2009

Asymptotics for General Multivariate Kernel Density Derivative Estimators

J. E. Chacon
Universidad de Extremadura, Spain

T. Duon
Institut Pasteur, France

M. P. Wand
University of Wollongong, mwand@uow.edu.au

Centre for Statistical and Survey Methodology

The University of Wollongong

Working Paper

08-09

Asymptotics for General Multivariate Kernel Density Derivative Estimators.

Chacon, J.E., Duong, T. and Wand, M.P.

Centre for Statistical & Survey Methodology, University of Wollongong, Wollongong NSW 2522. Phone +61 2 4221 5435, Fax +61 2 4221 4845. Email: anica@uow.edu.au

Asymptotics for general multivariate kernel density derivative estimators

José E. Chacón[*], Tarn Duong[†] and M. P. Wand[‡]

May 28, 2009

Abstract

We investigate general kernel density derivative estimators, that is, kernel estimators of multivariate density derivative functions using general (or unconstrained) bandwidth matrix selectors. These density derivative estimators have been relatively less well researched than their density estimator analogues. A major obstacle for progress has been the intractability of the matrix analysis when treating higher order multivariate derivatives. With an alternative vectorization of these higher order derivatives, these mathematical intractabilities are surmounted in an elegant and unified framework. The finite sample and asymptotic analysis of squared errors for density estimators are generalized to density derivative estimators. Moreover, we are able to exhibit a closed form expression for a normal scale bandwidth matrix for density derivative estimators. These normal scale bandwidths are employed in a numerical study to demonstrate the gain in performance of unconstrained selectors over their constrained counterparts.

Keywords: asymptotic mean integrated squared error, normal scale rule, optimal, unconstrained bandwidth matrices.

[*]Departamento de Matemáticas, Universidad de Extremadura, Spain. E-mail: jechacon@unex.es

[†]Institut Pasteur, Groupe Imagerie et Modélisation; CNRS, URA 2582, F-75015, Paris, France. Email: tduong@pasteur.fr

[‡]School of Mathematics and Applied Statistics, University of Wollongong, Wollongong, Australia. E-mail: mwand@uow.edu.au

1 Introduction

Estimating probability density functions with kernel functions has had notable success due to their ease of interpretation and visualization. On the other hand, estimating derivatives of density functions has received less attention. This is partially because it is a more challenging theoretical problem (especially for multivariate data). Nonetheless there remains much information about the structure of a density function which is not easily ascertained from the density function itself. For example, the local maxima and minima are where there are zero first derivatives and non-zero second derivatives. One of the original papers on kernel density estimation (Parzen, 1962) was also concerned with the estimating the global mode of the density function, though not from a density derivative point of view. We can recast this problem as a density derivative estimation problem: find the local maxima via derivative estimates and the global mode follows as the largest of these local maxima. The focus on a mode as a single point can be extended to the region immediately surrounding the mode, known as a bump or modal region. Modal regions can be used to determine the existence of multi-modality and/or clusters. Godtliebsen, Marron and Chaudhuri (2002)'s feature significance technique for bump-hunting relies on estimating and characterizing the first and second derivatives for bivariate data. In an econometrics setting, the Engel curve describes the demand for a good/service as a function of income. It classifies goods/services based on the slope of their Engel curve so the first derivative is an essential component for interpreting these curves, see Hildenbrand and Hildenbrand (1986). In a more general setting, Singh (1977) suggests as an application of the multivariate density derivatives to estimate the Fisher information matrix in its translation parameter form.

The first paper to be concerned with univariate kernel density derivative estimation appears to be Bhattacharya (1967), followed by Schuster (1969) and Singh (1979, 1987). Singh (1976) studies a multivariate estimator with a diagonal bandwidth matrix, and Härdle, Marron and Wand (1990) with the bandwidth parametrized as a constant times the identity matrix. This previous research has mostly focused on constrained parametrizations of the bandwidth matrix since they simplify the matrix analysis compared to the general, unconstrained parametrization. Analyzing squared error measures for general kernel density derivative estimators has reached an impasse using the traditional vectorization of higher order derivatives of multivariate functions (a vectorization transforms the higher order derivative tensor into a more tractable vector form). To tackle this problem, we introduce an alternative vectorization of higher order derivatives. This vectorization is a subtle rearrangement of the traditional vectorization which allows us to write down with the same ease all the usual error expressions from the density estimation case. Thus we generalize the usual squared error analysis for kernel density derivative estimators. Furthermore, we are able to write down a closed form expression for a normal scale bandwidth matrix, i.e., the optimal bandwidth for the rth order derivative of a normal density with normal kernels.

Normal scale selectors were the first step which eventually lead to the development of the now widely used bandwidth selectors for density estimation: we set up a similar starting point for future bandwidth selection in density derivative estimation.

In Section 2 we define a kernel estimator of a multivariate density derivative, and provide the results for mean integrated square convergence both asymptotically and for finite samples. The influence of the bandwidth matrix on convergence is established here. In Section 3 we focus on the class of normal mixture densities. Estimation of these densities with normal kernels produces further simplified special cases of the results in the Section 2, where we develop a normal scale bandwidth selector. In Section 4, we use these normal scale selectors to quantify the improvement in asymptotic performance when using unconstrained matrices. We illustrate the normal scale selectors on data arising from high throughput biotechnology in Section 5. The usual normal scale selectors based on the density function may lead to insufficient smoothing when estimating the density curvature (or second derivative). We conclude with a discussion in Section 6.

2 Kernel density derivative estimation

The current state of multivariate kernel density estimation has reached maturity, and recent advances there can be carried over to the density derivative case. To proceed, we use the linearity of the kernel density estimator to define a kernel density derivative estimator. The usual performance measure for kernel density estimation is the mean integrated squared error (MISE) which is easily extended to cover density derivatives.

We introduce some notation needed to state our problem. For a matrix $\mathbf{A}$ we denote

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

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 will adopt the convention $\mathbf{A}^{\otimes 1} = \mathbf{A}$ and $\mathbf{A}^{\otimes 0} = 1 \in \mathbb{R}$.

If $f\colon \mathbb{R}^d \to \mathbb{R}$ is a real function of a vector variable we denote $\mathsf{D}^{\otimes r} f(\boldsymbol{x}) \in \mathbb{R}^{d^r}$ the vector containing all the partial derivatives of order r of f at $\boldsymbol{x}$, arranged so that we can formally write

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

Thus we write the rth derivative of f as a vector of length d^r, and not as an r-fold tensor array or as a matrix. Moreover, if $f\colon \mathbb{R}^d \to \mathbb{R}^p$ is a vector function of a vector variable, with components $f = (f_1, \ldots, f_p)$ then we define $\mathsf{D}^{\otimes r} f(\boldsymbol{x}) \in \mathbb{R}^{pd^r}$ to be

$$\mathsf{D}^{\otimes r} f(\boldsymbol{x}) = \begin{pmatrix} \mathsf{D}^{\otimes r} f_1(\boldsymbol{x}) \\ \hline \vdots \\ \hline \mathsf{D}^{\otimes r} f_p(\boldsymbol{x}) \end{pmatrix}.$$

Notice that, using this notation, we have $\mathsf{D}(\mathsf{D}^{\otimes r}f) = \mathsf{D}^{\otimes(r+1)}f$. Also, the gradient of f is just $\mathsf{D}^{\otimes 1}f$ and the Hessian $\mathsf{H}f = \partial^2 f/(\partial \boldsymbol{x}\partial \boldsymbol{x}^T)$ is such that $\operatorname{vec}\mathsf{H}f = \mathsf{D}^{\otimes 2}f$, where vec denotes the vector operator (see Henderson and Searle (1979)). This vectorization carries some redundancy, for example, $\mathsf{D}^{\otimes 2}f$ contains repeated mixed partial derivatives whereas the usual vectorization $\operatorname{vech}\mathsf{H}f$ contains only the unique second order partial derivatives, with vech denoting the vector half operator (see Henderson and Searle (1979)). The latter is usually preferred since it is minimal and its matrix analysis is not more complicated than the former. However for $r > 2$, it appears that the matrix analysis using a generalization of the vector half operator becomes intractable. Other authors have used the same vectorization we propose to develop results for higher order derivatives: Holmquist (1996a) computes derivatives of the multivariate normal density function and Chacón and Duong (2008) compute kernel estimators of multivariate integrated density derivative functionals.

Suppose now that $f\colon \mathbb{R}^d \to \mathbb{R}$ is a density and we want to estimate $\mathsf{D}^{\otimes r}f(\boldsymbol{x})$ for some $r \in \mathbb{N}$. To this end, we use the kernel estimator $\widehat{\mathsf{D}^{\otimes r}f}(\boldsymbol{x};\mathbf{H}) = \mathsf{D}^{\otimes r}\hat{f}(\boldsymbol{x};\mathbf{H})$, where given a random sample $\mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_d$ drawn from f,

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

denotes the multivariate kernel density estimator with kernel K and bandwidth matrix $\mathbf{H}$, with $K_{\mathbf{H}}(\boldsymbol{x}) = |\mathbf{H}|^{-1/2}K(\mathbf{H}^{-1/2}\boldsymbol{x})$. The conditions on K and $\mathbf{H}$ will be given later. Notice that alternative expressions for $\widehat{\mathsf{D}^{\otimes r}f}(\boldsymbol{x};\mathbf{H})$ are

$$\begin{aligned}\widehat{\mathsf{D}^{\otimes r}f}(\boldsymbol{x};\mathbf{H}) &= n^{-1}\sum_{i=1}^{n} \mathsf{D}^{\otimes r}K_{\mathbf{H}}(\boldsymbol{x} - \mathbf{X}_i) \qquad (2)\\ &= n^{-1}(\mathbf{H}^{-1/2})^{\otimes r}\sum_{i=1}^{n}(\mathsf{D}^{\otimes r}K)_{\mathbf{H}}(\boldsymbol{x} - \mathbf{X}_i)\\ &= n^{-1}|\mathbf{H}|^{-1/2}(\mathbf{H}^{-1/2})^{\otimes r}\sum_{i=1}^{n}\mathsf{D}^{\otimes r}K\big(\mathbf{H}^{-1/2}(\boldsymbol{x} - \mathbf{X}_i)\big).\end{aligned}$$

The last expression in the previous display is quite helpful for implementing the estimator, because it separates the roles of K and $\mathbf{H}$. This is the multivariate generalization of the kernel estimator which appears, for instance, in Härdle, Marron and Wand (1990) and Jones (1994).

Here, the most general (unconstrained) form of the bandwidth matrix is used. In contrast, earlier papers considered this type of kernel estimator, but with constrained parametrizations of the bandwidth matrix, e.g. Härdle, Marron and Wand (1990) used a parametrization where $\mathbf{H}$ is h^2 multiplied by the identity matrix, and Singh (1976) used $\mathbf{H} = \operatorname{diag}(h_1^2, h_2^2, \ldots, h_d^2)$. However, in the case $r = 0$ (estimation of the density itself), Wand and Jones (1993) provide examples that show that a significant gain may be achieved by using unconstrained parametrizations over constrained ones; see also Chacón (2009). We generalize these results to arbitrary derivatives in Section 4.

We measure the error of the kernel density derivative estimator at the point $\boldsymbol{x}$ by using the mean squared error (MSE), defined as

$$\mathrm{MSE}(\boldsymbol{x};\mathbf{H}) \equiv \mathrm{MSE}\{\widehat{\mathsf{D}^{\otimes r}f}(\boldsymbol{x};\mathbf{H})\} = \mathbb{E}\|\widehat{\mathsf{D}^{\otimes r}f}(\boldsymbol{x};\mathbf{H}) - \mathsf{D}^{\otimes r}f(\boldsymbol{x})\|^2,$$

with $\|\cdot\|$ standing for the Euclidean norm; that is, $\|\boldsymbol{v}\|^2 = \boldsymbol{v}^T\boldsymbol{v} = \mathrm{tr}(\boldsymbol{v}\boldsymbol{v}^T)$ for a vector $\boldsymbol{v}$, where $\mathrm{tr}\,\mathbf{A}$ is short for the trace of a matrix $\mathbf{A}$. We have the two forms $\mathrm{MSE}(\boldsymbol{x};\mathbf{H})$ and $\mathrm{MSE}\{\widehat{\mathsf{D}^{\otimes r}f}(\boldsymbol{x};\mathbf{H})\}$, depending on whether we wish to suppress the explicit dependence on $\widehat{\mathsf{D}^{\otimes r}f}$ or not. It is easy to check that we can split $\mathrm{MSE}(\boldsymbol{x};\mathbf{H}) = \mathrm{B}^2(\boldsymbol{x};\mathbf{H}) + \mathrm{V}(\boldsymbol{x};\mathbf{H})$, where

$$\begin{aligned} \mathrm{B}^2(\boldsymbol{x};\mathbf{H}) &= \|\mathbb{E}\widehat{\mathsf{D}^{\otimes r}f}(\boldsymbol{x};\mathbf{H}) - \mathsf{D}^{\otimes r}f(\boldsymbol{x})\|^2 \\ \mathrm{V}(\boldsymbol{x};\mathbf{H}) &= \mathbb{E}\|\widehat{\mathsf{D}^{\otimes r}f}(\boldsymbol{x};\mathbf{H}) - \mathbb{E}\widehat{\mathsf{D}^{\otimes r}f}(\boldsymbol{x};\mathbf{H})\|^2 \end{aligned}$$

Analogously, as a global measure of the performance of the estimator we will use the mean integrated squared error, defined as

$$\mathrm{MISE}(\mathbf{H}) \equiv \mathrm{MISE}\{\widehat{\mathsf{D}^{\otimes r}f}(\cdot;\mathbf{H})\} = \int \mathrm{MSE}\{\widehat{\mathsf{D}^{\otimes r}f}(\boldsymbol{x};\mathbf{H})\}d\boldsymbol{x}.$$

All the results in this paper rely on the following assumptions on the bandwidth matrix, the density function and the kernel function:

(A1) $\mathbf{H}$ is a bandwidth matrix which is a symmetric and positive-definite matrix, and such that every element of $\mathbf{H} \to 0$ and $n^{-1}|\mathbf{H}|^{-1/2}(\mathbf{H}^{-1})^{\otimes r} \to 0$ as $n \to \infty$.

(A2) f is a density function for which all partial derivatives up to order $(r+2)$ inclusive exist, all its partial derivatives of order r are square integrable, and all its partial derivatives of order $(r+2)$ are bounded, continuous and square integrable.

(A3) K is a kernel which is a positive, symmetric, square integrable density function such that $\int \boldsymbol{x}\boldsymbol{x}^T K(\boldsymbol{x})d\boldsymbol{x} = m_2(K)\mathbf{I}_d$ for some real number $m_2(K)$ and $\mathbf{I}_d$ is the identity matrix of order d, and all its partial derivatives of order r are square integrable.

This is not the minimal set of assumptions but it provides a useful starting point for quantifying the following squared error results. We leave it to future research to reduce these assumptions.

For any vector function $g\colon \mathbb{R}^d \to \mathbb{R}^p$ denote

$$\mathbf{R}(g) = \int g(\boldsymbol{x})g(\boldsymbol{x})^T d\boldsymbol{x} \in \mathcal{M}_{p\times p}.$$

Also, for an arbitrary kernel L we write

$$\mathbf{R}_{L,\mathbf{H},r}(f) = \int L_{\mathbf{H}} * \mathsf{D}^{\otimes r}f(\boldsymbol{x})\mathsf{D}^{\otimes r}f(\boldsymbol{x})^T d\boldsymbol{x} \in \mathcal{M}_{d^r\times d^r},$$

where the convolution operator is applied to each component of $\mathsf{D}^{\otimes r} f(\boldsymbol{x})$ separately. Expanding all the terms in the bias-variance decomposition we obtain the following exact representation of the MISE function.

Theorem 1. *Assume that (A1)–(A3) hold. The MISE function can be expressed as*

$$\begin{aligned}\mathrm{MISE}\{\widehat{\mathsf{D}^{\otimes r} f}(\cdot;\mathbf{H})\} &= \operatorname{tr}\mathbf{R}(\mathsf{D}^{\otimes r} f) + n^{-1}|\mathbf{H}|^{-1/2}\operatorname{tr}\big((\mathbf{H}^{-1})^{\otimes r}\mathbf{R}(\mathsf{D}^{\otimes r}K)\big)\\ &\quad + (1-n^{-1})\operatorname{tr}\mathbf{R}_{K*K,\mathbf{H},r}(f) - 2\operatorname{tr}\mathbf{R}_{K,\mathbf{H},r}(f).\end{aligned}$$

The proof of this, along with the proofs for all other theorems in this paper, are deferred to the Appendix.

