidth matrix is not suitable for the sphered data, then the plug-in selector with unconstrained pilot bandwidth clearly outperforms the other two methods. This occurs for densities #7, #8, #11, and #12.

5.2 Multivariate study

In addition to the bivariate target densities, we want to test the new plug-in method for densities in higher dimensions. To this end, we will check the performance of the selectors DH and CD at the time of estimating a density presenting features similar to those of density #7 in the previous section. We do not include here the method by Wand and Jones (1994) because the really huge number of density functional estimations needed for its computation in higher dimensions makes this method not very manageable in practice.

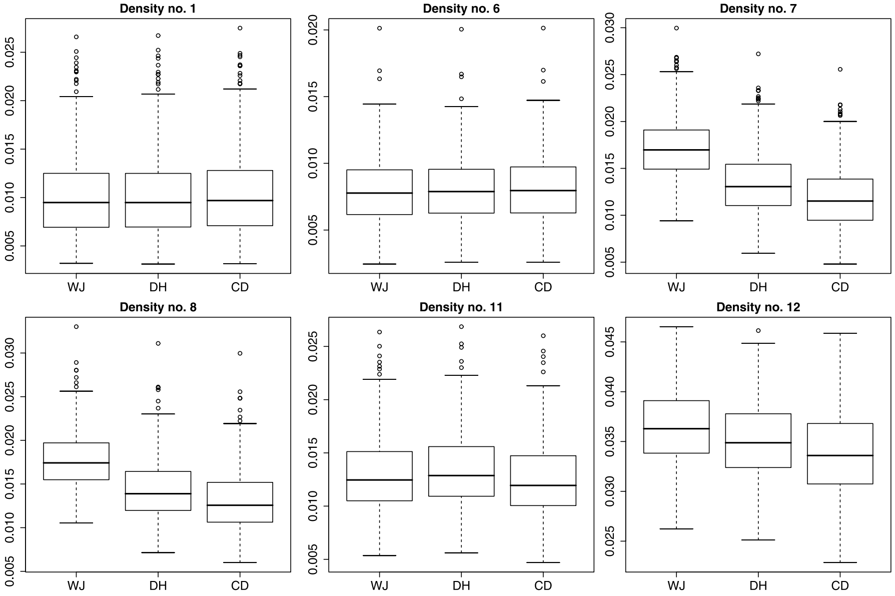

Fig. 3 Box-plots for the ISEs of the plug-in methods WJ, DH, and CD (*from left to right*) and $n = 100$

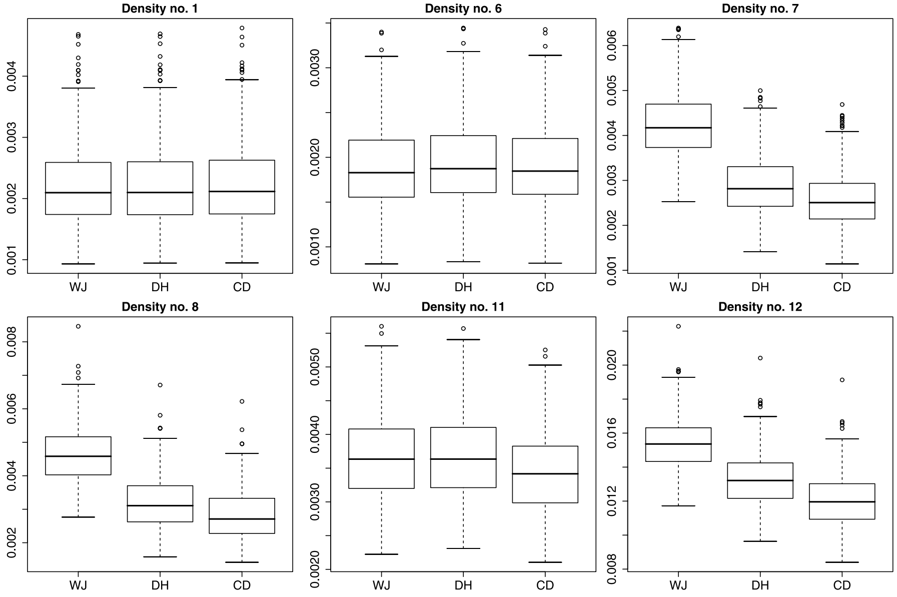

Fig. 4 Box-plots for the ISEs of the plug-in methods WJ, DH, and CD (*from left to right*) and $n = 1000$

Precisely, we want our target density to have every bivariate projection consisting of a separated bimodal density. For instance, such a density may be constructed as follows: for every $i = 2, 3, \dots, d$, consider $\mathbf{R}_i$, the 45-degree rotation matrix in the plane of $\mathbb{R}^d$ defined by the coordinates x_1 and x_i. Multiply these matrices to get

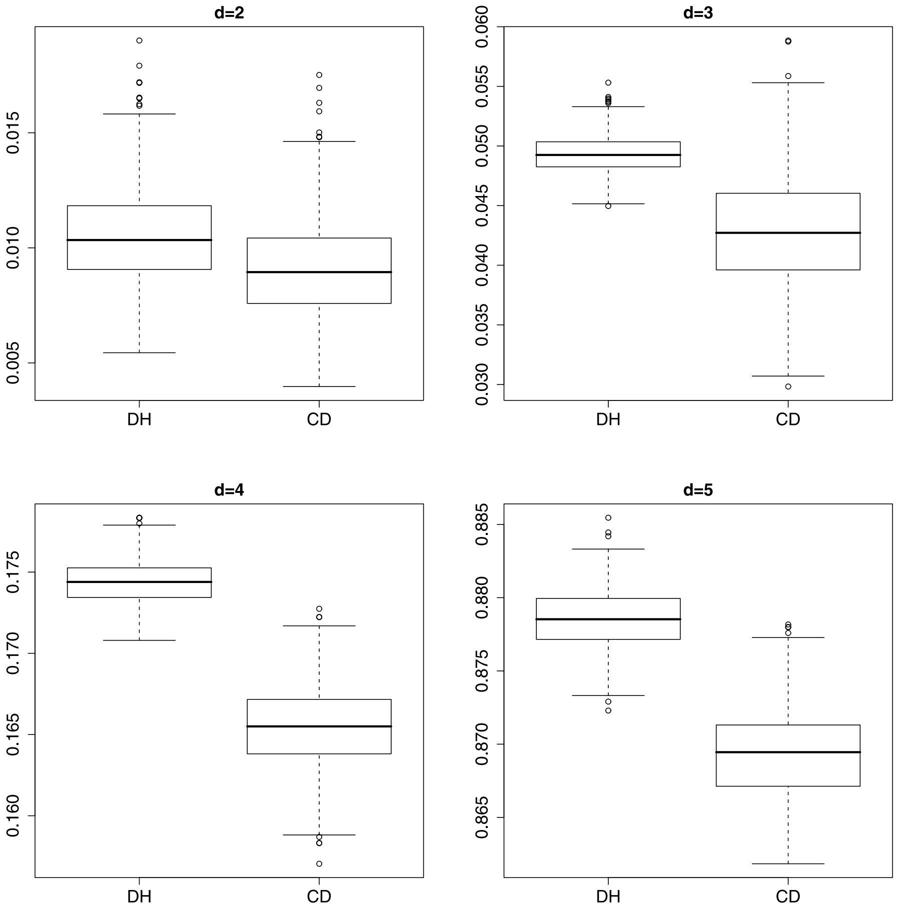

Fig. 5 Box-plots for the ISEs of the plug-in methods DH and CD for $d = 2, 3, 4, 5$

$\mathbf{R} = \mathbf{R}_d\mathbf{R}_{d-1}\cdots\mathbf{R}_2$. Then, the multivariate density that we will aim to estimate is an equal mixture of two d-variate normal densities having means $\boldsymbol{\mu}_1 = \mathbf{R}(d, 0, \ldots, 0)^T$ and $\boldsymbol{\mu}_2 = -\boldsymbol{\mu}_1$ and common variance matrix $\boldsymbol{\Sigma}_1 = \boldsymbol{\Sigma}_2 = \mathbf{R}\boldsymbol{\Delta}\mathbf{R}^T$, where $\boldsymbol{\Delta} \in \mathcal{M}_{d\times d}$ is the diagonal matrix with diagonal given by $(4^{-(d-1)}, 4^{-(d-2)}, \ldots, 4^{-1}, 1)$.

