© Board of the Foundation of the Scandinavian Journal of Statistics 2005. Published by Blackwell Publishing Ltd, 9600 Garsington
Road, Oxford OX4 2DQ, UK and 350 Main Street, Malden, MA 02148, USA Vol 32: 485–506, 2005

Cross-validation Bandwidth Matrices for Multivariate Kernel Density Estimation

TARN DUONG and MARTIN L. HAZELTON

School of Mathematics and Statistics, University of Western Australia

ABSTRACT. The performance of multivariate kernel density estimates depends crucially on the choice of bandwidth matrix, but progress towards developing good bandwidth matrix selectors has been relatively slow. In particular, previous studies of cross-validation (CV) methods have been restricted to biased and unbiased CV selection of diagonal bandwidth matrices. However, for certain types of target density the use of full (i.e. unconstrained) bandwidth matrices offers the potential for significantly improved density estimation. In this paper, we generalize earlier work from diagonal to full bandwidth matrices, and develop a smooth cross-validation (SCV) methodology for multivariate data. We consider optimization of the SCV technique with respect to a pilot bandwidth matrix. All the CV methods are studied using asymptotic analysis, simulation experiments and real data analysis. The results suggest that SCV for full bandwidth matrices is the most reliable of the CV methods. We also observe that experience from the univariate setting can sometimes be a misleading guide for understanding bandwidth selection in the multivariate case.

Key words: asymptotic, biased, mean integrated squared error, pilot bandwidth, smooth cross-validation, unbiased

1. Introduction

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

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

where $\boldsymbol{x} = (x_1, x_2, \ldots, x_d)^{\mathrm{T}}$ and $\boldsymbol{X}_i = (X_{i1}, X_{i2}, \ldots, X_{id})^{\mathrm{T}}$, $i = 1, 2, \ldots, n$. Here, $K(\boldsymbol{x})$ is the multivariate kernel, which we assume to be a spherically symmetric probability density function; $\mathbf{H}$ is the bandwidth matrix, which is symmetric and positive-definite; and $K_{\mathbf{H}}(\boldsymbol{x}) = |\mathbf{H}|^{-1/2}K(\mathbf{H}^{-1/2}\boldsymbol{x})$. The estimator (1) is a useful tool for exploratory data analysis (especially for bivariate data when $\hat{f}$ can be visualized using the familiar perspective or contour plots), and also has applications in areas such as non-parametric discriminant analysis and goodness-of-fit testing. It is well known that the choice of $\mathbf{H}$ is crucial to the performance of $\hat{f}$. (for example, see Wand & Jones, 1993 and Simonoff, 1996). It follows that the study of data-driven methods for selecting $\mathbf{H}$ is not only important in its own right, but also because it sheds light on more general multivariate kernel smoothing problems.

The problem of selecting the scalar bandwidth in univariate kernel density estimation is quite well understood. A number of methods exist that combine good theoretical properties with strong practical performance (see Jones *et al.*, 1996). Many of these techniques can be extended to the multivariate case in a relatively straightforward fashion if $\mathbf{H}$ is constrained to be a diagonal matrix (see Wand & Jones 1994; Sain *et al.*, 1994). However, imposing such a constraint on the bandwidth matrix can result in markedly suboptimal density estimates, even if the data are pre-sphered (Wand & Jones, 1993; Duong & Hazelton 2003).

The preceding discussion indicates a need for data-driven methods for choosing full (i.e. unconstrained) bandwidth matrices. The development of selectors for full $\mathbf{H}$ is rather more challenging than that for diagonal $\mathbf{H}$. In particular, the need to consider the orientation of kernel functions to the coordinate axes in the former case introduces a problem without a univariate analogue. The additional difficulties in selecting full (as opposed to diagonal) bandwidth matrices partly explain the relatively slow progress in this area. Of the two major approaches to bandwidth selection, namely plug-in methods and cross-validation (CV) techniques, only the former has received any attention to date in the context of full bandwidth matrices. Wand & Jones (1994) outlined a plug-in selector that can be applied to full bandwidth matrices, but they concentrated on diagonal **H** when presenting methodological particulars. A more detailed account of plug-in selectors for full **H** was provided recently by Duong & Hazelton (2003).