The form of the MISE given in the above theorem involves a complicated dependence on the bandwidth matrix $\mathbf{H}$ via the $\mathbf{R}$ functionals. To show this dependence more clearly, we search for a more mathematically tractable form of the MISE. The next result provides an asymptotic representation of the MISE function, and can be viewed as an extension of Formula (4.10) in Wand and Jones (1995, p. 98), which corresponds to the case $r = 0$.

Theorem 2. *Assume that (A1)–(A3) hold. We can expand*

$$\mathrm{MISE}\{\widehat{\mathsf{D}^{\otimes r} f}(\cdot;\mathbf{H})\} = \mathrm{AMISE}\{\widehat{\mathsf{D}^{\otimes r} f}(\cdot;\mathbf{H})\} + o\big(n^{-1}|\mathbf{H}|^{-1/2}\operatorname{tr}^r(\mathbf{H}^{-1}) + \operatorname{tr}^2\mathbf{H}\big),$$

where

$$\begin{aligned}\mathrm{AMISE}\{\widehat{\mathsf{D}^{\otimes r} f}(\cdot;\mathbf{H})\} &= n^{-1}|\mathbf{H}|^{-1/2}\operatorname{tr}\big((\mathbf{H}^{-1})^{\otimes r}\mathbf{R}(\mathsf{D}^{\otimes r}K)\big)\\ &\quad + \tfrac{m_2(K)^2}{4}\operatorname{tr}\big[(\mathbf{I}_{d^r}\otimes\operatorname{vec}^T\mathbf{H})\mathbf{R}(\mathsf{D}^{\otimes(r+2)}f)(\mathbf{I}_{d^r}\otimes\operatorname{vec}\mathbf{H})\big].\end{aligned}$$

The AMISE-optimal bandwidth matrix $\mathbf{H}_{\mathrm{AMISE}}$ is defined to be the matrix, amongst all symmetric positive definite matrices, which minimizes $\mathrm{AMISE}\{\widehat{\mathsf{D}^{\otimes r} f}(\cdot;\mathbf{H})\}$. Next we give the order of such a matrix, and the order of the resulting minimal AMISE.

Theorem 3. *Assume that (A1)–(A3) hold. Every entry of the optimal bandwidth matrix $\mathbf{H}_{\mathrm{AMISE}}$ is of order $O(n^{-2/(d+2r+4)})$. As a consequence, the minimal achievable AMISE, $\min_{\mathbf{H}}\mathrm{AMISE}\{\widehat{\mathsf{D}^{\otimes r} f}(\cdot;\mathbf{H})\}$, is of order $O(n^{-4/(d+2r+4)})$.*

Theorems 1, 2 and 3 generalize the existing MISE, AMISE and $\mathbf{H}_{\mathrm{AMISE}}$ results for constrained to unconstrained bandwidth matrices. These rates of convergence reveal the relationship with dimension d and derivative order r. As either of these increase, the minimum achievable of the AMISE increases. Though we note that, at least asymptotically, the marginal increase in difficulty in estimating a derivative an order higher is the same as estimating a density two dimensions higher.

To appreciate the ramifications of these three theorems, we briefly review the literature for kernel density derivative estimation. Bhattacharya (1967) shows that the rate of convergence in probability of a univariate kernel density derivative estimator with a second order kernel is bounded by $n^{-1/(2r+4)}$, without developing intermediate squared error convergence results. Schuster (1969) establishes the same rate for a wider class of kernels. Singh (1979, 1987) establish, for his specially constructed univariate kernel estimator, that $n^{-2(p-r)/(2p+1)}$ is the MSE and MISE rate of convergence respectively. This estimator employs kernels whose rth moment is 1 and all other jth moments are zero, $j = 0, 1, \dots, p-1$, $j \neq r$ for $p > r$. The order of these kernels is greater than or equal to r. Since we assume second order kernels (assumption (A3)), then Theorem 3 will not generalize this result for $r > 2$. For second order kernels, Wand and Jones (1995, p. 49) show that $p = r + 2$ which gives the MISE to be $O(n^{-4/(2r+5)})$, which is indeed Theorem 3 with $d = 1$.

For d-variate density derivative estimation, Stone (1980) states that for any linear nonparametric estimator, the minimum achievable MSE of an estimator of g, a scalar functional of $\mathsf{D}^{\otimes r} f$, is $O(n^{-2(p-r)/(2p+d)})$ where $p > r$ and the pth order derivative of g is bounded. From assumption (A2), we have that $p = r + 2$ so the MISE rate is $O(n^{-4/(d+2r+4)})$ which is exactly the same rate as Theorem 3. However this general result is unable to elucidate certain key questions specific to kernel estimators, such as the relationship between the convergence rate and the bandwidth matrix, which Theorems 1, 2 and 3 demonstrate clearly. Singh (1976) shows that a multivariate kernel density derivative estimator with a diagonal bandwidth matrix $\mathbf{H} = \operatorname{diag}(h_1^2, h_2^2, \dots, h_d^2)$ is mean square convergent, though he is not able to quantify the rate of convergence. More recently Duong, Cowling, Koch and Wand (2008) establish that the MISE convergence rate for kernel estimators with unconstrained bandwidth matrices, is $n^{-4/(d+6)}$ and $n^{-4/(d+8)}$ for $r = 1, 2$. Theorem 3 extends these two special cases to arbitrary r.

So far we have only considered scalar functionals of the expected value and the variance of $\widehat{\mathsf{D}^{\otimes r} f}$. For completeness, the following theorem gives these quantities in their vector and matrix form. This theorem is a generalization of the results obtained by Duong, Cowling, Koch and Wand (2008) for $r = 1, 2$ to arbitrary r.

Theorem 4. *Assume (A1)–(A3) hold. The expected value of* $\widehat{\mathsf{D}^{\otimes r} f}(\boldsymbol{x}; \mathbf{H})$ *is*

$$\mathbb{E}\, \widehat{\mathsf{D}^{\otimes r} f}(\boldsymbol{x}; \mathbf{H}) = \mathsf{D}^{\otimes r} f(\boldsymbol{x}) + \tfrac{1}{2} m_2(K)(\mathbf{I}_{d^r} \otimes \operatorname{vec}^T \mathbf{H}) \mathsf{D}^{\otimes (r+2)} f(\boldsymbol{x}) + O(\operatorname{tr}^2 \mathbf{H}) \mathbf{1}_{d^r}$$

and the variance is

$$\operatorname{Var} \widehat{\mathsf{D}^{\otimes r} f}(\boldsymbol{x}; \mathbf{H}) = n^{-1} |\mathbf{H}|^{-1/2} (\mathbf{H}^{-1/2})^{\otimes r} \mathbf{R}(\mathsf{D}^{\otimes r} K)(\mathbf{H}^{-1/2})^{\otimes r} f(\boldsymbol{x}) + o(n^{-1} |\mathbf{H}|^{-1/2} \operatorname{tr}^r \mathbf{H}^{-1}) \mathbf{J}_{d^r}$$

where the elements of $\mathbf{1}_p \in \mathbb{R}^p$ *and* $\mathbf{J}_p \in \mathcal{M}_{p \times p}$ *are all ones.*

3 Normal mixture densities

In this section we study in detail the problem in the normal mixture case. We start with a single normal density, when $K = \phi$ with ϕ the density of the standard d-variate normal distribution, $\phi(\boldsymbol{x}) = (2\pi)^{-d/2}\exp(-\boldsymbol{x}^T\boldsymbol{x}/2)$, and when $f = \phi_{\boldsymbol{\Sigma}}(\cdot - \boldsymbol{\mu})$ a normal density with mean $\boldsymbol{\mu}$ and variance $\boldsymbol{\Sigma}$.

The MISE and AMISE expressions in the normal case are closely related to the moments of quadratic forms in normal variables. Given two symmetric matrices $\mathbf{A}$ and $\mathbf{B}$ in $\mathcal{M}_{d\times d}$, we will write

$$\mu_{r,s}(\mathbf{A},\mathbf{B}) \equiv \mathbb{E}[(\boldsymbol{z}^T\mathbf{A}^{-1}\boldsymbol{z})^r(\boldsymbol{z}^T\mathbf{B}^{-1}\boldsymbol{z})^s] \quad \text{and} \quad \mu_r(\mathbf{A}) \equiv \mu_{r,0}(\mathbf{A},\mathbf{I})$$

where $\boldsymbol{z}$ is a d-variate random vector with standard normal distribution.

Theorem 5. *Assume that (A1) holds. Further assume that f is a normal density with mean $\boldsymbol{\mu}$ and variance $\boldsymbol{\Sigma}$; and that K is the normal kernel. The MISE function admits the explicit expression*

$$\begin{aligned}\mathrm{MISE}\{\widehat{\mathsf{D}^{\otimes r}f}(\cdot;\mathbf{H})\} = 2^{-(d+r)}\pi^{-d/2}\big\{&|\boldsymbol{\Sigma}|^{-1/2}\mu_r(\boldsymbol{\Sigma}) + n^{-1}|\mathbf{H}|^{-1/2}\mu_r(\mathbf{H}) \\ &+ (1-n^{-1})|\mathbf{H}+\boldsymbol{\Sigma}|^{-1/2}\mu_r(\mathbf{H}+\boldsymbol{\Sigma}) - 2^{(d+2r+2)/2}|\mathbf{H}+2\boldsymbol{\Sigma}|^{-1/2}\mu_r(\mathbf{H}+2\boldsymbol{\Sigma})\big\}.\end{aligned}$$

As expected, an explicit form of the minimizer of the MISE is not available.

To rewrite Theorem 5 without the μ_r functionals, we use Theorem 1 in Holmquist (1996b) which shows that

$$\mu_r(\mathbf{A}) = \mathrm{OF}(2r)(\mathrm{vec}^T\,\mathbf{A}^{-1})^{\otimes r}\boldsymbol{\mathcal{S}}_{d,2r}(\mathrm{vec}\,\mathbf{I}_d)^{\otimes r}$$

where $\mathrm{OF}(2r) = (2r-1)(2r-3)\cdots 5\cdot 3\cdot 1$ denotes the odd factorial and $\boldsymbol{\mathcal{S}}_{m,n} \in \mathcal{M}_{m^n\times m^n}$ is the symmetrizer matrix of order m, n; see Holmquist (1996a,b). These references contain long technical definitions of the symmetrizer matrix, which we do not reproduce here. Instead we focus on the action of the symmetrizer matrix on Kronecker products of vectors. Let $\boldsymbol{x}_i \in \mathbb{R}^p, i = 1,\ldots n$, and $\mathbf{X}^* = \{\boldsymbol{x}_1^*,\ldots,\boldsymbol{x}_n^*\}$ be a permutation of $\{\boldsymbol{x}_1,\ldots\boldsymbol{x}_n\}$. The symmetrizer matrix maps the product $\bigotimes_{i=1}^n \boldsymbol{x}_i$ to a linear combination of products of all possible permutations of $\boldsymbol{x}_1,\ldots,\boldsymbol{x}_n$

$$\boldsymbol{\mathcal{S}}_{m,n}\Big(\bigotimes_{i=1}^n \boldsymbol{x}_i\Big) = \frac{1}{n!}\sum_{\text{all }\mathbf{X}^*}\bigotimes_{i=1}^n \boldsymbol{x}_i^*.$$

More explicitly for a 3-fold product, $\boldsymbol{\mathcal{S}}_{m,3}(\boldsymbol{x}_1\otimes\boldsymbol{x}_2\otimes\boldsymbol{x}_3) = \frac{1}{6}[\boldsymbol{x}_1\otimes\boldsymbol{x}_2\otimes\boldsymbol{x}_3 + \boldsymbol{x}_1\otimes\boldsymbol{x}_3\otimes\boldsymbol{x}_2 + \boldsymbol{x}_2\otimes\boldsymbol{x}_1\otimes\boldsymbol{x}_3 + \boldsymbol{x}_2\otimes\boldsymbol{x}_3\otimes\boldsymbol{x}_1 + \boldsymbol{x}_3\otimes\boldsymbol{x}_1\otimes\boldsymbol{x}_2 + \boldsymbol{x}_3\otimes\boldsymbol{x}_2\otimes\boldsymbol{x}_1]$.

Corollary 1. *Under the conditions of Theorem 5, the MISE function admits the explicit expression*

$$\begin{aligned}\mathrm{MISE}\{\widehat{\mathsf{D}^{\otimes r}f}(\cdot;\mathbf{H})\} = 2^{-(d+r)}\pi^{-d/2}\mathrm{OF}(2r)\big\{&|\boldsymbol{\Sigma}|^{-1/2}(\mathrm{vec}^T\,\boldsymbol{\Sigma}^{-1})^{\otimes r} + |\mathbf{H}|^{-1/2}(\mathrm{vec}^T\,\mathbf{H}^{-1})^{\otimes r} \\ &+ (1-n^{-1})|\mathbf{H}+\boldsymbol{\Sigma}|^{-1/2}(\mathrm{vec}^T(\mathbf{H}+\boldsymbol{\Sigma})^{-1})^{\otimes r} \\ &- 2^{(d+2r+2)/2}|\mathbf{H}+2\boldsymbol{\Sigma}|^{-1/2}(\mathrm{vec}^T(\mathbf{H}+2\boldsymbol{\Sigma})^{-1})^{\otimes r}\big\}\boldsymbol{\mathcal{S}}_{d,2r}(\mathrm{vec}\,\mathbf{I}_d)^{\otimes r}.\end{aligned}$$

This corollary has the advantage of showing the explicit dependence of the MISE on the bandwidth matrix $\mathbf{H}$. However, the direct computation of $\boldsymbol{\mathcal{S}}_{d,2r} \in \mathcal{M}_{d^{2r}\times d^{2r}}$ may be an onerous task; for example, for $d = r = 4$, $\boldsymbol{\mathcal{S}}_{4,8}$ is a 65536×65536 matrix. In contrast, although Theorem 5 does not express the explicit dependence of the MISE on $\mathbf{H}$ due to the use of the $\boldsymbol{\mu}_r$ functionals, formula (6) in Holmquist (1996b) gives the computationally efficient recursive expression

$$\mu_r(\mathbf{A}) = (r-1)!\, 2^{r-1} \sum_{j=0}^{r-1} \frac{\operatorname{tr}\left(\mathbf{A}^{-(r-j)}\right)}{j!\, 2^j} \mu_j(\mathbf{A}). \tag{3}$$

Here, we understand $\mathbf{A}^{-p} = (\mathbf{A}^{-1})^p$ for $p > 0$ and, consequently, $\mathbf{A}^0 = \mathbf{I}_d$.

An analogous expression of the AMISE is given below. In this case, we can also write down an explicit expression for its minimizer.