We show in Fig. 5 the ISE box-plots for the estimation of such a density using the two methods with the dimension ranging from $d = 2$ to $d = 5$. In all cases the sample size is $n = 100$. Although we must observe that, for $d \geq 3$, the new method is indeed more variable than the DH method, the box-plots clearly suggest that even so it still has a better performance, at least in terms of ISE.

6 Concluding remarks

By generalizing the pilot bandwidth selection stages for a plug-in selector, we have developed the first multivariate bandwidth selector which is unconstrained at all levels of bandwidth selection. The mathematical framework for these unconstrained selectors involves a subtle yet important re-definition of the usual integrated density derivatives. These redefinitions allowed us to determine the asymptotic optimality properties of these unconstrained selectors.

For their finite sample behavior, we conducted simulation experiments on a range of bivariate densities. The unconstrained pilot selector performs at least as well as the constrained pilot selectors for those densities where the constrained parameterization is appropriate. For those densities whose structure is not appropriate for constrained selectors, the unconstrained selector performs markedly better. This latter conclusion is also verified in higher-dimensional simulations, where the advantage of the new unconstrained selector becomes even more evident. The trade-off for this increased flexibility is the increased computational load. For bivariate data, this is a small increase, though as the dimension increases, due to large symmetrizer matrices required to compute matrix derivatives, the load becomes more onerous.

We concentrate on plug-in selectors since they are a widely used class of bandwidth selectors whose performance results from the optimal tuning of pilot selectors. Unbiased cross validation selectors (Rudemo 1982; Bowman 1984) and biased cross validation selectors (Scott and Terrell 1987; Sain et al. 1994) do not use pilot selectors and consequently typically exhibit lower levels of performance compared to plug-in selectors. Smoothed cross validation (Hall et al. 1992; Duong and Hazelton 2003) rectifies this deficit in performance. In common with plug-in selectors, multivariate smoothed cross validation selectors use pilot selectors to tune the final bandwidth selectors. Future research would be to develop analogous unconstrained pilot selectors for smoothed cross validation and a comprehensive comparison of the different cross validation selectors with our unconstrained plug-in selectors.

Looking further afield, the asymptotic analyses of the unconstrained pilot selectors can be extended to derive optimal bandwidth selectors for kernel estimators of higher-order density derivatives. There is interest in especially the first (gradient) and second (Hessian) derivatives, since they characterize important information about the density function which is not immediately available from only the density itself. These characterizations are important in applications such as bump hunting and feature significance.

Acknowledgements Part of this work was developed while the first author was visiting the second one at the Institut Pasteur in Paris. He is most grateful to the Institut Pasteur for funding his visit. Also, J.E. Chacón has been partially supported by Spanish Ministerio de Ciencia y Tecnología project MTM2006-06172, and T. Duong was funded by Institut Pasteur through a "Programme Transversal de Recherches" grant (PTR No. 218).

Appendix: Proofs

Proof of Theorem 1

Using the bias-variance decomposition of the MSE, we just need to provide an asymptotic expansion for each of these two terms. This is done in the next two lemmas.

Lemma 1 (Bias) *Assume hypotheses* (L), (F), *and* (G) *from Theorem* 1. *Then,*

$$\mathsf{B}(\mathbf{G}) = \Big\| n^{-1}|\mathbf{G}|^{-1/2}\big(\mathbf{G}^{-1/2}\big)^{\otimes r}\mathsf{D}^{\otimes r}L(0) + \frac{m_2(L)}{2}\big(\operatorname{vec}^T\mathbf{G}\otimes\mathbf{I}_{d^r}\big)\boldsymbol{\psi}_{r+2} \\ + \big[O\big(n^{-1}\big) + o(\operatorname{tr}\mathbf{G})\big]\mathbf{1}_{d^r}\Big\|,$$

where $\mathbf{1}_{d^r}$ *denotes the vector in* $\mathbb{R}^{d^r}$ *with all elements equal to* 1.

Proof We can write

$$\mathbb{E}\hat{\boldsymbol{\psi}}_r(\mathbf{G}) = n^{-1}\mathsf{D}^{\otimes r}L_{\mathbf{G}}(0) + \big(1-n^{-1}\big)\mathbb{E}\big[\big(\mathsf{D}^{\otimes r}L_{\mathbf{G}}\big)(\mathbf{X}_1-\mathbf{X}_2)\big].$$

The first term in the right-hand side of the previous formula can be expanded using (6). For the expected value, using hypotheses (L) and (F), and standard techniques, it is not difficult to see that