In this paper, we consider CV selectors for full bandwidth matrices, including unbiased, biased and smoothed cross-validation (SCV) approaches. We describe the operation of each of these methods, and compare their performance using both asymptotic analyses and simulation experiments. More specifically, the following material is covered. In section 2 we provide necessary background on optimal bandwidth matrices, focusing on the mean integrated squared error (MISE) performance criterion and its asymptotic approximation. In the following section, we look at unbiased, biased and smoothed CV estimators of MISE. The performances of the corresponding full bandwidth matrix selectors are compared using asymptotic analyses in section 4. The analysis of the SCV selector is the most intricate, because it depends on the choice of pilot smoothing parameters. A significant body of theory exists for analogous pilot bandwidth selection problems in the univariate setting; see Jones *et al.* (1991), Hazelton (1996). However, many of the ideas do not transfer easily to the more difficult multivariate case, leading us to employ a number of simplifications en route to our development of a practical procedure for choosing the pilot bandwidth matrix. We find that SCV implemented in this fashion has better large sample performance than biased cross-validation (BCV) and unbiased cross-validation (UCV) in low dimensions, with the latter two methods turning out to have identical asymptotic rates. In section 5 we compare selectors of full and diagonal bandwidth matrices using numerical studies involving both simulated and real data. Results for the simulated bivariate data indicate that SCV selection of full bandwidth matrices is a good approach in practice, while UCV can perform surprisingly well for certain types of target density. We also see that BCV of full bandwidth matrices can be difficult to implement because of problems in optimizing over the relevant CV function. Results from simulated data in four and six dimensions again illustrate the attractions of SCV selection of full bandwidth matrices. These results also suggest that theoretical asymptotic advantages of UCV in high dimensions can translate into good practical performance for relatively modest sample sizes. Analysis of data on child mortality illustrates some of the differences in density estimates resulting from the different bandwidth matrix selectors, and also highlights the fact that intuition based on univariate kernel smoothing can be an unreliable guide to understanding the smoothing of multivariate data.

2. Optimal bandwidth matrices

It is usual to judge the performance of a bandwidth matrix **H** according to a global error criterion for $\hat{f}(\boldsymbol{x};\mathbf{H})$. Amongst such criteria, we prefer to work with MISE, given by

$$\begin{aligned}\mathrm{MISE}(\mathbf{H}) \equiv \mathrm{MISE}\hat{f}(\cdot;\mathbf{H}) &= \mathbb{E}\int_{\mathbb{R}^d}\left(\hat{f}(\boldsymbol{x};\mathbf{H}) - f(\boldsymbol{x})\right)^2 \mathrm{d}\boldsymbol{x}\\ &= \int_{\mathbb{R}^d}\mathrm{Bias}\{[\hat{f}(\boldsymbol{x};\mathbf{H})]\}^2\mathrm{d}\boldsymbol{x} + \int_{\mathbb{R}^d}\mathrm{Var}[\hat{f}(\boldsymbol{x};\mathbf{H})]\mathrm{d}\boldsymbol{x}.\end{aligned}$$

MISE is not mathematically tractable (except in special cases), so we employ a well known asymptotic approximation. The asymptotic mean integrated squared error (AMISE) of $\hat{f}$ is given by

$$\begin{aligned}\mathrm{AMISE}(\mathbf{H}) &\equiv \mathrm{AMISE}\hat{f}(\cdot;\mathbf{H})\\ &= \frac{1}{4}\mu_2(K)^2(\mathrm{vech}^{\mathrm{T}}\mathbf{H})\boldsymbol{\Psi}_4(\mathrm{vech}\,\mathbf{H}) + n^{-1}|\mathbf{H}|^{-1/2}R(K),\end{aligned} \tag{2}$$

where the first term on the right-hand side is the asymptotic version of integrated squared bias, and the second term is the asymptotic version of integrated variance. In (2), $R(K) = \int_{\mathbb{R}^d} K(\boldsymbol{x})^2 \mathrm{d}\boldsymbol{x} < \infty$, $\mu_2(K)\mathbf{I}_d = \int_{\mathbb{R}^d} \boldsymbol{x}\boldsymbol{x}^{\mathrm{T}}K(\boldsymbol{x})\mathrm{d}\boldsymbol{x}$ where $\mathbf{I}_d$ is the d dimensional identity matrix, $\mu_2(K) < \infty$, and vech is the vector half operator, so that vech $\mathbf{H}$ is the lower triangular half of $\mathbf{H}$ strung out column-wise into a vector (see Wand & Jones, 1995, chapter 4). The matrix $\boldsymbol{\Psi}_4$ is of dimensions $\frac{1}{2}d(d+1) \times \frac{1}{2}d(d+1)$, and is defined by