Theorem 6. *Under the conditions of Theorem 5, the AMISE function admits the explicit expression*

$$\mathrm{AMISE}\{\widehat{\mathsf{D}^{\otimes r} f}(\cdot;\mathbf{H})\} = 2^{-(d+r)}\pi^{-d/2}\big\{n^{-1}|\mathbf{H}|^{-1/2}\mu_r(\mathbf{H}) + \tfrac{1}{16}|\mathbf{\Sigma}|^{-1/2}\mu_{r,2}(\mathbf{\Sigma}, \mathbf{\Sigma}^{1/2}\mathbf{H}^{-1}\mathbf{\Sigma}^{1/2})\big\}.$$

The value of $\mathbf{H}$ *that minimizes this function is*

$$\mathbf{H}_{\mathrm{AMISE}} = \left(\frac{4}{d+2r+2}\right)^{2/(d+2r+4)} \mathbf{\Sigma} n^{-2/(d+2r+4)}.$$

The AMISE expression in Theorem 6 can be rewritten without μ_r functionals. From Theorem 5 in Holmquist (1996b),

$$\mu_{r,s}(\mathbf{A},\mathbf{B}) = \mathrm{OF}(2r+2s)[(\operatorname{vec}^T \mathbf{A}^{-1})^{\otimes r} \otimes (\operatorname{vec}^T \mathbf{B}^{-1})^{\otimes s}]\boldsymbol{\mathcal{S}}_{d,2r+2s}(\operatorname{vec}\mathbf{I}_d)^{\otimes(r+s)}.$$

Corollary 2. *Under the conditions of Theorem 5, the AMISE function admits the explicit expression*

$$\begin{aligned}\mathrm{AMISE}\{\widehat{\mathsf{D}^{\otimes r} f}(\cdot;\mathbf{H})\} = 2^{-(d+r)}\pi^{-d/2}\big\{&\mathrm{OF}(2r)n^{-1}|\mathbf{H}|^{-1/2}(\operatorname{vec}^T \mathbf{H}^{-1})^{\otimes r}\boldsymbol{\mathcal{S}}_{d,2r}(\operatorname{vec}\mathbf{I}_d)^{\otimes r} \\ &+ \tfrac{1}{16}|\mathbf{\Sigma}|^{-1/2}\mathrm{OF}(2r+4)[(\operatorname{vec}^T \mathbf{\Sigma}^{-1})^{\otimes r} \otimes \{\operatorname{vec}^T(\mathbf{\Sigma}^{-1/2}\mathbf{H}\mathbf{\Sigma}^{-1/2})\}^{\otimes 2}] \\ &\times \boldsymbol{\mathcal{S}}_{d,2r+4}(\operatorname{vec}\mathbf{I}_d)^{\otimes(r+2)}\big\}.\end{aligned}$$

To facilitate the comparison of the extra amount of smoothing induced for higher dimensions and higher derivatives for standard normal densities, we examine the ratio

$$\frac{h_{\mathrm{AMISE}}(d,r,n)}{h_{\mathrm{AMISE}}(1,0,n)} = \left(\frac{3}{4}\right)^{1/5} \left(\frac{4}{d+2r+2}\right)^{1/(d+2r+4)} n^{(d+2r-1)/(5d+10r+20)}$$

where $\mathbf{H}_{\mathrm{AMISE}}(d,r,n) = h^2_{\mathrm{AMISE}}(d,r,n)\mathbf{I}_d$ is the AMISE-optimal bandwidth for the standard normal density given d, r and n. These values are tabulated in Table 1 for $d = 1, 2, \ldots, 6$

		$d=1$	$d=2$	$d=3$	$d=4$	$d=5$	$d=6$
$n=1000$	$r=0$	1.00	1.19	1.36	1.51	1.64	1.76
	$r=1$	1.36	1.51	1.64	1.76	1.86	1.96
	$r=2$	1.64	1.76	1.86	1.96	2.04	2.12
	$r=3$	1.86	1.96	2.04	2.12	2.19	2.26
$n=10\ 000$	$r=0$	1.00	1.28	1.55	1.79	2.01	2.21
	$r=1$	1.55	1.79	2.01	2.21	2.40	2.56
	$r=2$	2.01	2.21	2.40	2.56	2.71	2.85
	$r=3$	2.40	2.56	2.71	2.85	2.98	3.10
$n=100\ 000$	$r=0$	1.00	1.39	1.77	2.13	2.47	2.79
	$r=1$	1.77	2.13	2.47	2.79	3.08	3.35
	$r=2$	2.47	2.79	3.08	3.35	3.60	3.84
	$r=3$	3.08	3.35	3.60	3.84	4.05	4.25

Table 1: Comparison of extra smoothing induced for higher dimensions and higher order derivatives for a standard normal density. Each table entry contains the value of $h_{\text{AMISE}}(d,r,n)/h_{\text{AMISE}}(1,0,n)$.

and $r = 0, 1, 2, 3$. These ratios are the same for different combinations of d and r whenever $d + 2r$ are the same.

The normal scale bandwidth selector is obtained by replacing the variance $\mathbf{\Sigma}$ in $\mathbf{H}_{\text{AMISE}}$ from Theorem 6 by an estimate $\hat{\mathbf{\Sigma}}$

$$\hat{\mathbf{H}}_{\text{NS}} = \left(\frac{4}{d+2r+2}\right)^{2/(d+2r+4)} \hat{\mathbf{\Sigma}} n^{-2/(d+2r+4)}. \tag{4}$$

We can use (4) as a starting point to develop consistent data-driven bandwidth matrices. Consistent univariate selectors for density derivatives include the unbiased cross validation selector of Härdle, Marron and Wand (1990), and the selector of Wu (1997) and Wu and Lin (2000). The performance of the multivariate versions of these selectors is yet to be established and we do not pursue this further in this paper.

We now consider general normal mixture densities $f(\boldsymbol{x}) = \sum_{\ell=1}^k w_\ell \phi_{\mathbf{\Sigma}_\ell}(\boldsymbol{x} - \boldsymbol{\mu}_\ell)$ where $w_\ell > 0$ and $\sum_{\ell=1}^k w_\ell = 1$. Normal mixture densities are widely employed in simulation studies since they provide a rich class of densities with tractable exact squared error expressions. Normal mixture densities were used in early attempts for data-based bandwidth selection for multivariate kernel density estimation, see Ćwik and Koronacki (1997). They proposed the following 2-step procedure: (i) a preliminary normal mixture is fitted to the data and (ii) the MISE- and AMISE-optimal bandwidths are computed from the closed form expressions for the MISE and AMISE for this normal mixture fit. We provide MISE and AMISE expressions for density derivatives to provide the basis for an analogous selector.

Theorem 7. *Assume (A1) holds. Further assume that f is the normal mixture density $\sum_{\ell=1}^{k} w_\ell \phi_{\boldsymbol{\Sigma}_\ell}(\cdot - \boldsymbol{\mu}_\ell)$ and that K is the normal kernel. The MISE function admits the explicit expression*

$$\mathrm{MISE}\{\widehat{\mathsf{D}^{\otimes r} f}(\cdot;\mathbf{H})\} = 2^{-r} n^{-1} (4\pi)^{-d/2} |\mathbf{H}|^{-1/2} \mu_r(\mathbf{H}) + \boldsymbol{w}^T \{(1-n^{-1})\boldsymbol{\Omega}_2 - 2\boldsymbol{\Omega}_1 + \boldsymbol{\Omega}_0\}\boldsymbol{w},$$

where $\boldsymbol{w} = (w_1, w_2, \dots, w_k)^T$ and $\boldsymbol{\Omega}_a \in \mathcal{M}_{k\times k}$ whose (ℓ,ℓ') entry is given by

$$(\boldsymbol{\Omega}_a)_{\ell,\ell'} = (-1)^r (\mathrm{vec}^T \mathbf{I}_{d^r}) \mathsf{D}^{\otimes 2r} \phi_{a\mathbf{H}+\boldsymbol{\Sigma}_{\ell\ell'}}(\boldsymbol{\mu}_{\ell\ell'})$$

with $\boldsymbol{\mu}_{\ell\ell'} = \boldsymbol{\mu}_\ell - \boldsymbol{\mu}_{\ell'}$, $\boldsymbol{\Sigma}_{\ell\ell'} = \boldsymbol{\Sigma}_\ell + \boldsymbol{\Sigma}_{\ell'}$. An equivalent expression of the MISE is

$$\begin{aligned}\mathrm{MISE}\{\widehat{\mathsf{D}^{\otimes r} f}(\cdot;\mathbf{H})\} &= 2^{-r} \mathrm{OF}(2r) n^{-1} (4\pi)^{-d/2} |\mathbf{H}|^{-1/2} (\mathrm{vec}^T \mathbf{H}^{-1})^{\otimes r} \boldsymbol{\mathcal{S}}_{d,2r} (\mathrm{vec}\, \mathbf{I}_d)^{\otimes r} \\ &\quad + \boldsymbol{w}^T \{(1-n^{-1})\boldsymbol{\Omega}_2 - 2\boldsymbol{\Omega}_1 + \boldsymbol{\Omega}_0\}\boldsymbol{w},\end{aligned}$$

where

$$\begin{aligned}(\boldsymbol{\Omega}_a)_{\ell,\ell'} &= (-1)^r \phi_{a\mathbf{H}+\boldsymbol{\Sigma}_{\ell\ell'}}(\boldsymbol{\mu}_{\ell\ell'}) (\mathrm{vec}^T (a\mathbf{H}+\boldsymbol{\Sigma}_{\ell\ell'})^{-2})^{\otimes r} \boldsymbol{\mathcal{S}}_{d,2r} \\ &\quad \times \sum_{j=0}^{r} (-1)^j \mathrm{OF}(2j) \binom{2r}{2j} \big[\boldsymbol{\mu}_{\ell\ell'}^{\otimes(2r-2j)} \otimes (\mathrm{vec}(a\mathbf{H}+\boldsymbol{\Sigma}_{\ell\ell'}))^{\otimes j}\big].\end{aligned}$$

Theorem 7 is the analogue of Theorem 1 in Wand and Jones (1993). The following AMISE formula is the analogue of Theorem 1 in Wand (1992).

Theorem 8. *Under the conditions of Theorem 7, the AMISE function admits the explicit expression*

$$\mathrm{AMISE}\{\widehat{\mathsf{D}^{\otimes r} f}(\cdot;\mathbf{H})\} = 2^{-r} n^{-1} (4\pi)^{-d/2} |\mathbf{H}|^{-1/2} \mu_r(\mathbf{H}) + \tfrac{1}{4} \boldsymbol{w}^T \widetilde{\boldsymbol{\Omega}} \boldsymbol{w}$$

where $\widetilde{\boldsymbol{\Omega}} \in \mathcal{M}_{k\times k}$ whose (ℓ,ℓ') entry is given by

$$\widetilde{\boldsymbol{\Omega}}_{\ell,\ell'} = (-1)^r \, \mathrm{vec}^T \big(\mathbf{I}_{d^r} \otimes (\mathrm{vec}\, \mathbf{H}\, \mathrm{vec}^T \mathbf{H})\big) \mathsf{D}^{\otimes 2r+4} \phi_{\boldsymbol{\Sigma}_{\ell\ell'}}(\boldsymbol{\mu}_{\ell\ell'})$$

with $\boldsymbol{\mu}_{\ell\ell'} = \boldsymbol{\mu}_\ell - \boldsymbol{\mu}_{\ell'}$, $\boldsymbol{\Sigma}_{\ell\ell'} = \boldsymbol{\Sigma}_\ell + \boldsymbol{\Sigma}_{\ell'}$. An equivalent expression for the AMISE is

$$\mathrm{AMISE}\{\widehat{\mathsf{D}^{\otimes r} f}(\cdot;\mathbf{H})\} = 2^{-r} \mathrm{OF}(2r) n^{-1} (4\pi)^{-d/2} |\mathbf{H}|^{-1/2} (\mathrm{vec}^T \mathbf{H}^{-1})^{\otimes r} \boldsymbol{\mathcal{S}}_{d,2r} (\mathrm{vec}\, \mathbf{I}_d)^{\otimes r} + \tfrac{1}{4} \boldsymbol{w}^T \widetilde{\boldsymbol{\Omega}} \boldsymbol{w}$$

where

$$\begin{aligned}\widetilde{\boldsymbol{\Omega}}_{\ell,\ell'} &= (-1)^r \phi_{\boldsymbol{\Sigma}_{\ell\ell'}}(\boldsymbol{\mu}_{\ell\ell'}) \big[(\mathrm{vec}^T \boldsymbol{\Sigma}_{\ell\ell'}^{-2})^{\otimes r} \otimes (\mathrm{vec}^T (\boldsymbol{\Sigma}_{\ell\ell'}^{-1} \mathbf{H} \boldsymbol{\Sigma}_{\ell\ell'}^{-1}))^{\otimes 2}\big] \\ &\quad \times \boldsymbol{\mathcal{S}}_{d,2r+4} \sum_{j=0}^{r+2} (-1)^j \mathrm{OF}(2j) \binom{2r+4}{2j} \big[\boldsymbol{\mu}_{\ell\ell'}^{\otimes(2r-2j+4)} \otimes (\mathrm{vec}\, \boldsymbol{\Sigma}_{\ell\ell'})^{\otimes j}\big].\end{aligned}$$

For $r = 0$, Theorem 8 gives $\mathrm{AMISE}\{\hat{f}(\cdot;\mathbf{H})\} = n^{-1}(4\pi)^{-d/2}|\mathbf{H}|^{-1/2} + \frac{1}{4}\boldsymbol{w}^T \widetilde{\boldsymbol{\Omega}} \boldsymbol{w}$ with $\widetilde{\boldsymbol{\Omega}}_{\ell,\ell'} = \phi_{\boldsymbol{\Sigma}_{\ell\ell'}}(\boldsymbol{\mu}_{\ell\ell'}) (\mathrm{vec}^T(\boldsymbol{\Sigma}_{\ell\ell'}^{-1}\mathbf{H}\boldsymbol{\Sigma}_{\ell\ell'}^{-1}))^{\otimes 2} \boldsymbol{\mathcal{S}}_{d,4} [\boldsymbol{\mu}_{\ell\ell'}^{\otimes 4} - 6\boldsymbol{\mu}_{\ell\ell'}^{\otimes 2} \otimes \mathrm{vec}\, \boldsymbol{\Sigma}_{\ell\ell'} + 3(\mathrm{vec}\, \boldsymbol{\Sigma}_{\ell\ell'})^{\otimes 2}]$. This appears to be completely different to, even though it is equivalent to, Theorem 1 in Wand (1992). The expressions in Theorems 5 to 8 are implemented in the `ks` library in `R`.

4 Asymptotic relative efficiency

We examine the gain in density estimation performance when using the added flexibility of unconstrained selectors. The usual measure of asymptotic performance is the minimal achievable AMISE. We compare this minimal AMISE for these parametrization classes: $\mathcal{F}$ the class of all positive-definite matrices, $\mathcal{D}$ the class of all positive-definite diagonal matrices, and $\mathcal{I}$ the class of positive scalar multiples of the identity matrix. We consider f to be a single normal density to simplify the mathematical analysis.

Corollary 3. *Assume that the conditions of Theorem 5 hold. For the class of unconstrained bandwidth matrices $\mathcal{F}$, the minimal achievable AMISE is*

$$\begin{aligned}\min_{\mathbf{H}\in\mathcal{F}} \mathrm{AMISE}(\mathbf{H}) = {} & 2^{-(d+r+4)} 2^{8/(d+2r+4)} \pi^{-d/2} (d+2r+4)(d+2r+2)^{(d+2r)/(d+2r+4)} \\ & \times |\boldsymbol{\Sigma}|^{-1/2} \mu_r(\boldsymbol{\Sigma}) n^{-4/(d+2r+4)}.\end{aligned}$$

The asymptotic rate of convergence of the minimal achievable AMISE was previously stated in Theorem 3 for general f. This corollary provides its constants when f is a single normal density, generalizing the result from Wand and Jones (1995, p. 112) to general r.

Corollary 4. *Assume that the conditions of Theorem 5 hold. For the class $\mathcal{I}$, $\mathbf{H} = h^2\mathbf{I}_d$, the AMISE admits the expression*

$$\mathrm{AMISE}(\mathbf{H}) = 2^{-(d+r)} \pi^{-d/2} \{ n^{-1} h^{-d-2r} \mu_r(\mathbf{I}_d) + \tfrac{1}{16} |\boldsymbol{\Sigma}|^{-1/2} \mu_{r+2}(\boldsymbol{\Sigma}) h^4 \}.$$

The bandwidth which minimizes the AMISE is $\mathbf{H}_{\mathrm{AMISE}} = h^2_{\mathrm{AMISE}}\mathbf{I}_d$ where

$$h_{\mathrm{AMISE}} = \left(\frac{4(d+2r)|\boldsymbol{\Sigma}|^{1/2}\mu_r(\mathbf{I}_d)}{\mu_{r+2}(\boldsymbol{\Sigma}) n} \right)^{1/(d+2r+4)}.$$

The minimal achievable AMISE is

$$\begin{aligned}\min_{\mathbf{H}\in\mathcal{I}} \mathrm{AMISE}(\mathbf{H}) = {} & 2^{-(d+r+4)} 2^{8/(d+2r+4)} \pi^{-d/2} \big\{ |\boldsymbol{\Sigma}|^{-(d+2r)/2} \mu_{r+2}(\boldsymbol{\Sigma})^{d+2r} \mu_{r+1}(\mathbf{I}_d)^4 \big\}^{1/(d+2r+4)} \\ & \times (d+2r+4)(d+2r)^{-1} n^{-4/(d+2r+4)}.\end{aligned}$$

Comparing this corollary to the previous one, the rate of AMISE convergence does not depend on the parametrization class. The difference in finite sample performance is due to the different coefficients of the minimal AMISE. The gain in density estimation efficiency using an unconstrained bandwidth matrix over constrained bandwidths was established in the bivariate case by Wand and Jones (1993). Their main measure of this gain is the Asymptotic Relative Efficiency (ARE), e.g.,

$$\mathrm{ARE}(\mathcal{F} : \mathcal{D}) = \left[\min_{\mathbf{H}\in\mathcal{F}} \mathrm{AMISE}(\mathbf{H}) / \min_{\mathbf{H}\in\mathcal{D}} \mathrm{AMISE}(\mathbf{H}) \right]^{(d+2r+4)/4}$$

The interpretation of ARE($\mathcal{F}:\mathcal{D}$) is that, for large n, the minimum error using n observations with a diagonal bandwidth can be achieved using only ARE($\mathcal{F}:\mathcal{D}$) $\times\, n$ observations with an unconstrained $\mathbf{H}$. Analogous definitions and interpretations apply to ARE($\mathcal{F}:\mathcal{I}$) and ARE($\mathcal{D}:\mathcal{I}$).