$$\begin{aligned}
&\mathbb{E}\big[\big(\mathsf{D}^{\otimes r}L_{\mathbf{G}}\big)(\mathbf{X}_1-\mathbf{X}_2)\big]\\
&\quad= \int_{\mathbb{R}^{2d}}\big(\mathsf{D}^{\otimes r}L_{\mathbf{G}}\big)(\boldsymbol{x}-\boldsymbol{y})f(\boldsymbol{x})f(\boldsymbol{y})\,d\boldsymbol{x}\,d\boldsymbol{y}\\
&\quad= \int_{\mathbb{R}^{2d}}L_{\mathbf{G}}(\boldsymbol{x}-\boldsymbol{y})f(\boldsymbol{x})\mathsf{D}^{\otimes r}f(\boldsymbol{y})\,d\boldsymbol{x}\,d\boldsymbol{y}\\
&\quad= \int_{\mathbb{R}^{2d}}L(\boldsymbol{z})f\big(\boldsymbol{y}+\mathbf{G}^{1/2}\boldsymbol{z}\big)\mathsf{D}^{\otimes r}f(\boldsymbol{y})\,d\boldsymbol{y}\,d\boldsymbol{z}\\
&\quad= \int_{\mathbb{R}^{d}}\mathsf{D}^{\otimes r}f(\boldsymbol{y})f(\boldsymbol{y})\,d\boldsymbol{y} + \int_{\mathbb{R}^{2d}}L(\boldsymbol{z})\boldsymbol{z}^T\mathbf{G}^{1/2}\mathsf{D}f(\boldsymbol{y})\mathsf{D}^{\otimes r}f(\boldsymbol{y})\,d\boldsymbol{y}\,d\boldsymbol{z}\\
&\qquad+ \frac{1}{2}\int_{\mathbb{R}^{2d}}L(\boldsymbol{z})\big(\boldsymbol{z}^T\mathbf{G}^{1/2}\big)^{\otimes 2}\mathsf{D}^{\otimes 2}f(\boldsymbol{y})\mathsf{D}^{\otimes r}f(\boldsymbol{y})\,d\boldsymbol{y}\,d\boldsymbol{z} + o(\operatorname{tr}\mathbf{G})\mathbf{1}_{d^r}\\
&\quad= \boldsymbol{\psi}_r + \frac{m_2(L)}{2}\int_{\mathbb{R}^{d}}\big(\operatorname{vec}^T\mathbf{I}_d\big)\big(\mathbf{G}^{1/2}\big)^{\otimes 2}\mathsf{D}^{\otimes 2}f(\boldsymbol{y})\mathsf{D}^{\otimes r}f(\boldsymbol{y})\,d\boldsymbol{y} + o(\operatorname{tr}\mathbf{G})\mathbf{1}_{d^r}\\
&\quad= \boldsymbol{\psi}_r + \frac{m_2(L)}{2}\bigg\{\int_{\mathbb{R}^{d}}\mathsf{D}^{\otimes r}f(\boldsymbol{y})\mathsf{D}^{\otimes 2}f(\boldsymbol{y})^T\,d\boldsymbol{y}\bigg\}\operatorname{vec}\mathbf{G} + o(\operatorname{tr}\mathbf{G})\mathbf{1}_{d^r}\\
&\quad= \boldsymbol{\psi}_r + \frac{m_2(L)}{2}\big(\operatorname{vec}^T\mathbf{G}\otimes\mathbf{I}_{d^r}\big)\operatorname{vec}\int_{\mathbb{R}^{d}}\mathsf{D}^{\otimes r}f(\boldsymbol{y})\mathsf{D}^{\otimes 2}f(\boldsymbol{y})^T\,d\boldsymbol{y} + o(\operatorname{tr}\mathbf{G})\mathbf{1}_{d^r}\\
&\quad= \boldsymbol{\psi}_r + \frac{m_2(L)}{2}\big(\operatorname{vec}^T\mathbf{G}\otimes\mathbf{I}_{d^r}\big)\operatorname{vec}\boldsymbol{\psi}_{r+2} + o(\operatorname{tr}\mathbf{G})\mathbf{1}_{d^r},
\end{aligned}$$

where, for the last two equalities, we have used that $\mathbf{A}\boldsymbol{v} = \operatorname{vec}\mathbf{A}\boldsymbol{v} = (\boldsymbol{v}^T\otimes\mathbf{I}_p)\operatorname{vec}\mathbf{A}$ for a matrix $\mathbf{A}\in\mathcal{M}_{p\times q}$ and a vector $\boldsymbol{v}\in\mathbb{R}^q$, and the fact that the usual properties of the Kronecker product and an element-wise application of the integral formula in

Wand and Jones (1995, p. 111), lead to

$$\begin{aligned}\operatorname{vec}\int_{\mathbb{R}^d}\mathsf{D}^{\otimes r}f(\boldsymbol{y})\mathsf{D}^{\otimes 2}f(\boldsymbol{y})^T\,d\boldsymbol{y}&=\int_{\mathbb{R}^d}\mathsf{D}^{\otimes 2}f(\boldsymbol{y})\otimes\mathsf{D}^{\otimes r}f(\boldsymbol{y})\,d\boldsymbol{y}\\&=\int_{\mathbb{R}^d}\mathsf{D}^{\otimes(r+2)}f(\boldsymbol{y})f(\boldsymbol{y})\,d\boldsymbol{y}.\end{aligned}$$

This yields the proof. □

Lemma 2 (Variance) *Assume hypotheses* (L), (F), *and* (G) *from Theorem* 1. *Then*,

$$\begin{aligned}\mathsf{V}(\mathbf{G})={}&4n^{-1}\left\{\int_{\mathbb{R}^d}\mathsf{D}^{\otimes r}f(\boldsymbol{x})^T\mathsf{D}^{\otimes r}f(\boldsymbol{x})f(\boldsymbol{x})\,d\boldsymbol{x}-\|\boldsymbol{\psi}_r\|^2\right\}\\&+2n^{-2}\psi_0|\mathbf{G}|^{-1/2}\operatorname{tr}\big((\mathbf{G}^{-1})^{\otimes r}\mathbf{R}(\mathsf{D}^{\otimes r}L)\big)\\&+o\big(n^{-1}\big)+o\big(n^{-2}|\mathbf{G}|^{-1/2}\operatorname{tr}^{-r}\mathbf{G}\big).\end{aligned}$$

Proof The dominant part of the trace of the covariance matrix of $\hat{\boldsymbol{\psi}}_r(\mathbf{G})$ is given by $4n^{-1}(\xi_1-\xi_0)+2n^{-2}\xi_2$, where

$$\begin{aligned}\xi_1&=\operatorname{tr}\mathbb{E}\big[\mathsf{D}^{\otimes r}L_{\mathbf{G}}(\mathbf{X}_1-\mathbf{X}_2)\mathsf{D}^{\otimes r}L_{\mathbf{G}}(\mathbf{X}_1-\mathbf{X}_3)^T\big],\\\xi_2&=\operatorname{tr}\mathbb{E}\big[\mathsf{D}^{\otimes r}L_{\mathbf{G}}(\mathbf{X}_1-\mathbf{X}_2)\mathsf{D}^{\otimes r}L_{\mathbf{G}}(\mathbf{X}_1-\mathbf{X}_2)^T\big],\\\xi_0&=\operatorname{tr}\mathbb{E}\big[\mathsf{D}^{\otimes r}L_{\mathbf{G}}(\mathbf{X}_1-\mathbf{X}_2)\big]\mathbb{E}\big[\mathsf{D}^{\otimes r}L_{\mathbf{G}}(\mathbf{X}_1-\mathbf{X}_2)\big]^T.\end{aligned}$$

The asymptotic expansion for ξ_0 follows immediately from the calculations in the proof of the previous lemma. We have

$$\xi_0=\operatorname{tr}\big(\boldsymbol{\psi}_r\boldsymbol{\psi}_r^T\big)+O(\operatorname{tr}\mathbf{G})=\|\boldsymbol{\psi}_r\|^2+O(\operatorname{tr}\mathbf{G}).$$

For ξ_1, arguing as in Wand and Jones (1995, p. 69), we have

$$\begin{aligned}\xi_1&=\operatorname{tr}\int_{\mathbb{R}^{3d}}\mathsf{D}^{\otimes r}L_{\mathbf{G}}(\boldsymbol{x}-\boldsymbol{y})\mathsf{D}^{\otimes r}L_{\mathbf{G}}(\boldsymbol{x}-\boldsymbol{z})^T f(\boldsymbol{x})f(\boldsymbol{y})f(\boldsymbol{z})\,d\boldsymbol{x}\,d\boldsymbol{y}\,d\boldsymbol{z}\\&=\operatorname{tr}\int_{\mathbb{R}^{3d}}L(\boldsymbol{u})L(\boldsymbol{v})f(\boldsymbol{x})\mathsf{D}^{\otimes r}f\big(\boldsymbol{x}-\mathbf{G}^{1/2}\boldsymbol{u}\big)\mathsf{D}^{\otimes r}f\big(\boldsymbol{x}-\mathbf{G}^{1/2}\boldsymbol{v}\big)^T\,d\boldsymbol{u}\,d\boldsymbol{v}\,d\boldsymbol{x}\\&=\operatorname{tr}\int_{\mathbb{R}^d}\mathsf{D}^{\otimes r}f(\boldsymbol{x})\mathsf{D}^{\otimes r}f(\boldsymbol{x})^T f(\boldsymbol{x})\,d\boldsymbol{x}+O(\operatorname{tr}\mathbf{G}).\end{aligned}$$