$$\boldsymbol{\Psi}_4 = \int_{\mathbb{R}^d} \mathrm{vech}(2D^2f(\boldsymbol{x}) - \mathrm{dg}\,D^2f(\boldsymbol{x}))\mathrm{vech}^{\mathrm{T}}(2D^2f(\boldsymbol{x}) - \mathrm{dg}\,D^2f(\boldsymbol{x}))\mathrm{d}\boldsymbol{x},$$

where $D^2f(\boldsymbol{x})$ is the Hessian matrix of f and dg $\mathbf{A}$ is matrix $\mathbf{A}$ with all of its non-diagonal elements set to zero. Sufficient conditions for the validity of the expansions defined by (2) are that all entries in $D^2f(\boldsymbol{x})$ are square integrable and all entries of $\mathbf{H} \to 0$ and $n^{-1}|\mathbf{H}|^{-1/2} \to 0$, as $n \to \infty$. The individual elements of $\boldsymbol{\Psi}_4$ can be written in terms of integrated density derivative functionals,

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

where $\boldsymbol{r} = (r_1, r_2, \ldots, r_d)$ for non-negative integers $r_1, r_2, \ldots, r_d$, $|\boldsymbol{r}| = r_1 + r_2 + \cdots + r_d$ and

$$f^{(\boldsymbol{r})}(\boldsymbol{x}) = \frac{\partial^{|\boldsymbol{r}|}}{\partial_{x_1}^{r_1}\partial_{x_2}^{r_2}\cdots\partial_{x_d}^{r_d}} f(\boldsymbol{x}).$$

In the bivariate case, for example,

$$\boldsymbol{\Psi}_4 = \begin{bmatrix} \psi_{40} & 2\psi_{31} & \psi_{22} \\ 2\psi_{31} & 4\psi_{22} & 2\psi_{13} \\ \psi_{22} & 2\psi_{13} & \psi_{04} \end{bmatrix}.$$

In this paper, we focus on full bandwidth matrix selectors which seek to estimate $\mathbf{H}_{\mathrm{MISE}}$, defined by

$$\mathbf{H}_{\mathrm{MISE}} = \mathrm{argmin}_{\mathbf{H}\in\mathcal{H}}\ \mathrm{MISE}\hat{f}(\cdot,\mathbf{H}), \tag{3}$$

where $\mathcal{H}$ is the set of $d \times d$ symmetric, positive-definite matrices. We will also have cause to consider diagonal bandwidth matrix selection, where optimization in (3) is over $\mathcal{D}$, the subset of diagonal matrices in $\mathcal{H}$. Mathematically it is simpler to work with an asymptotic version of $\mathbf{H}_{\mathrm{MISE}}$, namely

$$\mathbf{H}_{\mathrm{AMISE}} = \mathrm{argmin}_{\mathbf{H}\in\mathcal{H}}\ \mathrm{AMISE}\hat{f}(\cdot,\mathbf{H}).$$

As we shall see in section 4, the discrepancy between $\mathbf{H}_{\mathrm{MISE}}$ and $\mathbf{H}_{\mathrm{AMISE}}$ is asymptotically negligible in comparison with the random variation in the bandwidth matrix selectors that we consider. The problems of estimating $\mathbf{H}_{\mathrm{MISE}}$ and $\mathbf{H}_{\mathrm{AMISE}}$ are equivalent for most practical purposes, an observation that motivates some of the CV selectors described hereafter and facilitates their asymptotic analysis.

3. Unbiased, biased and smooth cross-validation

All CV bandwidth matrix selectors aim to estimate MISE, or AMISE, and some combine the two (modulo a constant) and then minimize the resulting function. UCV targets MISE and employs the objective function

$$\mathrm{UCV}(\mathbf{H}) = \int_{\mathbb{R}^d} \hat{f}(\boldsymbol{x};\mathbf{H})^2 \mathrm{d}\boldsymbol{x} - 2n^{-1}\sum_{i=1}^{n} \hat{f}_{-i}(\boldsymbol{X}_i;\mathbf{H}),$$

where

$$\hat{f}_{-i}(\boldsymbol{x};\mathbf{H}) = (n-1)^{-1}\sum_{\substack{j=1\\ j\neq i}}^{n} K_{\mathbf{H}}(\boldsymbol{x}-\boldsymbol{X}_j),$$