Computing these AREs analytically for general densities is mathematically intractable, so we focus on the case where f is a normal density, making use of Corollaries 3 and 4.

Corollary 5. *Assume that the conditions of Theorem 5 hold. The asymptotic relative efficiency of $\mathcal{F}$ compared to $\mathcal{I}$ is*

$$\mathrm{ARE}(\mathcal{F}:\mathcal{I}) = [(d+2r+2)(d+2r)]^{(d+2r)/4}|\mathbf{\Sigma}|^{-1/2}\mu_r(\mathbf{\Sigma})^{(d+2r+4)/4}\mu_{r+2}(\mathbf{\Sigma})^{-(d+2r)/4}\mu_r(\mathbf{I}_d)^{-1}.$$

In the case of a bivariate normal density with both variances equal we are able to obtain explicit analytic expressions for the asymptotic relative efficiencies, for all $r \geq 0$, in terms of the correlation coefficient ρ.

Corollary 6. *Assume that the conditions of Theorem 5 hold. Suppose that f is a bivariate normal density function having variances equal and with correlation coefficient ρ. Then*

$$\mathrm{ARE}(\mathcal{F}:\mathcal{D}) = \mathrm{ARE}(\mathcal{F}:\mathcal{I}) = [4(r+2)(r+1)]^{(r+1)/2}\frac{(1-\rho^2)^{1/2}Q(r,\rho)^{(r+3)/2}}{Q(r,0)Q(r+2,\rho)^{(r+1)/2}}$$

where

$$Q(r,\rho) = \sum_{j=0}^{r}\sum_{j'=0}^{j}\binom{r}{j}\binom{j}{j'}(-2\rho)^{j-j'}m_{j+j'}m_{2r-j-j'}$$

and $m_k = \frac{1}{2}\{(-1)^k+1\}\mathrm{OF}(k)$ for $k = 0,1,2,\ldots$.

In Figure 1, we compare the AREs for four families of a single bivariate normal density with mean zero and marginal variances σ_1^2, σ_2^2 and correlation coefficient ρ ranging from -1 to $+1$. Without loss of generality, we let $\sigma_1 = 1$, and we let $\sigma_2 = 1, 2, 5, 10$. For each value of ρ, we compute ARE($\mathcal{F}:\mathcal{D}$) numerically and ARE($\mathcal{F}:\mathcal{I}$) analytically. The former are plotted as the black curves, and the latter as grey curves. We consider derivatives of order $r = 0, 1, \ldots, 4$ which are drawn in the solid, short dashed, dotted, dot-dashed and long dashed lines respectively. This figure generalizes the plots in Wand and Jones (1993). The most immediate trend from these plots is, apart from equal marginal variances $\sigma_1 = \sigma_2$ with weak correlation, ARE($\mathcal{F}:\mathcal{I}$) is close to zero, indicating that the class $\mathcal{I}$ is inadequate for moderate to high correlation. The other striking trend is that the rate that both AREs tend to zero, as $|\rho|$ tends to 1, increases as the derivative order increases. This indicates that the gain from unconstrained bandwidths for higher derivatives exceeds the known gain for $r = 0$ (Wand and Jones, 1993).

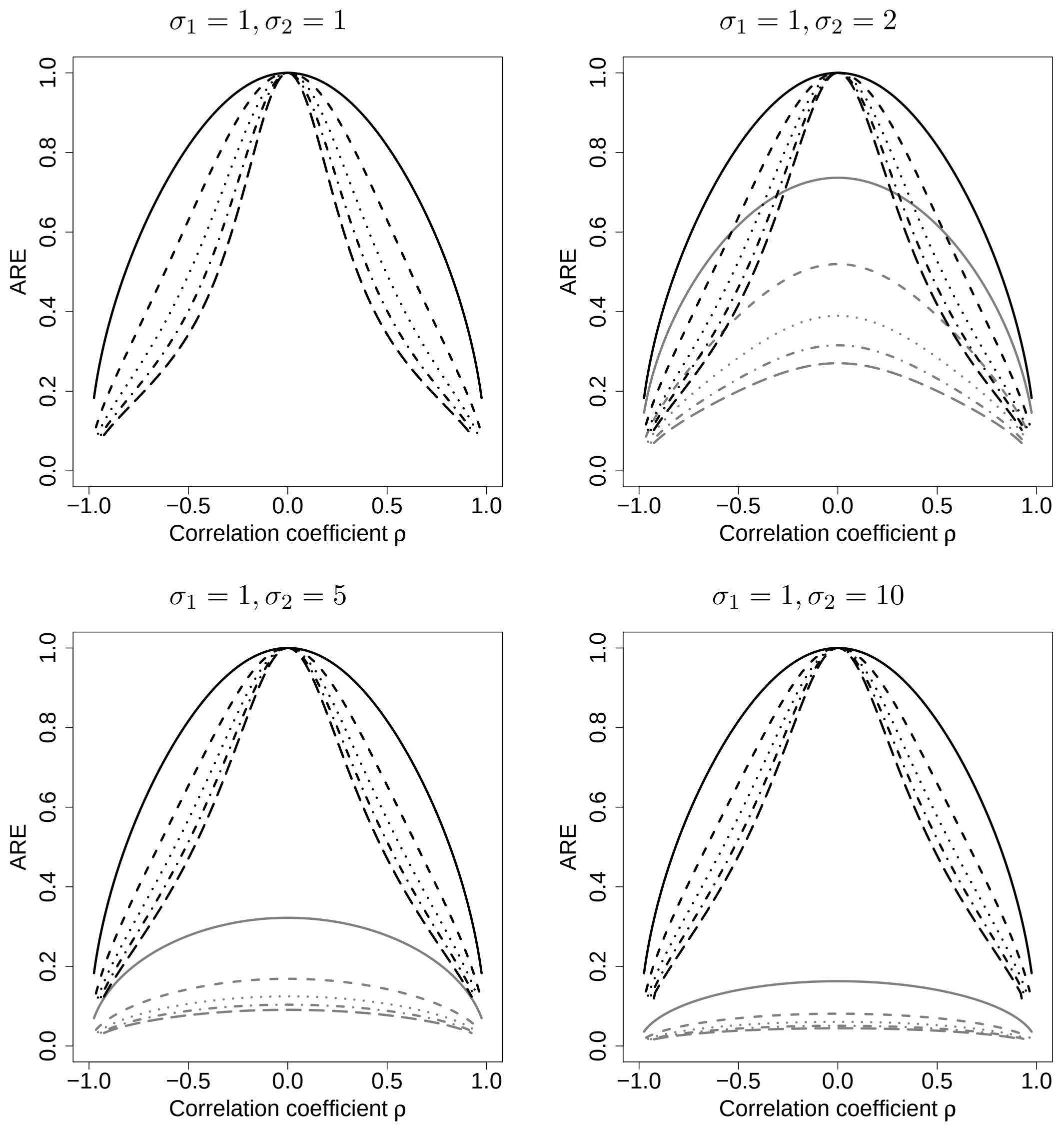

Figure 1: Each family of target densities is a single normal density with marginal variances σ_1 and σ_2, and correlation coefficient ρ. The black curves are $\text{ARE}(\mathcal{F} : \mathcal{D})$, the grey curves $\text{ARE}(\mathcal{F} : \mathcal{I})$. For $\sigma_1 = \sigma_2$ these two AREs coincide exactly. The derivatives are: $r = 0$ (solid), $r = 1$ (short dashed), $r = 2$ (dotted), $r = 3$ (dot-dashed), $r = 4$ (long dashed).

5 Application: high-throughput flow cytometry

Flow cytometry is a method by which multiple characteristics of single cells or other particles are simultaneously measured as they pass through a laser beam in a fluid stream (Shapiro, 2003). The last few years have seen a major change in flow cytometry technology, toward what has become known as *high-throughput* (or *high-content*) flow cytometry (e.g. Le Meur et al. (2007)). This modern technology combines robotic fluid handling, flow cytometric instrumentation and bioinformatics software so that relatively large numbers of cells can be processed and analyzed in a short period of time. With such massive amounts of data, there is a high premium on good automatic methods for pre-processing and extraction of clinically relevant information.

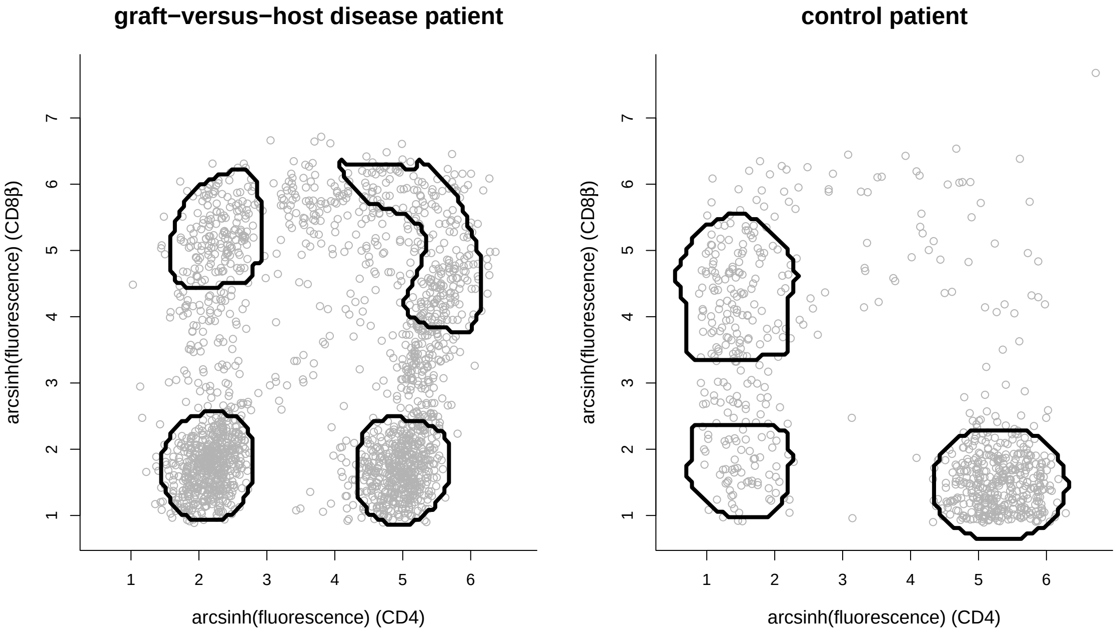

Figure 2: Cellular fluorescence measurements, after undergoing the arcsinh transformation, corresponding to antibodies CD4 and CD8β after subsetting on CD3-positive cells. The left panel is data from a patient who develops graft-versus-host disease. The right panel is data from a control patient. Further details about the data are given in Brinkman et al. (2007). The shapes correspond to significant negative density curvature regions using the methodology of Duong, Cowling, Koch and Wand (2008) with the bandwidth chosen via the normal scale rule (4).

Figure 2 is a subset of data from the flow cytometry experiment described in Brinkman et al. (2007). The left panel is cellular fluorescence measurements – corresponding to antibodies CD4 and CD8β, after gating on CD3-positive cells – on a patient who develops graft-versus-host disease. The right panel corresponds to a control. The data were collected 32 days after each patient had a blood and marrow transplant. The goal is to identify cell populations that differ between control and disease groups and, hence, constitute valid disease biomarkers, e.g. CD4-positive, CD8β-positive, CD3-positive; where 'positive' indicates fluorescence of the relevant antibody above a threshold. The shapes in Figure 2 correspond to regions of high significant negative curvature based on the methodology of Godtliebsen, Marron and Chaudhuri (2002) and refined by Duong, Cowling, Koch and Wand (2008).

The bandwidth matrix is chosen according to the normal scale rule (4) with $d = r = 2$:

$$\hat{\mathbf{H}}_{\text{NS}} = (1/2)^{1/5}\hat{\mathbf{\Sigma}}n^{-1/5} \tag{5}$$

where $\hat{\mathbf{\Sigma}}$ is the sample variance. Since the normal density is close to being the density which gives the largest optimal amount of smoothing given a fixed variance for density estimation (Terrell, 1990), then (5) corresponds approximately to the largest bandwidth matrix which should be considered for curvature estimation. The absence of significant curvature for CD4-positive and CD8β-positive cells in the control patient, despite use of this maximal bandwidth, represents an important clinical difference and gives rise to useful cellular signatures for graft-versus-host disease. Using the $r = 0$ normal scale rule, as illustrated in Table 1, could lead to insufficient smoothing for large sample sizes. In more comprehensive analyses of these data, described in Naumann and Wand (2009), more sophisticated filters for identifying cellular signatures are employed. The normal scale rule for second derivative estimation plays an important role in the initial phases of these filters, identifying candidate modal regions of possible interest. The plots in Figure 2 were computed using the `R` library `feature` whose main function uses (5) as the upper limit on the default bandwidth matrix range.

6 Discussion

Kernel smoothing is a widely used non-parametric method for multivariate density estimation. It has the potential to be as equally successful for density derivative estimation. The relative lack of theoretical development for density derivatives compared to densities has hindered this progress. One obstacle is the specification of higher order derivatives. By writing the rth order array of rth order differentials as an r-fold Kronecker product of first order differentials, we maintain an intuitive, systematic vectorization of all derivatives. This allows the derivation of the equivalent of standard quantities like MISE and AMISE for kernel density estimators for general derivatives.

The single most important factor in the performance of kernel estimators is the choice of the bandwidth. For density estimation, there is now a solid body of work for reliable bandwidth matrix selection. Using the theoretical simplifications afforded by our vector form derivatives, we can write down an unconstrained data-driven selector based on normal scales. These normal scale selectors facilitate the quantification of the possible gain in performance in using the unconstrained bandwidth matrices compared to more constrained parametrizations. These selectors are a starting point from which more advanced unconstrained bandwidth selectors can be now developed, and for the second derivative, they are a starting point from which to estimate modal regions.

Acknowledgments. J. E. Chacón has been partially supported by Spanish Ministerio de Ciencia y Tecnología project MTM2006-06172, T. Duong was funded by Institut Pasteur through a 'Programme Transversal de Recherches' grant (PTR No. 218), and M.P. Wand received support from Australian Research Council Grant DP055651. The authors are grateful for the assistance received from Nolwenn Le Meur and Richard White.