Therefore, $\xi_1=\int_{\mathbb{R}^d}\mathsf{D}^{\otimes r}f(\boldsymbol{x})^T\mathsf{D}^{\otimes r}f(\boldsymbol{x})f(\boldsymbol{x})\,d\boldsymbol{x}+O(\operatorname{tr}\mathbf{G})$. Finally, for ξ_2, we have

$$\begin{aligned}\xi_2&=\operatorname{tr}\int_{\mathbb{R}^{2d}}\mathsf{D}^{\otimes r}L_{\mathbf{G}}(\boldsymbol{x}-\boldsymbol{y})\mathsf{D}^{\otimes r}L_{\mathbf{G}}(\boldsymbol{x}-\boldsymbol{y})^T f(\boldsymbol{x})f(\boldsymbol{y})\,d\boldsymbol{x}\,d\boldsymbol{y}\\&=\operatorname{tr}\bigg[\big(\mathbf{G}^{-1/2}\big)^{\otimes r}\bigg\{\int_{\mathbb{R}^{2d}}\big(\mathsf{D}^{\otimes r}L\big)_{\mathbf{G}}(\boldsymbol{x}-\boldsymbol{y})\big(\mathsf{D}^{\otimes r}L\big)_{\mathbf{G}}(\boldsymbol{x}-\boldsymbol{y})^T f(\boldsymbol{x})f(\boldsymbol{y})\,d\boldsymbol{x}\,d\boldsymbol{y}\bigg\}\end{aligned}$$

$$\times \big(\mathbf{G}^{-1/2}\big)^{\otimes r}\Big]$$

$$= |\mathbf{G}|^{-1/2}\operatorname{tr}\Big[\big(\mathbf{G}^{-1}\big)^{\otimes r}\int_{\mathbb{R}^{2d}} \mathsf{D}^{\otimes r}L(\boldsymbol{z})\mathsf{D}^{\otimes r}L(\boldsymbol{z})^T f\big(\boldsymbol{y}+\mathbf{G}^{1/2}\boldsymbol{z}\big) f(\boldsymbol{y})\,d\boldsymbol{y}\,d\boldsymbol{z}\Big]$$

$$= \psi_0|\mathbf{G}|^{-1/2}\operatorname{tr}\big(\big(\mathbf{G}^{-1}\big)^{\otimes r}\mathbf{R}\big(\mathsf{D}^{\otimes r}L\big)\big) + O\big(|\mathbf{G}|^{-1/2}\operatorname{tr}\big[\big(\mathbf{G}^{-1}\big)^{\otimes r}\mathbf{G}\big]\big),$$

and the proof is complete. □

Proof of Theorem 2

In view of Theorem 1, let $\mathrm{AB}^2(\mathbf{G}) = \alpha_1(\mathbf{G}) + \alpha_2(\mathbf{G}) + \alpha_3(\mathbf{G})$, where

$$\alpha_1(\mathbf{G}) = n^{-2}|\mathbf{G}|^{-1}\operatorname{tr}\big[\boldsymbol{\ell}_r\boldsymbol{\ell}_r^T\big(\mathbf{G}^{-1}\big)^{\otimes r}\big]$$

$$\alpha_2(\mathbf{G}) = n^{-1}m_2(L)|\mathbf{G}|^{-1/2}\operatorname{tr}\big[\boldsymbol{\psi}_{r+2}\boldsymbol{\ell}_r^T\big(\mathbf{G}^{-1/2}\big)^{\otimes r}\big(\operatorname{vec}^T\mathbf{G}\otimes\mathbf{I}_{d^r}\big)\big]$$

$$\alpha_3(\mathbf{G}) = \frac{m_2(L)^2}{4}\operatorname{tr}\big[\big(\boldsymbol{\psi}_{r+2}\boldsymbol{\psi}_{r+2}^T\big)\big(\operatorname{vec}\mathbf{G}\operatorname{vec}^T\mathbf{G}\otimes\mathbf{I}_{d^r}\big)\big]$$

with $\boldsymbol{\ell}_r = \mathsf{D}^{\otimes r}L(0)$. We will find the derivative of AB^2 and the order of $\mathbf{G}$ which attains a zero derivative.

For α_1, the differential is

$$d\alpha_1(\mathbf{G}) = -\alpha_1(\mathbf{G})\big(\operatorname{vec}^T\mathbf{G}^{-1}\big)\,d\operatorname{vec}\mathbf{G} + n^{-2}|\mathbf{G}|^{-1}d\operatorname{tr}\big[\boldsymbol{\ell}_r\boldsymbol{\ell}_r^T\big(\mathbf{G}^{-1}\big)^{\otimes r}\big]$$

since $d|\mathbf{G}|^{-1} = -|\mathbf{G}|^{-2}d|\mathbf{G}| = -|\mathbf{G}|^{-2}|\mathbf{G}|\operatorname{tr}(\mathbf{G}^{-1}d\mathbf{G}) = -|\mathbf{G}|^{-1}(\operatorname{vec}^T\mathbf{G}^{-1}) \times (d\operatorname{vec}\mathbf{G})$. The trace can be expressed as $\operatorname{vec}^T(\boldsymbol{\ell}_r\boldsymbol{\ell}_r^T)\,d\operatorname{vec}((\mathbf{G}^{-1})^{\otimes r})$. We have $d\operatorname{vec}\mathbf{G}^{-1} = \operatorname{vec}(d\mathbf{G}^{-1}) = -\operatorname{vec}(\mathbf{G}^{-1}(d\mathbf{G})\mathbf{G}^{-1}) = -(\mathbf{G}^{-1}\otimes\mathbf{G}^{-1})\,d\operatorname{vec}\mathbf{G}$, which implies that

$$d\big(\operatorname{vec}\big(\mathbf{G}^{-1}\big)^{\otimes r}\big) = d\big(\operatorname{vec}\big(\mathbf{G}^{\otimes r}\big)^{-1}\big) = -\big(\mathbf{G}^{-1}\otimes\mathbf{G}^{-1}\big)^{\otimes r}\,d\operatorname{vec}\big(\mathbf{G}^{\otimes r}\big)$$

$$= -\big(\mathbf{G}^{-1}\big)^{\otimes 2r}\,d\operatorname{vec}\big(\mathbf{G}^{\otimes r}\big).$$