is a leave-one-out estimator of f. The function UCV(**H**) is unbiased in the sense that $\mathbb{E}[\mathrm{UCV}(\mathbf{H})] = \mathrm{MISE}[\hat{f}(\cdot;\mathbf{H})] - R(f)$. It can be expanded to give

$$\begin{aligned}\mathrm{UCV}(\mathbf{H}) &= n^{-2}\sum_{i=1}^{n}\sum_{j=1}^{n}(K_{\mathbf{H}} * K_{\mathbf{H}})(\boldsymbol{X}_i-\boldsymbol{X}_j) - 2n^{-1}(n-1)^{-1}\sum_{i=1}^{n}\sum_{\substack{j=1\\ j\neq i}}^{n} K_{\mathbf{H}}(\boldsymbol{X}_i-\boldsymbol{X}_j)\\ &= n^{-1}(n-1)^{-1}\sum_{i=1}^{n}\sum_{\substack{j=1\\ j\neq i}}^{n}(K_{\mathbf{H}} * K_{\mathbf{H}} - 2K_{\mathbf{H}})(\boldsymbol{X}_i-\boldsymbol{X}_j) + n^{-1}R(K)|\mathbf{H}|^{-1/2}, \qquad (4)\end{aligned}$$

where $*$ denotes a convolution. The UCV bandwidth matrix selector $\hat{\mathbf{H}}_{\mathrm{UCV}}$ is the minimizer of UCV(**H**).

BCV involves estimation of AMISE (and so is biased for MISE, hence the name). There are two versions of BCV, depending on the manner in which $\boldsymbol{\Psi}_4$ from (2) is estimated. The first version, which we denote BCV1, estimates AMISE$[\hat{f}(\cdot;\mathbf{H})]$ by the function

$$\mathrm{BCV1}(\mathbf{H}) = \frac{1}{4}\mu_2(K)^2(\mathrm{vech}^{\mathrm{T}}\mathbf{H})\check{\boldsymbol{\Psi}}_4(\mathrm{vech}\,\mathbf{H}) + n^{-1}R(K)|\mathbf{H}|^{-1/2} \qquad (5)$$

where

$$\check{\psi}_{\mathbf{r}}(\mathbf{H}) = n^{-2}\sum_{i=1}^{n}\sum_{\substack{j=1\\ j\neq i}}^{n}(K_{\mathbf{H}}^{(\boldsymbol{r})} * K_{\mathbf{H}})(\boldsymbol{X}_i-\boldsymbol{X}_j). \qquad (6)$$

The BCV2 method employs the objective function

$$\mathrm{BCV2}(\mathbf{H}) = \frac{1}{4}\mu_2(K)^2(\mathrm{vech}^{\mathrm{T}}\mathbf{H})\tilde{\boldsymbol{\Psi}}_4(\mathrm{vech}\,\mathbf{H}) + n^{-1}R(K)|\mathbf{H}|^{-1/2}, \qquad (7)$$

with

$$\tilde{\psi}_{\boldsymbol{r}}(\mathbf{H}) = n^{-1}\sum_{i=1}^{n}\hat{f}_{-i}^{(\boldsymbol{r})}(\boldsymbol{X}_i;\mathbf{H}) = n^{-1}(n-1)^{-1}\sum_{i=1}^{n}\sum_{\substack{j=1\\ j\neq i}}^{n} K_{\mathbf{H}}^{(\boldsymbol{r})}(\boldsymbol{X}_i-\boldsymbol{X}_j). \qquad (8)$$

The BCV selectors $\hat{\mathbf{H}}_{\mathrm{BCV1}}$ and $\hat{\mathbf{H}}_{\mathrm{BCV2}}$ are the minimizers of (5) and (7), respectively.

SCV can be thought of as a hybrid of UCV and BCV, in that it is based on explicit estimation of the exact integrated squared bias (like UCV) and the asymptotic integrated variance (like BCV). The SCV objective function is

$$\mathrm{SCV}(\mathbf{H}) = n^{-2}\sum_{i=1}^{n}\sum_{j=1}^{n}(K_{\mathbf{H}} * K_{\mathbf{H}} * L_{\mathbf{G}} * L_{\mathbf{G}} - 2K_{\mathbf{H}} * L_{\mathbf{G}} * L_{\mathbf{G}} + L_{\mathbf{G}} * L_{\mathbf{G}})(\boldsymbol{X}_i - \boldsymbol{X}_j) + n^{-1}R(K)|\mathbf{H}|^{-1/2}, \tag{9}$$

where $L_{\mathbf{G}}(\cdot)$ is the pilot kernel with pilot bandwidth matrix $\mathbf{G}$. If there are no replications in the data, then SCV with $\mathbf{G} = \mathbf{0}$ is identical to UCV (as $L_{\mathbf{0}}$ can be thought of as the Dirac delta function). We can therefore think of SCV as UCV applied to data presmoothed with $L_{\mathbf{G}}$. An alternative motivation for the squared bias component of SCV is provided by the smoothed bootstrap (Hall *et al.*, 1992). The SCV selector $\hat{\mathbf{H}}_{\mathrm{SCV}}$ is the minimizer of SCV($\mathbf{H}$).