A Appendix: Proofs

A.1 Proof of the results in Section 2

A.1.1 Proof of Theorem 1

Proof of Theorem 1. First notice that we can write $\mathbb{E}\widehat{\mathsf{D}^{\otimes r}f}(\boldsymbol{x};\mathbf{H}) = \mathsf{D}^{\otimes r}K_{\mathbf{H}} * f(\boldsymbol{x}) = K_{\mathbf{H}} * \mathsf{D}^{\otimes r}f(\boldsymbol{x})$. Therefore,

$$\begin{aligned}\int \mathrm{B}^2(\boldsymbol{x};\mathbf{H})d\boldsymbol{x} &= \int \|K_{\mathbf{H}} * \mathsf{D}^{\otimes r}f(\boldsymbol{x}) - \mathsf{D}^{\otimes r}f(\boldsymbol{x})\|^2 d\boldsymbol{x} \\ &= \operatorname{tr}\mathbf{R}(\mathsf{D}^{\otimes r}f) + \operatorname{tr}\mathbf{R}_{K*K,\mathbf{H},r}(f) - 2\operatorname{tr}\mathbf{R}_{K,\mathbf{H},r}(f),\end{aligned}$$

as it is not difficult to check that $\int K_{\mathbf{H}} * \mathsf{D}^{\otimes r}f(\boldsymbol{x})K_{\mathbf{H}} * \mathsf{D}^{\otimes r}f(\boldsymbol{x})^T d\boldsymbol{x} = \mathbf{R}_{K*K,\mathbf{H},r}(f)$. About the variance term, it is clear that

$$\int \mathrm{V}(\boldsymbol{x};\mathbf{H})d\boldsymbol{x} = n^{-1}\int \mathbb{E}\|\mathsf{D}^{\otimes r}K_{\mathbf{H}}(\boldsymbol{x}-\mathbf{X}_1)\|^2 d\boldsymbol{x} - n^{-1}\int \|\mathbb{E}\mathsf{D}^{\otimes r}K_{\mathbf{H}}(\boldsymbol{x}-\mathbf{X}_1)\|^2 d\boldsymbol{x}. \quad (6)$$

The second integral in the right hand side is easily recognized as $\mathbf{R}_{K*K,\mathbf{H},r}(f)$ also, and for the first one we have

$$\begin{aligned}\int \mathbb{E}\|\mathsf{D}^{\otimes r}K_{\mathbf{H}}(\boldsymbol{x}-\mathbf{X}_1)\|^2 d\boldsymbol{x} &= \operatorname{tr}\iint \mathsf{D}^{\otimes r}K_{\mathbf{H}}(\boldsymbol{x}-\boldsymbol{y})\mathsf{D}^{\otimes r}K_{\mathbf{H}}(\boldsymbol{x}-\boldsymbol{y})^T f(\boldsymbol{y})d\boldsymbol{x}d\boldsymbol{y} \\ &= \operatorname{tr}\int \mathsf{D}^{\otimes r}K_{\mathbf{H}}(\boldsymbol{x})\mathsf{D}^{\otimes r}K_{\mathbf{H}}(\boldsymbol{x})^T d\boldsymbol{x} \\ &= \operatorname{tr}\Big[(\mathbf{H}^{-1/2})^{\otimes r}\int (\mathsf{D}^{\otimes r}K)_{\mathbf{H}}(\boldsymbol{x})(\mathsf{D}^{\otimes r}K)_{\mathbf{H}}(\boldsymbol{x})^T d\boldsymbol{x}(\mathbf{H}^{-1/2})^{\otimes r}\Big] \\ &= \operatorname{tr}\Big[(\mathbf{H}^{-1})^{\otimes r}|\mathbf{H}|^{-1/2}\int \mathsf{D}^{\otimes r}K(\boldsymbol{z})\mathsf{D}^{\otimes r}K(\boldsymbol{z})^T d\boldsymbol{z}\Big] \\ &= |\mathbf{H}|^{-1/2}\operatorname{tr}\big((\mathbf{H}^{-1})^{\otimes r}\mathbf{R}(\mathsf{D}^{\otimes r}K)\big). \qquad \square\end{aligned}$$

A.1.2 Proof of Theorem 2

Notice that, for a function $f\colon \mathbb{R}^d \to \mathbb{R}^p$ such that every element in $\mathsf{D}^{\otimes q}f(\boldsymbol{x})$ is piecewise continuous, we can write its Taylor polynomial expansion as

$$f(\boldsymbol{x}+\boldsymbol{h}) = \sum_{r=0}^{q}\frac{1}{r!}\big[\mathbf{I}_p \otimes (\boldsymbol{h}^T)^{\otimes r}\big]\mathsf{D}^{\otimes r}f(\boldsymbol{x}) + o(\|\boldsymbol{h}\|^q)\mathbf{1}_p, \qquad \boldsymbol{x},\boldsymbol{h} \in \mathbb{R}^d.$$

See Baxandall and Liebeck (1986, p. 164). The proof of Theorem 2 then follows from Lemmas 1 and 2 below, together with the bias-variance decomposition of the MSE.

Denote $\mathrm{IB}^2(\mathbf{H}) = \int \mathrm{B}^2(\boldsymbol{x};\mathbf{H})d\boldsymbol{x}$ and $\mathrm{IV}(\mathbf{H}) = \int \mathrm{V}(\boldsymbol{x};\mathbf{H})d\boldsymbol{x}$ the integrated squared bias and integrated variance of the kernel estimator, respectively, so that we can write $\mathrm{MISE}(\mathbf{H}) = \mathrm{IB}^2(\mathbf{H}) + \mathrm{IV}(\mathbf{H})$.

Lemma 1. *Assume that (A1)–(A3) hold. We can expand*

$$\mathrm{IB}^2(\mathbf{H}) = \tfrac{m_2(K)^2}{4}\,\mathrm{tr}\left[(\mathbf{I}_{d^r} \otimes \mathrm{vec}^T\,\mathbf{H})\mathbf{R}(\mathsf{D}^{\otimes(r+2)}f)(\mathbf{I}_{d^r} \otimes \mathrm{vec}\,\mathbf{H})\right] + o(\mathrm{tr}^2\,\mathbf{H})$$

Proof. We can write $\mathbb{E}\widehat{\mathsf{D}^{\otimes r}f}(\boldsymbol{x};\mathbf{H}) = \int K(\boldsymbol{z})\mathsf{D}^{\otimes r}f(\boldsymbol{x}-\mathbf{H}^{1/2}\boldsymbol{z})d\boldsymbol{z}$. Now, make use of a Taylor expansion to get

$$\begin{aligned}\mathsf{D}^{\otimes r}f(\boldsymbol{x} - \mathbf{H}^{1/2}\boldsymbol{z}) = \mathsf{D}^{\otimes r}f(\boldsymbol{x}) &- \left[\mathbf{I}_{d^r} \otimes (\boldsymbol{z}^T\mathbf{H}^{1/2})\right]\mathsf{D}^{\otimes(r+1)}f(\boldsymbol{x})\\ &+ \tfrac{1}{2}\left[\mathbf{I}_{d^r} \otimes (\boldsymbol{z}^T\mathbf{H}^{1/2})^{\otimes 2}\right]\mathsf{D}^{\otimes(r+2)}f(\boldsymbol{x}) + o(\mathrm{tr}\,\mathbf{H})\mathbf{1}_{d^r}.\end{aligned}$$

Substitute this in the previous formula and use assumption (A3) to obtain

$$\begin{aligned}\mathrm{B}(\boldsymbol{x};\mathbf{H}) &= \tfrac{m_2(K)}{2}\left\|\left[\mathbf{I}_{d^r} \otimes \{(\mathrm{vec}^T\,\mathbf{I}_d)(\mathbf{H}^{1/2})^{\otimes 2}\}\right]\mathsf{D}^{\otimes(r+2)}f(\boldsymbol{x}) + o(\mathrm{tr}\,\mathbf{H})\right\|\\ &= \tfrac{m_2(K)}{2}\left\|\left(\mathbf{I}_{d^r} \otimes \mathrm{vec}^T\,\mathbf{H}\right)\mathsf{D}^{\otimes(r+2)}f(\boldsymbol{x}) + o(\mathrm{tr}\,\mathbf{H})\right\|.\end{aligned}$$

We finish the proof by squaring and integrating the previous expression, taking into account assumption (A2). □

Lemma 2. *Assume that (A1) holds. We can expand*

$$\mathrm{IV}(\mathbf{H}) = n^{-1}|\mathbf{H}|^{-1/2}\,\mathrm{tr}\left((\mathbf{H}^{-1})^{\otimes r}\mathbf{R}(\mathsf{D}^{\otimes r}K)\right) + o(n^{-1}|\mathbf{H}|^{-1/2}\,\mathrm{tr}^r(\mathbf{H}^{-1})).$$

Proof. From the proof of Theorem 1 and the arguments in the previous lemma we have

$$\begin{aligned}\int \mathrm{V}(\boldsymbol{x};\mathbf{H})d\boldsymbol{x} &= n^{-1}\int \mathbb{E}\|\mathsf{D}^{\otimes r}K_{\mathbf{H}}(\boldsymbol{x} - \mathbf{X}_1)\|^2 d\boldsymbol{x} + O(n^{-1})\\ &= n^{-1}|\mathbf{H}|^{-1/2}\,\mathrm{tr}\left((\mathbf{H}^{-1})^{\otimes r}\mathbf{R}(\mathsf{D}^{\otimes r}K)\right) + o(n^{-1}|\mathbf{H}|^{-1/2}\,\mathrm{tr}^r(\mathbf{H}^{-1})).\end{aligned}$$ □

A.1.3 Proof of Theorem 3

Proof. Similar to the decomposition $\mathrm{MISE}(\mathbf{H}) = \mathrm{IB}^2(\mathbf{H})+\mathrm{IV}(\mathbf{H})$, we can write $\mathrm{AMISE}(\mathbf{H}) = \mathrm{AIB}^2(\mathbf{H})+\mathrm{AIV}(\mathbf{H})$ where $\mathrm{AIB}^2(\mathbf{H})$ and $\mathrm{AIV}(\mathbf{H})$ are the leading terms of $\mathrm{IB}^2(\mathbf{H})$ and $\mathrm{IV}(\mathbf{H})$ respectively from Lemmas 1 and 2.

Denote $\mathbf{K}_{r,s} \in \mathcal{M}_{rs\times rs}$ the commutation matrix of order r, s; see Magnus and Neudecker (1979). The commutation matrix allows us to commute the order of the matrices in a Kronecker product e.g., if $\mathbf{A} \in \mathcal{M}_{n\times r}$ and $\mathbf{B} \in \mathcal{M}_{m\times s}$, then $\mathbf{K}_{m,n}(\mathbf{A} \otimes \mathbf{B})\mathbf{K}_{r,s} = \mathbf{B} \otimes \mathbf{A}$.

To determine the derivative, we first find the differentials. Differentials of a function $f : \mathbb{R}^d \to \mathbb{R}^p$ have the advantage that they are always the same dimension as f itself, as opposed to derivatives whose dimension depends on the order of the derivative. So higher order differentials are easier to manipulate. The first identification theorem of Magnus and

Neudecker (1999, p. 87) states if the differential of $f(\boldsymbol{y})$ can be expressed as $df(\boldsymbol{y}) = A(\boldsymbol{y})\boldsymbol{y}$ for some matrix $A(\boldsymbol{y}) \in \mathcal{M}_{p\times d}$ then the derivative is $\mathsf{D}f(\boldsymbol{y}) = A(\boldsymbol{y})$. The differential of $\mathrm{AIB}^2(\mathbf{H})$ is

$$\begin{aligned} d\mathrm{AIB}^2(\mathbf{H}) = {} & \tfrac{m_2(K)^2}{4}(\operatorname{vec}^T \mathbf{R}(\mathsf{D}^{\otimes(r+2)}f))\mathbf{K}^{\otimes 2}_{d^r,d^2}[(\mathbf{K}_{d^{r+2},d^2}\otimes \mathbf{I}_{d^r}) + \mathbf{I}_{d^{2r+4}}] \\ & \times (\mathbf{I}_{d^2}\otimes \operatorname{vec}\mathbf{H}\otimes \operatorname{vec}\mathbf{I}_{d^r})d\operatorname{vec}\mathbf{H} \end{aligned}$$

since

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

where the last line follows by using a similar reasoning to determine Equation (11) in the proof of Theorem 2 in Chacón and Duong (2008).

The differential of $\mathrm{AIV}(\mathbf{H})$ is

$$\begin{aligned} d\mathrm{AIV}(\mathbf{H}) = -\Big\{ & \tfrac{1}{2}\mathrm{AIV}(\mathbf{H})(\operatorname{vec}^T\mathbf{H}^{-1}) + n^{-1}|\mathbf{H}|^{-1/2}(\operatorname{vec}^T\mathbf{R}(\mathsf{D}^{\otimes r}K))(\mathbf{H}^{-1})^{\otimes 2r} \\ & \times \boldsymbol{\Lambda}_r[(\mathbf{I}_{d^{r-1}}\otimes \mathbf{K}_{d,d^{r-1}})(\operatorname{vec}\mathbf{H}^{\otimes(r-1)}\otimes \mathbf{I}_d)\otimes \mathbf{I}_d]\Big\}\, d\operatorname{vec}\mathbf{H} \end{aligned}$$

where $\boldsymbol{\Lambda}_r = \sum_{i=1}^r \mathbf{K}^{\otimes 2}_{d^i,d^{r-i}}$. The reasoning follows similar lines to computing Equations (9) and (10) in the proof of Theorem 2 in Chacón and Duong (2008).

Let every entry of $\mathbf{H}$ be $O(n^{-\beta})$ for $\beta > 0$. Then $d\mathrm{AIV}(\mathbf{H}) = O(n^{\beta(d/2+r+1)-1})\ d\operatorname{vec}\mathbf{H}$ and $d\mathrm{AIB}^2(\mathbf{H}) = O(n^{-\beta})\ d\operatorname{vec}\mathbf{H}$. Equating powers gives $\beta = 2/(d+2r+4)$ and thus $d\mathrm{AMISE}(\mathbf{H}) = O(n^{-2/(d+2r+4)})\ d\operatorname{vec}\mathbf{H}$. The optimal $\mathbf{H}$ is a solution of the equation $\partial\mathrm{AMISE}(\mathbf{H})/(\partial\operatorname{vec}\mathbf{H}) = 0$, so all its entries are $O(n^{-2/(d+2r+4)})$, which implies that $\min_{\mathbf{H}}\mathrm{AMISE}(\mathbf{H}) = O(n^{-4/(d+2r+4)})$. □

A.1.4 Proof of Theorem 4

The proof follows directly from Lemmas 1 and 2.

Proof. From Lemma 1, $\mathbb{E}\,\widehat{\mathsf{D}^{\otimes r}f}(\boldsymbol{x};\mathbf{H}) = \mathsf{D}^{\otimes r}f(\boldsymbol{x}) + \frac{m_2(K)}{2}(\mathbf{I}_{d^r}\otimes \operatorname{vec}^T\mathbf{H})\mathsf{D}^{\otimes(r+2)}f(\boldsymbol{x})[1 + O(\operatorname{tr}\mathbf{H})]$. For the variance, we have

$$\operatorname{Var}\widehat{\mathsf{D}^{\otimes r}f}(\boldsymbol{x};\mathbf{H}) = n^{-1}\mathsf{D}^{\otimes r}K_{\mathbf{H}}\mathsf{D}^{\otimes r}K_{\mathbf{H}}^T * f(\boldsymbol{x}) - n^{-1}[K_{\mathbf{H}} * \mathsf{D}^{\otimes r}f(\boldsymbol{x})][K_{\mathbf{H}} * \mathsf{D}^{\otimes r}f(\boldsymbol{x})^T].$$

From Lemma 2, the convolution in the first term dominates the convolution in the second term since the value of the former is

$$\mathsf{D}^{\otimes r}K_{\mathbf{H}}\mathsf{D}^{\otimes r}K_{\mathbf{H}}^T * f(\boldsymbol{x}) = |\mathbf{H}|^{-1/2}(\mathbf{H}^{-1/2})^{\otimes r}\mathbf{R}(\mathsf{D}^{\otimes r}K)(\mathbf{H}^{-1/2})^{\otimes r}f(\boldsymbol{x})[1 + o(1)]$$

and the proof is complete. □

A.2 Proof of the results in Section 3

The proofs in Sections A.2.1 and A.2.2 assume without loss of generality that $f = \phi_{\mathbf{\Sigma}}$ to simplify the presentation of the results. These results for the general normal density $f = \phi_{\mathbf{\Sigma}}(\cdot - \boldsymbol{\mu})$ remain valid since they are invariant under this translation.