This leaves us to find $d\operatorname{vec}(\mathbf{G}^{\otimes r})$:

$$d\operatorname{vec}\big(\mathbf{G}^{\otimes r}\big) = \sum_{i=1}^{r}\operatorname{vec}\big(\mathbf{G}^{\otimes(i-1)}\otimes d\mathbf{G}\otimes\mathbf{G}^{\otimes(r-i)}\big)$$

$$= \sum_{i=1}^{r}\operatorname{vec}\big[\mathbf{K}_{d^i,d^{r-i}}\big(\mathbf{G}^{\otimes(r-i)}\otimes\mathbf{G}^{\otimes(i-1)}\otimes d\mathbf{G}\big)\mathbf{K}_{d^{r-i},d^i}\big]$$

$$= \sum_{i=1}^{r}\mathbf{K}_{d^i,d^{r-i}}^{\otimes 2}\operatorname{vec}\big(\mathbf{G}^{\otimes(r-1)}\otimes d\mathbf{G}\big)$$

$$= \sum_{i=1}^{r}\mathbf{K}_{d^i,d^{r-i}}^{\otimes 2}\big[\big\{(\mathbf{I}_{d^{r-1}}\otimes\mathbf{K}_{d,d^{r-1}})\big(\operatorname{vec}\mathbf{G}^{\otimes(r-1)}\otimes\mathbf{I}_d\big)\big\}\otimes\mathbf{I}_d\big]d\operatorname{vec}\mathbf{G},$$

where the second line follows from Schott (2005, p. 311) for commuting a 3-fold Kronecker product, with $\mathbf{K}_{m,n}$ the commutation matrix of orders m,n, and the fourth line follows from Magnus and Neudecker (1999, p. 48). Thus,

$$\begin{aligned} d\alpha_1(\mathbf{G}) = -\big\{&\alpha_1(\mathbf{G})\big(\mathrm{vec}^T\,\mathbf{G}^{-1}\big) + n^{-2}|\mathbf{G}|^{-1}\big(\boldsymbol{\ell}_r^T\otimes\boldsymbol{\ell}_r^T\big)\big(\mathbf{G}^{-1}\big)^{\otimes 2r} \\ &\times \boldsymbol{\Lambda}_r\big[\big\{(\mathbf{I}_{d^{r-1}}\otimes\mathbf{K}_{d,d^{r-1}})\big(\mathrm{vec}\,\mathbf{G}^{\otimes(r-1)}\otimes\mathbf{I}_d\big)\big\}\otimes\mathbf{I}_d\big]\big\}d\,\mathrm{vec}\,\mathbf{G}, \qquad (9)\end{aligned}$$

where $\boldsymbol{\Lambda}_r=\sum_{i=1}^r\mathbf{K}_{d^i,d^{r-i}}^{\otimes 2}$, since $\mathrm{vec}(\boldsymbol{a}\boldsymbol{b}^T)=\boldsymbol{b}\otimes\boldsymbol{a}$ for vectors $\boldsymbol{a}$ and $\boldsymbol{b}$.

For α_2, its differential is

$$\begin{aligned} d\alpha_2(\mathbf{G}) = -&\frac{1}{2}\alpha_2(\mathbf{G})\big(\mathrm{vec}^T\,\mathbf{G}^{-1}\big)\,d\,\mathrm{vec}\,\mathbf{G} \\ &+n^{-1}m_2(L)|\mathbf{G}|^{-1/2}d\,\mathrm{tr}\big[\boldsymbol{\psi}_{r+2}\boldsymbol{\ell}_r^T\big(\mathbf{G}^{-1/2}\big)^{\otimes r}\big(\mathrm{vec}^T\,\mathbf{G}\otimes\mathbf{I}_{d^r}\big)\big]\end{aligned}$$

since we have $d|\mathbf{G}|^{-1/2}=-\frac{1}{2}|\mathbf{G}|^{-3/2}d|\mathbf{G}|=-\frac{1}{2}|\mathbf{G}|^{-1/2}(\mathrm{vec}^T\,\mathbf{G}^{-1})(d\,\mathrm{vec}\,\mathbf{G})$. The trace can be expressed as $\mathrm{vec}^T(\boldsymbol{\ell}_r\boldsymbol{\psi}_{r+2}^T)\,d\,\mathrm{vec}[(\mathbf{G}^{-1/2})^{\otimes r}(\mathrm{vec}^T\,\mathbf{G}\otimes\mathbf{I}_{d^r})]$. Then

$$\begin{aligned} d\,\mathrm{vec}&\big[\big(\mathbf{G}^{-1/2}\big)^{\otimes r}\big(\mathrm{vec}^T\,\mathbf{G}\otimes\mathbf{I}_{d^r}\big)\big] \\ &=\mathrm{vec}\big[\big(\mathbf{G}^{-1/2}\big)^{\otimes r}\big(d\,\mathrm{vec}^T\,\mathbf{G}\otimes\mathbf{I}_{d^r}\big)\big]+\mathrm{vec}\big[d\big(\big(\mathbf{G}^{-1/2}\big)^{\otimes r}\big)\big(\mathrm{vec}^T\,\mathbf{G}\otimes\mathbf{I}_{d^r}\big)\big] \\ &=\big(\mathbf{I}_{d^{r+2}}\otimes\big(\mathbf{G}^{-1/2}\big)^{\otimes r}\big)\,\mathrm{vec}\big(d\,\mathrm{vec}^T\,\mathbf{G}\otimes\mathbf{I}_{d^r}\big)+(\mathrm{vec}\,\mathbf{G}\otimes\mathbf{I}_{d^{2r}})\,d\,\mathrm{vec}\big(\big(\mathbf{G}^{-1/2}\big)^{\otimes r}\big) \\ &=\big(\mathbf{I}_{d^{r+2}}\otimes\big(\mathbf{G}^{-1/2}\big)^{\otimes r}\big)(\mathbf{I}_{d^2}\otimes\mathrm{vec}\,\mathbf{I}_{d^r})\,d\,\mathrm{vec}\,\mathbf{G} \\ &\quad+(\mathrm{vec}\,\mathbf{G}\otimes\mathbf{I}_{d^{2r}})\,d\,\mathrm{vec}\big(\big(\mathbf{G}^{-1/2}\big)^{\otimes r}\big),\end{aligned}$$

where the last lines follow from Magnus and Neudecker (1999, p. 48): for $\mathbf{B}\in\mathcal{M}_{m\times n}$ and $\boldsymbol{b}\in\mathbb{R}^p$, $\mathrm{vec}(\boldsymbol{b}^T\otimes\mathbf{B})=(\mathbf{I}_p\otimes\mathrm{vec}\,\mathbf{B})\boldsymbol{b}$. To evaluate $d\,\mathrm{vec}((\mathbf{G}^{-1/2})^{\otimes r})$, we start with the identity $\mathbf{G}^{-1/2}\mathbf{G}^{-1/2}=\mathbf{G}^{-1}$; then taking differentials and applying the vec operator, we obtain $d\,\mathrm{vec}\,\mathbf{G}^{-1/2}=-(\mathbf{G}^{1/2}\otimes\mathbf{G}+\mathbf{G}\otimes\mathbf{G}^{1/2})^{-1}\,d\,\mathrm{vec}\,\mathbf{G}$, which implies that

$$\begin{aligned} d\,\mathrm{vec}\big(\big(\mathbf{G}^{-1/2}\big)^{\otimes r}\big)&=d\,\mathrm{vec}\big(\big(\mathbf{G}^{\otimes r}\big)^{-1/2}\big) \\ &=-\big[\big(\mathbf{G}^{1/2}\big)^{\otimes r}\otimes\mathbf{G}^{\otimes r}+\mathbf{G}^{\otimes r}\otimes\big(\mathbf{G}^{1/2}\big)^{\otimes r}\big]^{-1}\,d\,\mathrm{vec}\big(\mathbf{G}^{\otimes r}\big).\end{aligned}$$