A.2.1 Proof of Theorem 5

The proof of Theorem 5 is based on the exact formula given in Theorem 1. Notice that, in the normal case, we have $\mathbf{R}_{\phi*\phi,\mathbf{H},r}(\phi_{\mathbf{\Sigma}}) = \mathbf{R}_{\phi,2\mathbf{H},r}(\phi_{\mathbf{\Sigma}})$ and $\mathbf{R}(\mathsf{D}^{\otimes r}\phi_{\mathbf{\Sigma}}) = \mathbf{R}_{\phi,\mathbf{0},r}(\phi_{\mathbf{\Sigma}})$, so that it follows that all we need to have an explicit expression for the MISE function in the normal case is just to obtain explicit formulas for $\operatorname{tr}\mathbf{R}_{\phi,\mathbf{H},r}(\phi_{\mathbf{\Sigma}})$ and $\operatorname{tr}\big((\mathbf{H}^{-1})^{\otimes r}\mathbf{R}(\mathsf{D}^{\otimes r}\phi)\big)$. These are provided in the following lemma.

Lemma 3. *For any symmetric positive definite matrix* $\mathbf{H}$ *we have*

i) $\mathbf{R}_{\phi,\mathbf{H},r}(\phi_{\mathbf{\Sigma}}) = 2^{d/2+r}\mathbf{R}(\mathsf{D}^{\otimes r}\phi_{\mathbf{H}+2\mathbf{\Sigma}})$.

ii) $\operatorname{tr}\big((\mathbf{H}^{-1})^{\otimes r}\mathbf{R}(\mathsf{D}^{\otimes r}\phi_{\mathbf{\Sigma}})\big) = 2^{-(d+r)}\pi^{-d/2}|\mathbf{\Sigma}|^{-1/2}\mu_r(\mathbf{\Sigma}^{1/2}\mathbf{H}\mathbf{\Sigma}^{1/2})$.

From i) and ii) we immediately obtain

iii) $\operatorname{tr}\mathbf{R}_{\phi,\mathbf{H},r}(\phi_{\mathbf{\Sigma}}) = (2\pi)^{-d/2}|\mathbf{H}+2\mathbf{\Sigma}|^{-1/2}\mu_r(\mathbf{H}+2\mathbf{\Sigma})$.

iv) $\operatorname{tr}\big((\mathbf{H}^{-1})^{\otimes r}\mathbf{R}(\mathsf{D}^{\otimes r}\phi)\big) = 2^{-(d+r)}\pi^{-d/2}\mu_r(\mathbf{H})$.

Proof. *i*) Reasoning as in Chacón and Duong (2008), it is easy to check that

$$\operatorname{vec}\mathbf{R}(\mathsf{D}^{\otimes r}\phi_{\mathbf{\Sigma}}) = (-1)^r\mathsf{D}^{\otimes 2r}\phi_{2\mathbf{\Sigma}}(\mathbf{0}) = (-1)^r 2^{-(d/2+r)}\mathsf{D}^{\otimes 2r}\phi_{\mathbf{\Sigma}}(\mathbf{0}).$$

With this in mind, an element-wise application of some of the results in Appendix C of Wand and Jones (1995) leads to

$$\begin{aligned}
\operatorname{vec}\mathbf{R}_{\phi,\mathbf{H},r}(\phi_{\mathbf{\Sigma}}) &= \operatorname{vec}\int_{\mathbb{R}^d}(\phi_{\mathbf{H}} * \mathsf{D}^{\otimes r}\phi_{\mathbf{\Sigma}})(\boldsymbol{x})\mathsf{D}^{\otimes r}\phi_{\mathbf{\Sigma}}(\boldsymbol{x})^T d\boldsymbol{x} \\
&= \operatorname{vec}\int_{\mathbb{R}^d}\mathsf{D}^{\otimes r}\phi_{\mathbf{H}+\mathbf{\Sigma}}(\boldsymbol{x})\mathsf{D}^{\otimes r}\phi_{\mathbf{\Sigma}}(\boldsymbol{x})^T d\boldsymbol{x} \\
&= \int_{\mathbb{R}^d}\mathsf{D}^{\otimes r}\phi_{\mathbf{\Sigma}}(\boldsymbol{x}) \otimes \mathsf{D}^{\otimes r}\phi_{\mathbf{H}+\mathbf{\Sigma}}(\boldsymbol{x})d\boldsymbol{x} \\
&= (-1)^r\int_{\mathbb{R}^d}\mathsf{D}^{\otimes 2r}\phi_{\mathbf{\Sigma}}(\boldsymbol{x})\phi_{\mathbf{H}+\mathbf{\Sigma}}(\boldsymbol{x})d\boldsymbol{x} \\
&= (-1)^r\mathsf{D}^{\otimes 2r}\phi_{\mathbf{H}+2\mathbf{\Sigma}}(\mathbf{0}) \\
&= 2^{d/2+r}\operatorname{vec}\mathbf{R}(\mathsf{D}^{\otimes r}\phi_{\mathbf{H}+2\mathbf{\Sigma}}),
\end{aligned}$$

which yields the result.

ii) Chacón and Duong (2008) also show that

$$\operatorname{vec}\mathbf{R}(\mathsf{D}^{\otimes r}\phi_{\boldsymbol{\Sigma}}) = 2^{-(d+r)}\pi^{-d/2}\mathrm{OF}(2r)|\boldsymbol{\Sigma}|^{-1/2}\boldsymbol{\mathcal{S}}_{d,2r}(\operatorname{vec}\boldsymbol{\Sigma}^{-1})^{\otimes r}. \tag{7}$$

Moreover, it is not hard to check that the symmetrizer matrix fulfills $\boldsymbol{\mathcal{S}}_{d,2r}\operatorname{vec}[(\mathbf{H}^{-1})^{\otimes r}] = \boldsymbol{\mathcal{S}}_{d,2r}(\operatorname{vec}\mathbf{H}^{-1})^{\otimes r}$. This is because $(\operatorname{vec}\mathbf{H}^{-1})^{\otimes r}$ can be obtained form $\operatorname{vec}[(\mathbf{H}^{-1})^{\otimes r}]$ by multiplying it by Kronecker products of commutation and identity matrices, and multiplication of this kind of matrices by the symmetrizer matrix has no effect, as seen from part (iv) of Theorem 1 in Schott (2003). Therefore, if $\boldsymbol{z}$ denotes a d-variate vector with standard normal distribution and $\boldsymbol{x} = \boldsymbol{\Sigma}^{-1/2}\boldsymbol{z}$, then

$$\begin{aligned}
\operatorname{tr}\big((\mathbf{H}^{-1})^{\otimes r}\mathbf{R}(\mathsf{D}^{\otimes r}\phi_{\boldsymbol{\Sigma}})\big) &= \operatorname{vec}^T[(\mathbf{H}^{-1})^{\otimes r}]\operatorname{vec}\mathbf{R}(\mathsf{D}^{\otimes r}\phi_{\boldsymbol{\Sigma}}) \\
&= 2^{-(d+r)}\pi^{-d/2}\mathrm{OF}(2r)|\boldsymbol{\Sigma}|^{-1/2}\operatorname{vec}^T[(\mathbf{H}^{-1})^{\otimes r}]\boldsymbol{\mathcal{S}}_{d,2r}(\operatorname{vec}\boldsymbol{\Sigma}^{-1})^{\otimes r} \\
&= 2^{-(d+r)}\pi^{-d/2}\mathrm{OF}(2r)|\boldsymbol{\Sigma}|^{-1/2}(\operatorname{vec}^T\mathbf{H}^{-1})^{\otimes r}\boldsymbol{\mathcal{S}}_{d,2r}(\operatorname{vec}\boldsymbol{\Sigma}^{-1})^{\otimes r} \\
&= 2^{-(d+r)}\pi^{-d/2}|\boldsymbol{\Sigma}|^{-1/2}\mathbb{E}[(\boldsymbol{x}^T\mathbf{H}^{-1}\boldsymbol{x})^r] \\
&= 2^{-(d+r)}\pi^{-d/2}|\boldsymbol{\Sigma}|^{-1/2}\mathbb{E}[(\boldsymbol{z}^T\boldsymbol{\Sigma}^{-1/2}\mathbf{H}^{-1}\boldsymbol{\Sigma}^{-1/2}\boldsymbol{z})^r],
\end{aligned}$$

Here, the fourth line follows from Theorem 1 in Holmquist (1996b). □

A.2.2 Proof of Theorem 6

The proof of Theorem 6 starts from the AMISE expression given in Theorem 2. The term appearing in the asymptotic integrated variance was already computed in Lemma 3 above. For the asymptotic integrated squared bias, it is clear that for the normal kernel we have $m_2(K) = 1$. From (7) and the results in Holmquist (1996a) it follows that

$$\operatorname{vec}\mathbf{R}(\mathsf{D}^{\otimes r}\phi_{\boldsymbol{\Sigma}}) = 2^{-(d+r)}\pi^{-d/2}|\boldsymbol{\Sigma}|^{-1/2}(\boldsymbol{\Sigma}^{-1/2})^{\otimes 2r}\mathbb{E}[\boldsymbol{z}^{\otimes 2r}]$$

with $\boldsymbol{z}$ a d-variate standard normal random vector. Or, in matrix form,

$$\mathbf{R}(\mathsf{D}^{\otimes r}\phi_{\boldsymbol{\Sigma}}) = 2^{-(d+r)}\pi^{-d/2}|\boldsymbol{\Sigma}|^{-1/2}(\boldsymbol{\Sigma}^{-1/2})^{\otimes r}\mathbb{E}[(\boldsymbol{z}\boldsymbol{z}^T)^{\otimes r}](\boldsymbol{\Sigma}^{-1/2})^{\otimes r}.$$

Therefore, using $(\boldsymbol{\Sigma}^{-1/2}\otimes\boldsymbol{\Sigma}^{-1/2})\operatorname{vec}\mathbf{H} = \operatorname{vec}(\boldsymbol{\Sigma}^{-1/2}\mathbf{H}\boldsymbol{\Sigma}^{-1/2})$ and some other matrix results from Magnus and Neudecker (1999, p. 48), we come to

$$\begin{aligned}
&\operatorname{tr}\big[(\mathbf{I}_{d^r}\otimes\operatorname{vec}^T\mathbf{H})\mathbf{R}(\mathsf{D}^{\otimes(r+2)}\phi_{\boldsymbol{\Sigma}})(\mathbf{I}_{d^r}\otimes\operatorname{vec}\mathbf{H})\big] \\
&\qquad = 2^{-(d+r+2)}\pi^{-d/2}|\boldsymbol{\Sigma}|^{-1/2}\operatorname{tr}\Big[\big\{(\boldsymbol{\Sigma}^{-1})^{\otimes r}\otimes(\operatorname{vec}\mathbf{B}\operatorname{vec}^T\mathbf{B})\big\}\mathbb{E}[(\boldsymbol{z}\boldsymbol{z}^T)^{\otimes(r+2)}]\Big]
\end{aligned}$$

with $\mathbf{B} = \boldsymbol{\Sigma}^{-1/2}\mathbf{H}\boldsymbol{\Sigma}^{-1/2}$. Now, the trace in the right hand side can be written as

$$\begin{aligned}
\mathbb{E}\operatorname{tr}\Big[(\boldsymbol{\Sigma}^{-1}\boldsymbol{z}\boldsymbol{z}^T)^{\otimes r}\otimes\{\operatorname{vec}\mathbf{B}\operatorname{vec}^T\mathbf{B}(\boldsymbol{z}\boldsymbol{z}^T)^{\otimes 2}\}\Big] &= \mathbb{E}\Big[\operatorname{tr}^r(\boldsymbol{\Sigma}^{-1}\boldsymbol{z}\boldsymbol{z}^T)\operatorname{tr}\{\operatorname{vec}\mathbf{B}\operatorname{vec}^T(\boldsymbol{z}\boldsymbol{z}^T\mathbf{B}\boldsymbol{z}\boldsymbol{z}^T)\}\Big] \\
&= \mathbb{E}\Big[(\boldsymbol{z}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{z})^r\{\operatorname{vec}^T(\boldsymbol{z}\boldsymbol{z}^T\mathbf{B}\boldsymbol{z}\boldsymbol{z}^T)\operatorname{vec}\mathbf{B}\}\Big] \\
&= \mathbb{E}\Big[(\boldsymbol{z}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{z})^r(\boldsymbol{z}^T\mathbf{B}\boldsymbol{z})^2\Big].
\end{aligned}$$

This yields the proof for the AMISE formula.

If we evaluate the AMISE formula in Theorem 6 at $\mathbf{H} = c\boldsymbol{\Sigma}$ for some $c > 0$ we obtain

$$\mathrm{AMISE}(c\boldsymbol{\Sigma}) = 2^{-(d+r)}\pi^{-d/2}|\boldsymbol{\Sigma}|^{-1/2}\big\{n^{-1}c^{-(d+2r)/2}\mu_r(\boldsymbol{\Sigma}) + \tfrac{1}{16}c^2\mu_{r,2}(\boldsymbol{\Sigma}, \mathbf{I}_d)\big\}.$$

But we will show below that

$$\mu_{r,2}(\boldsymbol{\Sigma}, \mathbf{I}_d) = (d + 2r + 2)(d + 2r)\mu_r(\boldsymbol{\Sigma}), \tag{8}$$

leading to

$$\mathrm{AMISE}(c\boldsymbol{\Sigma}) = 2^{-(d+r)}\pi^{-d/2}|\boldsymbol{\Sigma}|^{-1/2}\mu_r(\boldsymbol{\Sigma})\big\{n^{-1}c^{-(d+2r)/2} + \tfrac{1}{16}c^2(d + 2r + 2)(d + 2r)\big\}.$$

and this function is minimized by setting

$$c = \left(\frac{4}{(d + 2r + 2)n}\right)^{2/(d+2r+4)}.$$

Therefore, to finish the proof the only thing left is to show equality (8).

This task, however, is harder than it may seem at first sight. It is relatively easy if $\boldsymbol{\Sigma} = \mathbf{I}_d$ because, in this case, it suffices to show that $\mu_{r+1}(\mathbf{I}_d) = (d + 2r)\mu_r(\mathbf{I}_d)$, and that is an immediate consequence of the recursive formula (3). Therefore, to show (8) we will need a recursive formula similar to (3), but for the joint moments $\mu_{r,s}(\mathbf{A}, \mathbf{B})$. To that end, we first derive a technical lemma.

Lemma 4. *Consider the real function* $g_\alpha(t) \equiv g_\alpha(t; \mathbf{A}, \mathbf{B}, \mathbf{C}) = \operatorname{tr}\big[\{\mathbf{B}(\mathbf{C} + t\mathbf{A})^{-1}\}^\alpha\big]$ *for suitable matrices* $\mathbf{A}, \mathbf{B}, \mathbf{C}$ *and arbitrary* $\alpha \in \mathbb{N}$*. Then, the pth derivative of g is given by*

$$g_\alpha^{(p)}(t; \mathbf{A}, \mathbf{B}, \mathbf{C}) = (-1)^p\frac{(\alpha + p - 1)!}{(\alpha - 1)!}\operatorname{tr}\big[\{\mathbf{A}(\mathbf{C} + t\mathbf{A})^{-1}\}^p\{\mathbf{B}(\mathbf{C} + t\mathbf{A})^{-1}\}^\alpha\big]$$

so that $g_\alpha^{(p)}(0; \mathbf{A}, \mathbf{B}, \mathbf{C}) = (-1)^p\frac{(\alpha+p-1)!}{(\alpha-1)!}\operatorname{tr}\big[(\mathbf{A}\mathbf{C}^{-1})^p(\mathbf{B}\mathbf{C}^{-1})^\alpha\big]$.

Proof. The result is proved by induction on p. For $p = 1$, noting that the differential of $\mathbf{B}(\mathbf{C} + t\mathbf{A})^{-1}$ is $d\big[\mathbf{B}(\mathbf{C} + t\mathbf{A})^{-1}\big] = -\mathbf{B}(\mathbf{C} + t\mathbf{A})^{-1}\mathbf{A}(\mathbf{C} + t\mathbf{A})^{-1}dt$, we have

$$\begin{aligned}
d\operatorname{tr}\big[\{\mathbf{B}(\mathbf{C} + t\mathbf{A})^{-1}\}^\alpha\big] &= \operatorname{tr} d\big[\{\mathbf{B}(\mathbf{C} + t\mathbf{A})^{-1}\}^\alpha\big] \\
&= \operatorname{tr}\sum_{i=1}^{\alpha}\{\mathbf{B}(\mathbf{C} + t\mathbf{A})^{-1}\}^{i-1}\cdot d\big[\mathbf{B}(\mathbf{C} + t\mathbf{A})^{-1}\big]\cdot\{\mathbf{B}(\mathbf{C} + t\mathbf{A})^{-1}\}^{\alpha-i} \\
&= -\operatorname{tr}\sum_{i=1}^{\alpha}\{\mathbf{B}(\mathbf{C} + t\mathbf{A})^{-1}\}^{i}\cdot\mathbf{A}(\mathbf{C} + t\mathbf{A})^{-1}\cdot\{\mathbf{B}(\mathbf{C} + t\mathbf{A})^{-1}\}^{\alpha-i}dt \\
&= -\alpha\operatorname{tr}\big[\mathbf{A}(\mathbf{C} + t\mathbf{A})^{-1}\{\mathbf{B}(\mathbf{C} + t\mathbf{A})^{-1}\}^\alpha\big]dt
\end{aligned}$$

and we are done. The case of arbitrary p follows easily by considering $g_\alpha^{(p)}(t) = \frac{d}{dt}g_\alpha^{(p-1)}(t)$. $\square$

The recursive formula for $\mu_{r,s}(\mathbf{A},\mathbf{B})$ is given in the next theorem.

Theorem 9. *We can write*

$$\mu_{r,s}(\mathbf{A},\mathbf{B}) = \sum_{i=0}^{r}\sum_{j=0}^{s-1}\binom{r}{i}\binom{s-1}{j}(r+s-i-j-1)!2^{r+s-i-j-1}\operatorname{tr}(\mathbf{A}^{-(r-i)}\mathbf{B}^{-(s-j)})\mu_{i,j}(\mathbf{A},\mathbf{B}).$$

Proof. For ease of notation we will prove the result for $\mu_{r,s}(\mathbf{A}^{-1},\mathbf{B}^{-1})$; that is, we will show that, for $q_{\mathbf{A}} = \boldsymbol{z}^T\mathbf{A}\boldsymbol{z}$ and $q_{\mathbf{B}} = \boldsymbol{z}^T\mathbf{B}\boldsymbol{z}$ with $\boldsymbol{z}$ a d-variate standard normal random vector, we have

$$\mathbb{E}[q_{\mathbf{A}}^r q_{\mathbf{B}}^s] = \sum_{i=0}^{r}\sum_{j=0}^{s-1}\binom{r}{i}\binom{s-1}{j}(r+s-i-j-1)!2^{r+s-i-j-1}\operatorname{tr}(\mathbf{A}^{r-i}\mathbf{B}^{s-j})\mathbb{E}[q_{\mathbf{A}}^i q_{\mathbf{B}}^j].$$

It is well known that the joint moment generating function of $q_{\mathbf{A}}$ and $q_{\mathbf{B}}$ is

$$M(t_1,t_2) = \mathbb{E}[e^{t_1 q_{\mathbf{A}} + t_2 q_{\mathbf{B}}}] = |\mathbf{I}_d - 2t_1\mathbf{A} - 2t_2\mathbf{B}|^{-1/2},$$

see Magnus (1986). From that, we can write

$$\mathbb{E}[q_{\mathbf{A}}^r q_{\mathbf{B}}^s] = \frac{\partial^{r+s}M}{\partial t_1^r \partial t_2^s}(0,0),$$

so that all we need is to find a recursive formula for the partial derivatives of M. With the notations of the previous lemma, it is easy to show that

$$\frac{\partial M}{\partial t_2}(t_1,t_2) = M(t_1,t_2)\cdot g_1(t_2;-2\mathbf{B},\mathbf{B},\mathbf{I}_d - 2t_1\mathbf{A}).$$

This way, using the formulas for the derivatives of g_1 and Leibniz formula for the derivatives of a product,

$$\begin{aligned}
\frac{\partial^s M}{\partial t_2^s}(t_1,t_2) &= \frac{\partial^{s-1}}{\partial t_2^{s-1}}\Big(M(t_1,t_2)\cdot g_1(t_2;-2\mathbf{B},\mathbf{B},\mathbf{I}_d - 2t_1\mathbf{A})\Big)\\
&= \sum_{j=1}^{s-1}\binom{s-1}{j}\frac{\partial^j M}{\partial t_2^j}(t_1,t_2)\cdot g_1^{(s-j-1)}(t_2;-2\mathbf{B},\mathbf{B},\mathbf{I}_d - 2t_1\mathbf{A})\\
&= \sum_{j=1}^{s-1}\binom{s-1}{j}(s-j-1)!2^{s-j-1}\frac{\partial^j M}{\partial t_2^j}(t_1,t_2)\cdot\operatorname{tr}\left[\{\mathbf{B}(\mathbf{I}_d - 2t_1\mathbf{A} - 2t_2\mathbf{B})^{-1}\}^{s-j}\right]\\
&= \sum_{j=1}^{s-1}\frac{(s-1)!}{j!}2^{s-j-1}\frac{\partial^j M}{\partial t_2^j}(t_1,t_2)\cdot g_{s-j}(t_1;-2\mathbf{A},\mathbf{B},\mathbf{I}_d - 2t_2\mathbf{B}).
\end{aligned}$$

Now, if we compute the rth partial derivative with respect to t_1 we have

$$\frac{\partial^{r+s}M}{\partial t_1^r\partial t_2^s}(t_1,t_2) = \sum_{i=0}^{r}\sum_{j=1}^{s-1}\binom{r}{i}\frac{(s-1)!}{j!}2^{s-j-1}\frac{\partial^{i+j}M}{\partial t_1^i\partial t_2^j}(t_1,t_2)\cdot g_{s-j}^{(r-i)}(t_1;-2\mathbf{A},\mathbf{B},\mathbf{I}_d - 2t_2\mathbf{B})$$

Substituting (t_1,t_2) in this expression for $(0,0)$ and using again the previous lemma, we get the desired formula. □

As a consequence of this result, we finally are able to prove formula (8).

Corollary 7. *For any symmetric matrix* $\mathbf{A}$ *we have*

$$\mu_{r,2}(\mathbf{A},\mathbf{I}_d) = (d+2r+2)(d+2r)\mu_r(\mathbf{A}).$$

Proof. First, notice that from the previous theorem, taking into account that $\mathbf{A}^0 = \mathbf{I}_d$,

$$\begin{aligned}\mu_{r,1}(\mathbf{A},\mathbf{I}_d) &= \sum_{i=0}^{r}\binom{r}{i}(r-i)!2^{r-i}\operatorname{tr}(\mathbf{A}^{-(r-i)})\mu_{i,0}(\mathbf{A},\mathbf{I}_d)\\ &= d\mu_r(\mathbf{A}) + \sum_{i=0}^{r-1}\frac{r!}{i!}2^{r-i}\operatorname{tr}(\mathbf{A}^{-(r-i)})\mu_i(\mathbf{A})\\ &= (d+2r)\mu_r(\mathbf{A}),\end{aligned}$$

where the last equality follows from (3). Using this and Theorem 9,

$$\begin{aligned}\mu_{r,2}(\mathbf{A},\mathbf{I}_d) &= \sum_{i=0}^{r}\sum_{j=0}^{1}\binom{r}{i}\binom{1}{j}(r-i-j+1)!2^{r-i-j+1}\operatorname{tr}(\mathbf{A}^{-(r-i)})\mu_{i,j}(\mathbf{A},\mathbf{I}_d)\\ &= \sum_{i=0}^{r}\bigg[\binom{r}{i}(r-i+1)!2^{r-i+1}\operatorname{tr}(\mathbf{A}^{-(r-i)})\mu_{i,0}(\mathbf{A},\mathbf{I}_d)\\ &\qquad + \binom{r}{i}(r-i)!2^{r-i}\operatorname{tr}(\mathbf{A}^{-(r-i)})\mu_{i,1}(\mathbf{A},\mathbf{I}_d)\bigg]\\ &= \sum_{i=0}^{r}\bigg[\frac{r!}{i!}(r-i+1)2^{r-i+1}\operatorname{tr}(\mathbf{A}^{-(r-i)})\mu_i(\mathbf{A}) + \frac{r!}{i!}2^{r-i}\operatorname{tr}(\mathbf{A}^{-(r-i)})(d+2i)\mu_i(\mathbf{A})\bigg]\\ &= \sum_{i=0}^{r}\frac{r!}{i!}[2(r-i+1)+d+2i]2^{r-i}\operatorname{tr}(\mathbf{A}^{-(r-i)})\mu_i(\mathbf{A})\\ &= (d+2r+2)\mu_{r,1}(\mathbf{A},\mathbf{I}_d)\\ &= (d+2r+2)(d+2r)\mu_r(\mathbf{A}).\end{aligned}$$

□

The proofs for the exact formulas for the MISE and AMISE for normal mixture densities are derived in the next two sections.

A.2.3 Proof of Theorem 7

Proof. From the proof of Theorem 5, we have that for $K=\phi$, $\mathbf{R}_{\phi*\phi,\mathbf{H},r}(f) = \mathbf{R}_{\phi,2\mathbf{H},r}(f)$ and $\mathbf{R}(\mathsf{D}^{\otimes r}f) = \mathbf{R}_{\phi,\mathbf{0},r}(f)$. Combining this with Theorem 1 and part *iv*) of Lemma 3 we come to

$$\operatorname{MISE}\{\widehat{(\mathsf{D}^{\otimes r}f}(\cdot;\mathbf{H})\} = 2^{-r}n^{-1}(4\pi)^{-d/2}|\mathbf{H}|^{-1/2}\mu_r(\mathbf{H}) + (1-n^{-1})\varpi_2 - 2\varpi_1 + \varpi_0,$$

where $\varpi_a = \operatorname{tr} \mathbf{R}_{\phi,a\mathbf{H},r}(f) = (\operatorname{vec}^T \mathbf{I}_{d^r}) \operatorname{vec} \mathbf{R}_{\phi,a\mathbf{H},r}(f)$. Now,

$$\begin{aligned}
\operatorname{vec} \mathbf{R}_{\phi,a\mathbf{H},r}(f) &= \operatorname{vec} \int \phi_{a\mathbf{H}} * \mathsf{D}^{\otimes r} f(\boldsymbol{x}) \mathsf{D}^{\otimes r} f(\boldsymbol{x})^T d\boldsymbol{x} \\
&= \sum_{\ell,\ell'=1}^{k} w_\ell w_{\ell'} \int \mathsf{D}^{\otimes r} \phi_{\boldsymbol{\Sigma}_{\ell'}}(\boldsymbol{x} - \boldsymbol{\mu}_{\ell'})(\boldsymbol{x}) \otimes \mathsf{D}^{\otimes r} \phi_{a\mathbf{H}+\boldsymbol{\Sigma}_\ell}(\boldsymbol{x} - \boldsymbol{\mu}_\ell) d\boldsymbol{x} \\
&= \sum_{\ell,\ell'=1}^{k} w_\ell w_{\ell'} (-1)^r \mathsf{D}^{\otimes 2r} \phi_{a\mathbf{H}+\boldsymbol{\Sigma}_\ell+\boldsymbol{\Sigma}_{\ell'}}(\boldsymbol{\mu}_\ell - \boldsymbol{\mu}_{\ell'})
\end{aligned}$$

so that we can write $\varpi_a = \boldsymbol{w}^T \boldsymbol{\Omega}_a \boldsymbol{w}$, where $\boldsymbol{\Omega}_a$ is the $k \times k$ matrix with (ℓ, ℓ') entry given by $(\boldsymbol{\Omega}_a)_{\ell,\ell'} = (-1)^r (\operatorname{vec}^T \mathbf{I}_{d^r}) \mathsf{D}^{\otimes 2r} \phi_{a\mathbf{H}+\boldsymbol{\Sigma}_\ell+\boldsymbol{\Sigma}_{\ell'}}(\boldsymbol{\mu}_\ell - \boldsymbol{\mu}_{\ell'})$.

The second expression of $(\boldsymbol{\Omega}_a)_{\ell,\ell'}$ is derived from the following identity in Holmquist (1996a):

$$\mathsf{D}^{\otimes 2r} \phi_{\boldsymbol{\Sigma}}(\boldsymbol{\mu}) = \phi_{\boldsymbol{\Sigma}}(\boldsymbol{\mu}) (\boldsymbol{\Sigma}^{-1})^{\otimes 2r} \boldsymbol{\mathcal{S}}_{d,2r} \sum_{j=0}^{r} (-1)^j \operatorname{OF}(2j) \binom{2r}{2j} \big(\boldsymbol{\mu}^{\otimes(2r-2j)} \otimes (\operatorname{vec} \boldsymbol{\Sigma})^{\otimes j}\big) \qquad (9)$$

and using $(\operatorname{vec}^T \mathbf{I}_{d^r})(\boldsymbol{\Sigma}^{-1})^{\otimes 2r} = \operatorname{vec}^T (\boldsymbol{\Sigma}^{-2})^{\otimes r}$. □

A.2.4 Proof of Theorem 8

Proof. To determine the AMISE formula, as we already have the integrated variance from Theorem 7, it suffices to find an expression for the asymptotic integrated squared bias

$$\begin{aligned}
\operatorname{AIB}^2(\mathbf{H}) &= \tfrac{1}{4} \operatorname{tr} \big[(\mathbf{I}_{d^r} \otimes \operatorname{vec}^T \mathbf{H}) \mathbf{R}(\mathsf{D}^{\otimes(r+2)} f) (\mathbf{I}_{d^r} \otimes \operatorname{vec} \mathbf{H}) \big] \\
&= \tfrac{1}{4} \operatorname{vec}^T \big(\mathbf{I}_{d^r} \otimes (\operatorname{vec} \mathbf{H} \operatorname{vec}^T \mathbf{H}) \big) \operatorname{vec} \mathbf{R}(\mathsf{D}^{\otimes(r+2)} f).
\end{aligned}$$

We can write

$$\operatorname{vec} \mathbf{R}(\mathsf{D}^{\otimes(r+2)} f) = \operatorname{vec} \mathbf{R}_{\phi,0\cdot\mathbf{H},r+2}(f) = \sum_{\ell,\ell'=1}^{k} w_\ell w_{\ell'} (-1)^{r+2} \mathsf{D}^{\otimes 2r+4} \phi_{\boldsymbol{\Sigma}_\ell+\boldsymbol{\Sigma}_{\ell'}}(\boldsymbol{\mu}_\ell - \boldsymbol{\mu}_{\ell'})$$

so that $4\operatorname{AIB}^2(\mathbf{H}) = \boldsymbol{w}^T \widetilde{\boldsymbol{\Omega}} \boldsymbol{w}$, with $\boldsymbol{\mu}_{\ell\ell'} = \boldsymbol{\mu}_\ell - \boldsymbol{\mu}_{\ell'}$, $\boldsymbol{\Sigma}_{\ell\ell'} = \boldsymbol{\Sigma}_\ell + \boldsymbol{\Sigma}_{\ell'}$,

$$\begin{aligned}
\widetilde{\boldsymbol{\Omega}}_{\ell,\ell'} &= (-1)^r \operatorname{vec}^T \big(\mathbf{I}_{d^r} \otimes (\operatorname{vec} \mathbf{H} \operatorname{vec}^T \mathbf{H}) \big) \mathsf{D}^{\otimes 2r+4} \phi_{\boldsymbol{\Sigma}_{\ell\ell'}}(\boldsymbol{\mu}_{\ell\ell'}) \\
&= (-1)^r \phi_{\boldsymbol{\Sigma}_{\ell\ell'}}(\boldsymbol{\mu}_{\ell\ell'}) \big[(\operatorname{vec}^T \mathbf{I}_d)^{\otimes r} \otimes (\operatorname{vec}^T \mathbf{H})^{\otimes 2} \big] (\boldsymbol{\Sigma}_{\ell\ell'}^{-1})^{\otimes(2r+4)} \boldsymbol{\mathcal{S}}_{d,2r+4} \\
&\quad \times \sum_{j=0}^{r+2} (-1)^j \operatorname{OF}(2j) \binom{2r+4}{2j} \big[\boldsymbol{\mu}_{\ell\ell'}^{\otimes(2r-2j+4)} \otimes (\operatorname{vec} \boldsymbol{\Sigma}_{\ell\ell'})^{\otimes j} \big] \\
&= (-1)^r \phi_{\boldsymbol{\Sigma}_{\ell\ell'}}(\boldsymbol{\mu}_{\ell\ell'}) \big[(\operatorname{vec}^T \boldsymbol{\Sigma}_{\ell\ell'}^{-2})^{\otimes r} \otimes (\operatorname{vec}^T (\boldsymbol{\Sigma}_{\ell\ell}^{-1} \mathbf{H} \boldsymbol{\Sigma}_{\ell\ell}^{-1}))^{\otimes 2} \big] \boldsymbol{\mathcal{S}}_{d,2r+4} \\
&\quad \times \sum_{j=0}^{r+2} (-1)^j \operatorname{OF}(2j) \binom{2r+4}{2j} \big[\boldsymbol{\mu}_{\ell\ell'}^{\otimes(2r-2j+4)} \otimes (\operatorname{vec} \boldsymbol{\Sigma}_{\ell\ell'})^{\otimes j} \big],
\end{aligned}$$