So we have

$$\begin{aligned} d\alpha_2&(\mathbf{G}) \\ &=\Big(-\frac{1}{2}\alpha_2(\mathbf{G})\big(\mathrm{vec}^T\,\mathbf{G}^{-1}\big)+n^{-1}m_2(L)|\mathbf{G}|^{-1/2}\big(\boldsymbol{\psi}_{r+2}^T\otimes\boldsymbol{\ell}_r^T\big) \\ &\quad\times\big\{\big[\mathbf{I}_{d^2}\otimes\mathrm{vec}\big(\big(\mathbf{G}^{-1/2}\big)^{\otimes r}\big)\big]-(\mathrm{vec}\,\mathbf{G}\otimes\mathbf{I}_{d^{2r}})\big[\big(\mathbf{G}^{1/2}\big)^{\otimes r}\otimes\mathbf{G}^{\otimes r} \\ &\quad+\mathbf{G}^{\otimes r}\otimes\big(\mathbf{G}^{1/2}\big)^{\otimes r}\big]^{-1}\boldsymbol{\Lambda}_r\big[\big\{(\mathbf{I}_{d^{r-1}}\otimes\mathbf{K}_{d,d^{r-1}})\big(\mathrm{vec}\,\mathbf{G}^{\otimes(r-1)}\otimes\mathbf{I}_d\big)\big\}\otimes\mathbf{I}_d\big]\big\}\Big) \\ &\quad\times d\,\mathrm{vec}\,\mathbf{G}. \qquad (10)\end{aligned}$$

For α_3, the differential is $d\alpha_3(\mathbf{G}) = \frac{m_2(L)^2}{4}\, d\,\mathrm{tr}[(\boldsymbol{\psi}_{r+2}\boldsymbol{\psi}_{r+2}^T)(\mathrm{vec}\,\mathbf{G}\,\mathrm{vec}^T\,\mathbf{G} \otimes \mathbf{I}_{d^r})]$, and the trace can be rewritten as $\mathrm{vec}^T(\boldsymbol{\psi}_{r+2}\boldsymbol{\psi}_{r+2}^T)\,\mathrm{vec}(\mathrm{vec}\,\mathbf{G}\,\mathrm{vec}^T\,\mathbf{G} \otimes \mathbf{I}_{d^r})$. The differential of the latter part is

$$\begin{aligned}
&d\,\mathrm{vec}\big(\mathrm{vec}\,\mathbf{G}\,\mathrm{vec}^T\,\mathbf{G} \otimes \mathbf{I}_{d^r}\big)\\
&\quad = \mathrm{vec}\big(d\,\mathrm{vec}\,\mathbf{G} \otimes \mathrm{vec}^T\,\mathbf{G} \otimes \mathbf{I}_{d^r} + \mathrm{vec}\,\mathbf{G} \otimes d\,\mathrm{vec}^T\,\mathbf{G} \otimes \mathbf{I}_{d^r}\big)\\
&\quad = \mathrm{vec}\big(d\,\mathrm{vec}\,\mathbf{G} \otimes \big(\mathrm{vec}^T\,\mathbf{G} \otimes \mathbf{I}_{d^r}\big) + d\,\mathrm{vec}^T\,\mathbf{G} \otimes (\mathrm{vec}\,\mathbf{G} \otimes \mathbf{I}_{d^r})\big)\\
&\quad = (\mathbf{K}_{d^{r+2},d^2} \otimes \mathbf{I}_{d^r})\big[\mathbf{I}_{d^2} \otimes \mathrm{vec}\big(\mathrm{vec}^T\,\mathbf{G} \otimes \mathrm{vec}\,\mathbf{I}_{d^r}\big)\big]d\,\mathrm{vec}\,\mathbf{G}\\
&\qquad + \big[\mathbf{I}_{d^2} \otimes \mathrm{vec}(\mathrm{vec}\,\mathbf{G} \otimes \mathrm{vec}\,\mathbf{I}_{d^r})\big]d\,\mathrm{vec}\,\mathbf{G}
\end{aligned}$$

from Magnus and Neudecker (1999, p. 48): for $\mathbf{B} \in \mathcal{M}_{m\times n}$ and $\boldsymbol{b} \in \mathbb{R}^p$, $\mathrm{vec}(\boldsymbol{b} \otimes \mathbf{B}) = (\mathbf{K}_{np} \otimes \mathbf{I}_m)(\mathbf{I}_p \otimes \mathrm{vec}\,\mathbf{B})\boldsymbol{b}$. This can be further expanded, to give

$$\begin{aligned}
&d\,\mathrm{vec}\big(\mathrm{vec}\,\mathbf{G}\,\mathrm{vec}^T\,\mathbf{G} \otimes \mathbf{I}_{d^r}\big)\\
&\quad = \big[(\mathbf{K}_{d^{r+2},d^2} \otimes \mathbf{I}_{d^r}) + \mathbf{I}_{d^{2r+4}}\big](\mathbf{I}_{d^2} \otimes \mathrm{vec}\,\mathbf{G} \otimes \mathrm{vec}\,\mathbf{I}_{d^r})d\,\mathrm{vec}\,\mathbf{G},
\end{aligned}$$

from which we find that

$$\begin{aligned}
d\alpha_3(\mathbf{G}) = \frac{m(L)^2}{4}\big(\boldsymbol{\psi}_{r+2}^T \otimes \boldsymbol{\psi}_{r+2}^T\big)\big[(\mathbf{K}_{d^{r+2},d^2} \otimes \mathbf{I}_{d^r}) + \mathbf{I}_{d^{2r+4}}\big]&\\
\times (\mathbf{I}_{d^2} \otimes \mathrm{vec}\,\mathbf{G} \otimes \mathrm{vec}\,\mathbf{I}_{d^r})d\,\mathrm{vec}\,\mathbf{G}.& \qquad (11)
\end{aligned}$$

If we assume that $\mathbf{G} = O(\mathbf{J}_d n^{-\beta})$ for some $\beta > 0$, then $d\alpha_1(\mathbf{G}) = O(n^{\beta(d+r+1)-2})d\,\mathrm{vec}\,\mathbf{G}$, $d\alpha_2(\mathbf{G}) = O(n^{\beta(d+r)/2-1})d\,\mathrm{vec}\,\mathbf{G}$ and $d\alpha_3(\mathbf{G}) = O(n^{-\beta})d\,\mathrm{vec}\,\mathbf{G}$. Equating powers of n gives $\beta = 2/(r+d+2)$, which means that the solution to $\partial \mathrm{AB}^2(\mathbf{G})/\partial\,\mathrm{vec}\,\mathbf{G} = 0$, $\mathbf{G}_{\mathrm{AMSE},r}$, is of order $n^{-2/(r+d+2)}$. To complete the proof, $\mathrm{AB}^2(\mathbf{G}_{\mathrm{AMSE},r}) = O(n^{-4/(r+d+2)})$ and $\mathrm{AV}(\mathbf{G}_{\mathrm{AMSE},r}) = O(n^{-\min\{r+d+2,d+4\}/(r+d+2)})$, i.e., $\mathrm{MSE}(\mathbf{G}_{\mathrm{AMSE},r}) = O(n^{-\min\{r+d+2,4\}/(r+d+2)})$.