using (9). □

A.3 Proof of the results in Section 4

A.3.1 Proof of Corollary 3

Proof. From Theorem 6, $\mathbf{H}_{\text{AMISE}} = c_{\text{AMISE}}\boldsymbol{\Sigma}$ where $c_{\text{AMISE}} = \{4/[(d+2r+2)n]\}^{2/(d+2r+4)}$. Substituting this into the equation immediately following Eq. 8,

$$\begin{aligned}
\min_{\mathbf{H}\in\mathcal{F}} \text{AMISE}(\mathbf{H}) &= 2^{-(d+r)}\pi^{-d/2}|\boldsymbol{\Sigma}|^{-1/2}\mu_r(\boldsymbol{\Sigma})c^2_{\text{AMISE}}\big[n^{-1}c_{\text{AMISE}}^{-(d+2r+4)/2} + \tfrac{1}{16}(d+2r+2)(d+2r)\big] \\
&= 2^{-(d+r)}\pi^{-d/2}|\boldsymbol{\Sigma}|^{-1/2}\mu_r(\boldsymbol{\Sigma})\left(\frac{4}{d+2r+2}\right)^{4/(d+2r+4)} \\
&\quad \times \big[\tfrac{1}{4}(d+2r+2) + \tfrac{1}{16}(d+2r+2)(d+2r)\big]n^{-4/(d+2r+4)} \\
&= 2^{-(d+r+4)}\pi^{-d/2}(d+2r+2)(d+2r+4)\left(\frac{4}{d+2r+2}\right)^{4/(d+2r+4)} \\
&\quad \times |\boldsymbol{\Sigma}|^{-1/2}\mu_r(\boldsymbol{\Sigma})n^{-4/(d+2r+4)} \\
&= 2^{-(d+r+4)}2^{8/(d+2r+4)}\pi^{-d/2}(d+2r+4)(d+2r+2)^{(d+2r)/(d+2r+4)} \\
&\quad \times |\boldsymbol{\Sigma}|^{-1/2}\mu_r(\boldsymbol{\Sigma})n^{-4/(d+2r+4)}.
\end{aligned}$$

□

A.3.2 Proof of Corollary 4

Proof. Substituting $\mathbf{H} = h^2\mathbf{I}_d$ into the AMISE formula in Theorem 6,

$$\begin{aligned}
\text{AMISE}(h^2\mathbf{I}_d) &= 2^{-(d+r)}\pi^{-d/2}\{n^{-1}h^{-d-2r}\mu_r(\mathbf{I}_d) + \tfrac{1}{16}|\boldsymbol{\Sigma}|^{-1/2}\mu_{r,2}(\boldsymbol{\Sigma}, \boldsymbol{\Sigma}^{1/2}(h^{-2}\mathbf{I}_d)\boldsymbol{\Sigma}^{1/2})\} \\
&= 2^{-(d+r)}\pi^{-d/2}\{n^{-1}h^{-d-2r}\mu_r(\mathbf{I}_d) + \tfrac{1}{16}|\boldsymbol{\Sigma}|^{-1/2}\mu_{r+2}(\boldsymbol{\Sigma})h^4\}
\end{aligned}$$

since $\mu_{r,2}(\boldsymbol{\Sigma}, \boldsymbol{\Sigma}) = \mu_{r+2}(\boldsymbol{\Sigma})$. Differentiating with respect to h and setting to zero

$$-(d+2r)n^{-1}h^{-d-2r-1}\mu_r(\mathbf{I}_d) + \tfrac{1}{4}|\boldsymbol{\Sigma}|^{-1/2}\mu_{r+2}(\boldsymbol{\Sigma})h^3 = 0$$

gives

$$h_{\text{AMISE}} = \left(\frac{4(d+2r)|\boldsymbol{\Sigma}|^{1/2}\mu_r(\mathbf{I}_d)}{\mu_{r+2}(\boldsymbol{\Sigma})n}\right)^{1/(d+2r+4)}.$$

The minimal AMISE is

$$\begin{aligned}
&\min_{\mathbf{H}\in\mathcal{I}} \text{AMISE}(\mathbf{H}) \\
&\quad = 2^{-(d+r)}\pi^{-d/2} h_{\text{AMISE}}^4 \{n^{-1} h_{\text{AMISE}}^{-d-2r-4} + \tfrac{1}{16}|\boldsymbol{\Sigma}|^{-1/2}\mu_{r+2}(\boldsymbol{\Sigma})\} \\
&\quad = 2^{-(d+r)}\pi^{-d/2} \left(\frac{4(d+2r)|\boldsymbol{\Sigma}|^{1/2}\mu_r(\mathbf{I}_d)}{\mu_{r+2}(\boldsymbol{\Sigma})}\right)^{4/(d+2r+4)} \left(\frac{\mu_{r+2}(\boldsymbol{\Sigma})}{4(d+2r)|\boldsymbol{\Sigma}|^{1/2}} + \frac{\mu_{r+2}(\boldsymbol{\Sigma})}{16|\boldsymbol{\Sigma}|^{1/2}}\right) \\
&\qquad \times n^{-4/(d+2r+4)} \\
&\quad = 2^{-(d+r+4)}\pi^{-d/2}|\boldsymbol{\Sigma}|^{-1/2}\mu_{r+2}(\boldsymbol{\Sigma}) \left(\frac{d+2r+4}{d+2r}\right) \left(\frac{4(d+2r)|\boldsymbol{\Sigma}|^{1/2}\mu_r(\mathbf{I}_d)}{\mu_{r+2}(\boldsymbol{\Sigma})}\right)^{4/(d+2r+4)} \\
&\qquad \times n^{-4/(d+2r+4)} \\
&\quad = 2^{-(d+r+4)} 2^{8/(d+2r+4)} \pi^{-d/2} (d+2r+4)(d+2r)^{-1} \\
&\qquad \times \left\{|\boldsymbol{\Sigma}|^{-(d+2r)/2}\mu_{r+2}(\boldsymbol{\Sigma})^{d+2r}\mu_{r+1}(\mathbf{I}_d)^4\right\}^{1/(d+2r+4)} n^{-4/(d+2r+4)}
\end{aligned}$$

since $\mu_{r+1}(\mathbf{I}_d) = (d+2r)\mu_r(\mathbf{I}_d)$. □

A.3.3 Proof of Corollary 5

Proof. From Corollaries 3 and 4, the ARE is

$$\begin{aligned}
\text{ARE}(\mathcal{F}:\mathcal{I}) &= \frac{(d+2r+2)^{(d+2r)/4}|\boldsymbol{\Sigma}|^{-(d+2r+4)/8}\mu_r(\boldsymbol{\Sigma})^{(d+2r+4)/4}}{(d+2r)^{-(d+2r+4)/4}|\boldsymbol{\Sigma}|^{-(d+2r)/8}\mu_{r+2}(\boldsymbol{\Sigma})^{(d+2r)/4}\mu_{r+1}(\mathbf{I}_d)} \\
&= [(d+2r+2)(d+2r)]^{(d+2r)/4}|\boldsymbol{\Sigma}|^{-1/2}\mu_r(\boldsymbol{\Sigma})^{(d+2r+4)/4}\mu_{r+2}(\boldsymbol{\Sigma})^{-(d+2r)/4}\mu_r(\mathbf{I}_d)^{-1}
\end{aligned}$$

since $\mu_{r+1}(\mathbf{I}_d) = (d+2r)\mu_r(\mathbf{I}_d)$. □

A.3.4 Proof of Corollary 6

Proof. Let the variance be $\boldsymbol{\Sigma} = \sigma^2\boldsymbol{\Sigma}_\rho$ where $\boldsymbol{\Sigma}_\rho = \begin{bmatrix} 1 & \rho \\ \rho & 1 \end{bmatrix}$. With this form of the variance, using symmetry arguments, the bandwidth which minimizes the AMISE is in the form $\mathbf{H}_{\text{AMISE}} = c\mathbf{I}_2$, for some positive constant c. So we can apply Corollary 5. Thus

$$\begin{aligned}
\text{ARE}(\mathcal{F}:\mathcal{D}) &= \text{ARE}(\mathcal{F}:\mathcal{I}) \\
&= [4(r+1)(r+2)]^{(r+1)/2}|\boldsymbol{\Sigma}|^{-1/2}\mu_r(\boldsymbol{\Sigma})^{(r+3)/2}\mu_{r+2}(\boldsymbol{\Sigma})^{-(r+1)/2}\mu_r(\mathbf{I}_2)^{-1} \\
&\quad \times \mu_{r+2}(\boldsymbol{\Sigma})^{-(r+1)/2}\mu_r(\mathbf{I}_2)^{-1} \\
&= [4(r+1)(r+2)]^{(r+1)/2}\sigma^{-2}(1-\rho^2)^{-1/2}\sigma^{-r(r+3)}(1-\rho^2)^{-r(r+3)/2}Q(r,\rho)^{(r+3)/2} \\
&\quad \times \sigma^{(r+1)(r+2)}(1-\rho^2)^{(r+1)(r+2)/2}Q(r+2,\rho)^{-(r+1)/2}Q(r,0)^{-1} \\
&= \frac{(1-\rho^2)^{1/2}Q(r,\rho)^{(r+3)/2}}{Q(r,0)Q(r+2,\rho)^{(r+1)/2}}
\end{aligned}$$

where $Q(r,\rho) = (1-\rho^2)^r \mu_r(\mathbf{\Sigma}_\rho) = \sigma^{2r}(1-\rho^2)^r \mu_r(\mathbf{\Sigma})$. Since $\mathbf{\Sigma}^{-1} = \sigma^{-2}(1-\rho^2)^{-1}\begin{bmatrix} 1 & -\rho \\ -\rho & 1 \end{bmatrix}$, and if z_1, z_2 are independent standard normal random variables, then

$$\begin{aligned}
Q(r,\rho) &= \mathbb{E}\{(z_1^2 + z_2^2 - 2\rho z_1 z_2)^2\} \\
&= \mathbb{E}\left\{\sum_{j=0}^{r} \binom{r}{j} (z_1^2 - 2\rho z_1 z_2)^j (z_2)^{2(r-j)}\right\} \\
&= \mathbb{E}\left\{\sum_{j=0}^{r} \binom{r}{j} \sum_{j'=0}^{j} \binom{j}{j'} z_1^{2j'} (-2\rho z_1 z_2)^{j-j'} (z_2)^{2(r-j)}\right\} \\
&= \mathbb{E}\left\{\sum_{j=0}^{r} \binom{r}{j} \sum_{j'=0}^{j} \binom{j}{j'} (-2\rho)^{j-j'} z_1^{j+j'} z_2^{2r-j-j'}\right\} \\
&= \sum_{j=0}^{r} \sum_{j'=0}^{j} \binom{r}{j} \binom{j}{j'} (-2\rho)^{j-j'} m_{j+j'} m_{2r-j-j'}
\end{aligned}$$

where $m_k = \frac{1}{2}\{(-1)^k + 1\}\mathrm{OF}(k)$ is the kth central moment of a standard normal variable. □

References

Baxandall, P. and Liebeck, H. (1986) *Vector Calculus.* Oxford, Clarendon Press.

Bhattacharya, P.K. (1967) Estimation of a probability density function and its derivatives. *Sankhyā Ser. A* **29**, 373–382.

Brinkman, R.R, Gasparetto, M., Lee, S.-J.J., Ribickas, A.J., Perkins, J., Janssen, W., Smiley, R. and Smith, C. (2007). High-content flow cytometry and temporal data analysis for defining a cellular signature of graft-versus-host disease. *Biology of Blood and Marrow Transplantation* **13**, 691–700.

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

Chacón, J.E. and Duong, T. (2008) Multivariate plug-in bandwidth selection with unconstrained pilot bandwidth matrices. Submitted.

Ćwik, J. and Koronacki, J., (1997) A combined adaptive-mixtures/plug-in estimator of multivariate probability densities. *Comput. Statist. Data Anal.* **26**, 199–218.

Duong, T. and Hazelton, M.L. (2003) Plug-in bandwidth matrices for bivariate kernel density estimation *J. Nonparametr. Statist.* **15**, 17–30.

Duong, T., Cowling, A., Koch, I. and Wand, M.P. (2008) Feature significance for multivariate kernel density estimation. *Comput. Statist. Data Anal.* **52**, 4225–4242.

Gasser, T., Muller H-.G. and Mammitzsch, V. (1985) Kernels for nonparametric curve estimation. *J. Roy. Statist. Soc. Ser. B* **47**, 238–252.

Godtliebsen, F., Marron, J. S. and Chaudhuri, P. (2002) Significance in scale space for bivariate density estimation. *J. Comput. Graph. Statist.* **11**, 1–21.

Härdle, W., Marron, J. S. and Wand, M. P. (1990) Bandwidth choice for density derivatives. *J. Roy. Statist. Soc. Ser. B* **52**, 223–232.

Henderson, H.V. and Searle, S.R. (1979) Vec and vech operators for matrices, with some uses in Jacobians and multivariate statistics. *Canad. J. Statist.* **7**, 65–81.

Hildenbrand, K. and Hildenbrand, W. (1986) On the mean income effect: a data analysis of the U.K. family expenditure survey. In *Contributions to Mathematical Economics, in Honor of Gerard Debreu* (Edited by W. Hildenbrand and A. Mas-Colell), 247–268. Amsterdam, North-Holland.

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. *Stochastic Anal. Appl.* **14**, 149–164.

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

Magnus, J.R. (1986) The exact moments of a ratio of quadratic forms in normal variables. *Ann. Econom. Statist.* **4**, 95–109.

Le Meur, N., Rossini, A., Gasparetto, M., Smith, C., Brinkman, R.R. and Gentleman, R. (2007). Data quality assessment of ungated flow cytometry data in high throughput experiments. *Cytometry Part A* **71A**, 393–403.

Magnus, J.R. and Neudecker, H. (1979) The commutation matrix: some properties and applications. *Ann. Statist.* **7**, 381–394.

Magnus, J.R. and Neudecker, H. (1999) *Matrix Differential Calculus with Applications in Statistics and Econometrics*, revised edition. John Wiley & Sons, Chichester.

Naumann, U. and Wand, M.P. (2009) Automation in high-content flow cytometry screening. *Cytometry A*. In press.

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

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

Schuster, E.F (1969) Estimation of a probability function and its derivatives. *Ann. Math. Statist.* **40**, 1187–1195.

Shapiro, H.M. (2003) *Practical Flow Cytometry,* 4th edition. John Wiley & Sons, New York.

Singh, R.S. (1976) Nonparametric estimation of mixed partial derivatives of a multivariate density. *J. Multivariate Anal.* **6**, 111–122.

Singh, R.S. (1977) Applications of estimators of a density and its derivatives to certain statistical problems *J. Roy. Statist. Soc. Ser. B* **39**, 357-363.

Singh, R.S. (1979) Mean squared errors of estimates of a density and its derivatives. *Biometrika* **66**, 177–180.

Singh, R.S. (1987) MISE of kernel estimates of a density and its derivatives. *Statist. Prob. Letters* **5**, 153–159.

Stone, C.J. (1980) Optimal rates of convergence for nonparametric estimators *Ann. Statist.* **8**, 1348-1360.

Terrell, G.R. (1990) The maximal smoothing principle in density estimation. *J. Amer. Statist. Assoc.* **85**, 470–477.

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

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

Wand, M.P. and Jones, M.C. (1994) Multivariate plug-in bandwidth selection. *Comput. Stat.* **9**, 97–116.

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

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

Wu, T-.J. and Lin, Y. (2000) Information bound for bandwidth selection in kernel estimation of density derivatives. *Statistica Sinica* **10**, 457–473.