Proof of the normal reference formulas

Proof of formula (7) Applying Fact C.2.3 in Wand and Jones (1995) to every element in $\boldsymbol{\psi}_r^{\mathrm{NR}}$, we can show that

$$\begin{aligned}
\boldsymbol{\psi}_r^{\mathrm{NR}} &= \int_{\mathbb{R}^d} \mathsf{D}^{\otimes r}\phi_{\boldsymbol{\Sigma}}(\boldsymbol{x})\phi_{\boldsymbol{\Sigma}}(\boldsymbol{x})\,d\boldsymbol{x} = (-1)^r\mathsf{D}^{\otimes r}\phi_{2\boldsymbol{\Sigma}}(0)\\
&= (-1)^r|2\boldsymbol{\Sigma}|^{-1/2}\big[(2\boldsymbol{\Sigma})^{-1/2}\big]^{\otimes r}\mathsf{D}^{\otimes r}\phi(0)\\
&= 2^{-(r+d)/2}|\boldsymbol{\Sigma}|^{-1/2}\big(\boldsymbol{\Sigma}^{-1/2}\big)^{\otimes r}\phi(0)\boldsymbol{\mathcal{H}}_r(0)\\
&= (-1)^{r/2}(2\pi)^{-d/2}\mathrm{OF}(r)2^{-(r+d)/2}|\boldsymbol{\Sigma}|^{-1/2}\big(\boldsymbol{\Sigma}^{-1/2}\big)^{\otimes r}\boldsymbol{\mathcal{S}}_{d,r}(\mathrm{vec}\,\mathbf{I}_d)^{\otimes(r/2)},
\end{aligned}$$

and we are done, as we can interchange $(\boldsymbol{\Sigma}^{-1/2})^{\otimes r}\boldsymbol{\mathcal{S}}_{d,r} = \boldsymbol{\mathcal{S}}_{d,r}(\boldsymbol{\Sigma}^{-1/2})^{\otimes r}$ by part (vii) of Theorem 1 in Schott (2003) and obtain

$$(\boldsymbol{\Sigma}^{-1/2})^{\otimes r}(\operatorname{vec}\mathbf{I}_d)^{\otimes(r/2)} = [(\boldsymbol{\Sigma}^{-1/2}\otimes\boldsymbol{\Sigma}^{-1/2})\operatorname{vec}\mathbf{I}_d]^{\otimes(r/2)} = \big(\operatorname{vec}\boldsymbol{\Sigma}^{-1}\big)^{\otimes(r/2)},$$

using aforementioned properties of the Kronecker product. □

The proof of formula (8) is more laborious. The techniques used here are similar to those needed for the normal calculations included in Chacón et al. (2009) For a d-variate random vector z with standard normal distribution, denote by $\boldsymbol{\mu}_p = \mathbb{E}[z^{\otimes p}] \in \mathbb{R}^{d^p}$ its pth order vector moment and by $\nu_p(\mathbf{A}) = \mathbb{E}[(z^T\mathbf{A}z)^p]$ the pth order moment of the quadratic form $z^T\mathbf{A}z$ for a symmetric matrix $\mathbf{A}\in\mathcal{M}_{d\times d}$.

Lemma 3 *For even r, the following relations hold*:

(i) $\boldsymbol{\mu}_r^T(\boldsymbol{\Sigma}^{-1})^{\otimes r}\boldsymbol{\mu}_r = \mathrm{OF}(r)\nu_{r/2}(\boldsymbol{\Sigma}^{-2})$.
(ii) $\boldsymbol{\mu}_r^T[\operatorname{vec}^T\mathbf{I}_d\otimes(\boldsymbol{\Sigma}^{-1})^{\otimes r}]\boldsymbol{\mu}_{r+2} = (r+d)\mathrm{OF}(r)\nu_{r/2}(\boldsymbol{\Sigma}^{-2})$.

Proof (i) According to Holmquist (1996a), it is possible to write

$$\boldsymbol{\mu}_r = \mathrm{OF}(r)\boldsymbol{\mathcal{S}}_{d,r}(\operatorname{vec}\mathbf{I}_d)^{\otimes(r/2)}. \tag{12}$$

Therefore, using Theorem 1 in Schott (2003), we have

$$\begin{aligned}\boldsymbol{\mu}_r^T\big(\boldsymbol{\Sigma}^{-1}\big)^{\otimes r}\boldsymbol{\mu}_r &= \mathrm{OF}(r)^2\big(\operatorname{vec}^T\mathbf{I}_d\big)^{\otimes(r/2)}\boldsymbol{\mathcal{S}}_{d,r}\big(\boldsymbol{\Sigma}^{-1}\big)^{\otimes r}\boldsymbol{\mathcal{S}}_{d,r}(\operatorname{vec}\mathbf{I}_d)^{\otimes(r/2)}\\ &= \mathrm{OF}(r)^2\big(\operatorname{vec}^T\boldsymbol{\Sigma}^{-2}\big)^{\otimes(r/2)}\boldsymbol{\mathcal{S}}_{d,r}(\operatorname{vec}\mathbf{I}_d)^{\otimes(r/2)}\\ &= \mathrm{OF}(r)\nu_{r/2}\big(\boldsymbol{\Sigma}^{-2}\big),\end{aligned}$$

where the last line follows from Theorem 1 in Holmquist (1996b).

(ii) As $\boldsymbol{\mu}_r = (-1)^{r/2}\boldsymbol{\mathcal{H}}_r(0)$, applying the recurrence relation in Theorem 7.2 of Holmquist (1996a), we get $\boldsymbol{\mu}_{r+2} = (r+1)\boldsymbol{\mathcal{S}}_{d,r+2}(\operatorname{vec}\mathbf{I}_d\otimes\boldsymbol{\mu}_r)$. Thus, by comparison with (12) it follows that $\boldsymbol{\mathcal{S}}_{d,r+2}(\operatorname{vec}\mathbf{I}_d\otimes\boldsymbol{\mu}_r) = \mathrm{OF}(r)\boldsymbol{\mathcal{S}}_{d,r+2}(\operatorname{vec}\mathbf{I}_d)^{\otimes(r/2+1)}$. Then,

$$\begin{aligned}&\boldsymbol{\mu}_r^T\big[\operatorname{vec}^T\mathbf{I}_d\otimes\big(\boldsymbol{\Sigma}^{-1}\big)^{\otimes r}\big]\boldsymbol{\mu}_{r+2}\\ &\quad= \mathrm{OF}(r+2)\big(\operatorname{vec}^T\mathbf{I}_d\otimes\boldsymbol{\mu}_r^T\big)\boldsymbol{\mathcal{S}}_{d,r+2}\big[\mathbf{I}_{d^2}\otimes\big(\boldsymbol{\Sigma}^{-1}\big)^{\otimes r}\big](\operatorname{vec}\mathbf{I}_d)^{\otimes(r/2+1)}\\ &\quad= \mathrm{OF}(r)\mathrm{OF}(r+2)\big[\operatorname{vec}^T\mathbf{I}_d\otimes\big(\operatorname{vec}^T\boldsymbol{\Sigma}^{-2}\big)^{\otimes(r/2)}\big]\boldsymbol{\mathcal{S}}_{d,r+2}(\operatorname{vec}\mathbf{I}_d)^{\otimes(r/2+1)}\\ &\quad= \mathrm{OF}(r)\mathbb{E}\big[\big(z^T\boldsymbol{\Sigma}^{-2}z\big)^{r/2}(z^Tz)\big] = (r+d)\mathrm{OF}(r)\nu_{r/2}\big(\boldsymbol{\Sigma}^{-2}\big),\end{aligned}$$

where the third equality follows from Theorem 5 in Holmquist (1996b), and the fourth one from Chacón et al. (2009). □

Now we are ready to prove formula (8).

Proof of formula (8) First, using (7) and (12), it is not hard to show that, in the normal case,

$$\mathrm{AB}^2(\mathbf{G}) = (2\pi)^{-d}\left\|n^{-1}|\mathbf{G}|^{-1/2}\left(\mathbf{G}^{-1/2}\right)^{\otimes r}\boldsymbol{\mu}_r - 2^{-(r+d+4)/2}|\boldsymbol{\Sigma}|^{-1/2}\right.$$
$$\left.\times\left[\mathrm{vec}^T\left(\boldsymbol{\Sigma}^{-1/2}\mathbf{G}\boldsymbol{\Sigma}^{-1/2}\right)\otimes\left(\boldsymbol{\Sigma}^{-1/2}\right)^{\otimes r}\right]\boldsymbol{\mu}_{r+2}\right\|^2.$$

For $\mathbf{G} = c\boldsymbol{\Sigma}$, writing $\boldsymbol{v}_1 = (\boldsymbol{\Sigma}^{-1/2})^{\otimes r}\boldsymbol{\mu}_r$ and $\boldsymbol{v}_2 = [\mathrm{vec}^T\,\mathbf{I}_d \otimes (\boldsymbol{\Sigma}^{-1/2})^{\otimes r}]\boldsymbol{\mu}_{r+2}$, the previous formula reduces to

$$\begin{aligned}\mathrm{AB}^2(c\boldsymbol{\Sigma}) &= (2\pi)^{-d}|\boldsymbol{\Sigma}|^{-1}\left\|n^{-1}c^{-(r+d)/2}\boldsymbol{v}_1 - 2^{-(r+d+4)/2}c\boldsymbol{v}_2\right\|^2\\ &= (2\pi)^{-d}|\boldsymbol{\Sigma}|^{-1}\left\{n^{-2}c^{-(r+d)}\boldsymbol{v}_1^T\boldsymbol{v}_1 - 2n^{-1}2^{-(r+d+4)/2}c^{(2-r-d)/2}\boldsymbol{v}_1^T\boldsymbol{v}_2\right.\\ &\quad\left.+ 2^{-(r+d+4)}c^2\boldsymbol{v}_2^T\boldsymbol{v}_2\right\}\\ &= (2\pi)^{-d}|\boldsymbol{\Sigma}|^{-1}c^2\left\{n^{-2}c^{-(r+d+2)}\boldsymbol{v}_1^T\boldsymbol{v}_1\right.\\ &\quad\left.- 2n^{-1}2^{-(r+d+4)/2}c^{-(r+d+2)/2}\boldsymbol{v}_1^T\boldsymbol{v}_2 + 2^{-(r+d+4)}\boldsymbol{v}_2^T\boldsymbol{v}_2\right\},\end{aligned}$$

and the term inside the braces is quadratic in $\Theta = c^{-(r+d+2)/2}$ with positive coefficient $n^{-2}\boldsymbol{v}_1^T\boldsymbol{v}_1$ for Θ^2. Therefore, its minimizer is given by $\Theta_0 = 2^{-(r+d+4)/2} \times n(\boldsymbol{v}_1^T\boldsymbol{v}_2)/(\boldsymbol{v}_1^T\boldsymbol{v}_1)$. We finish the proof by noting that, according to the previous lemma, $(\boldsymbol{v}_1^T\boldsymbol{v}_2)/(\boldsymbol{v}_1^T\boldsymbol{v}_1) = r + d$. □

Proof of Theorem 3

The proof of this theorem is straightforward since it relies on previous results from both previous research and this paper. Reasoning as in Wand and Jones (1994, pp. 106–107), and making use of Lemma 1 in Duong and Hazelton (2005a), we obtain that the relative rate of convergence of $\hat{\mathbf{H}}_{\mathrm{PI}}$ to $\mathbf{H}_{\mathrm{AMISE}}$ equals $n^{-\alpha}$ whenever $\mathrm{MSE}(\mathbf{G})$ is of order $n^{-2\alpha}$. As Theorem 2 for $r = 4$ gives $\mathrm{MSE}(\mathbf{G}) = O(n^{-4/(d+6)})$, we are done.

References

Bowman AW (1984) An alternative method of cross-validation for the smoothing of density estimates. Biometrika 71:353–360

Chacón JE (2009) Data-driven choice of the smoothing parametrization for kernel density estimators. Can J Stat 37:249–265

Chacón JE, Duong T, Wand MP (2009) Asymptotics for general multivariate kernel density derivative estimators (submitted)

Duong T (2009) ks: Kernel smoothing. R package version 1.6.5

Duong T, Hazelton ML (2003) Plug-in bandwidth matrices for bivariate kernel density estimation. J Nonparametr Stat 15:17–30

Duong T, Hazelton ML (2005a) Convergence rates for unconstrained bandwidth matrix selectors in multivariate kernel density estimation. J Multivar Anal 93:417–433

Duong T, Hazelton ML (2005b) Cross-validation bandwidth matrices for multivariate kernel density estimation. Scand J Stat 32:485–506

Hall P, Marron JS (1987) Estimation of integrated squared density derivatives. Stat Probab Lett 6:109–115

Hall P, Marron JS, Park BU (1992) Smoothed cross-validation. Probab Theory Relat Fields 92:1–20

Henderson HV, Searle SR (1979) Vec and vech operators for matrices, with some uses in Jacobians and multivariate statistics. Can J Stat 7:65–81

Holmquist B (1985) The direct product permuting matrices. Linear Multilinear Algebra 17:117–141

Holmquist B (1996a) The d-variate vector Hermite polynomial of order k. Linear Algebra Appl 237–238:155–190

Holmquist B (1996b) Expectations of products of quadratic forms in normal variables. Stoch Anal Appl 14:149–164

Jones MC, Sheather SJ (1991) Using nonstochastic terms to advantage in kernel-based estimation of integrated squared density derivatives. Stat Probab Lett 11:511–514

Magnus JR, Neudecker H (1980) The elimination matrix: some lemmas and applications. SIAM J Algebra Discrete Methods 1:422–449

Magnus JR, Neudecker H (1999) Matrix differential calculus with applications in statistics and econometrics, revised edn. Wiley, Chichester

Meijer E (2005) Matrix algebra for higher order moments. Linear Algebra Appl 410:112–134

Park BU, Marron JS (1990) Comparison of data-driven bandwidth selectors. J Am Stat Assoc 85:66–72

Rosenblatt M (1956) Remarks on some nonparametric estimates of a density function. Ann Math Stat 27:832–837

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

Sain SR, Baggerly KA, Scott DW (1994) Cross-validation of multivariate densities. J Am Stat Assoc 89:807–817

Schott JR (2003) Kronecker product permutation matrices and their application to moment matrices of the normal distribution. J Multivar Anal 87:177–190

Schott JR (2005) Matrix analysis for statics, 2nd edn. Wiley, New York

Scott DW, Terrell GR (1987) Biased and unbiased cross-validation in density estimation. J Am Stat Assoc 82:1131–1146

Sheather SJ, Jones MC (1991) A reliable data-based bandwidth selection method for kernel density estimation. J R Stat Soc Ser B Stat Methodol 53:683–690

Silverman BW (1986) Density estimation for statics and data analysis. Chapman & Hall, London

Simonoff JS (1996) Smoothing methods in statics. Springer, Berlin

Wand MP (1992) Error analysis for general multivariate kernel estimators. J Nonparametr Stat 2:2–15

Wand MP, Jones MC (1993) Comparison of smoothing parameterizations in bivariate kernel density estimation. J Am Stat Assoc 88:520–528

Wand MP, Jones MC (1994) Multivariate plug-in bandwidth selection. Comput Stat 9:97–117

Wand MP, Jones MC (1995) Kernel smoothing. Chapman & Hall, London