The various CV functions simplify if we use a normal kernel, since the convolutions are then easy to evaluate. Specifically, if K and L are both set equal to the d-variate normal density with zero mean vector and identity covariance matrix, ϕ, then:

$$\mathrm{UCV}(\mathbf{H}) = n^{-1}(n-1)^{-1}\sum_{i=1}^{n}\sum_{\substack{j=1\\ j\neq i}}^{n}(\phi_{2\mathbf{H}} - 2\phi_{\mathbf{H}})(\boldsymbol{X}_i - \boldsymbol{X}_j) + n^{-1}(4\pi)^{-d/2}|\mathbf{H}|^{-1/2}, \tag{10}$$

where $\phi_{\mathbf{H}}(\boldsymbol{x}) = |\mathbf{H}|^{-1/2}\phi(\mathbf{H}^{-1/2}\boldsymbol{x})$ and

$$\mathrm{SCV}(\mathbf{H}) = n^{-2}\sum_{i=1}^{n}\sum_{j=1}^{n}(\phi_{2\mathbf{H}+2\mathbf{G}} - 2\phi_{\mathbf{H}+2\mathbf{G}} + \phi_{2\mathbf{G}})(\boldsymbol{X}_i - \boldsymbol{X}_j) + n^{-1}(4\pi)^{-d/2}|\mathbf{H}|^{-1/2}. \tag{11}$$

For the BCV methods, (8) and (10) become

$$\check{\psi}_{\boldsymbol{r}}(\mathbf{H}) = n^{-2}\sum_{i=1}^{n}\sum_{\substack{j=1\\ j\neq i}}^{n}(-1)^{|\boldsymbol{r}|}\phi_{2\mathbf{H}}^{(\boldsymbol{r})}(\boldsymbol{X}_i - \boldsymbol{X}_j)$$

and

$$\tilde{\psi}_{\boldsymbol{r}}(\mathbf{H}) = n^{-1}(n-1)^{-1}\sum_{i=1}^{n}\sum_{\substack{j=1\\ j\neq i}}^{n}\phi_{\mathbf{H}}^{(\boldsymbol{r})}(\boldsymbol{X}_i - \boldsymbol{X}_j),$$

respectively. Unless stated otherwise, we shall henceforth assume that all kernels are normal.

All of the methods described above are direct extensions of techniques for univariate density estimation (see Rudemo (1982) and Bowman (1984) for univariate UCV, Scott & Terrell (1987) and Jones & Kappenman (1992) for univariate BCV and Hall *et al.* (1992) for univariate SCV). Sain *et al.* (1994) investigated UCV, BCV1 and BCV2 for diagonal $\mathbf{H}$. These authors derived asymptotic properties of the methods, and compared their finite sample performance using a small simulation experiment. In the remainder of this paper we generalize Sain *et al.*'s (1994) asymptotic calculations for UCV and BCV to the case of full bandwidth matrices, examine the large sample behaviour of SCV, and conduct an empirical study of finite sample performance for each method.

4. Asymptotic analyses

4.1. Convergence rates for UCV and BCV selectors

The asymptotic performance of a bandwidth matrix selector can be assessed by its relative rate of convergence. The selector $\hat{\mathbf{H}}$ is said to converge to $\mathbf{H}_{\mathrm{AMISE}}$ with relative rate $n^{-\alpha}$ if

$$\text{vech}(\hat{\mathbf{H}} - \mathbf{H}_{\text{AMISE}}) = O_p(\mathbf{J}_{d'} n^{-\alpha})\text{vech } \mathbf{H}_{\text{AMISE}}, \tag{12}$$

where $\mathbf{J}_{d'}$ is the $d' \times d'$ matrix of ones and $d' = \frac{1}{2}d(d+1)$. Here the asymptotic order notation has been extended to matrices so that for sequences $(\mathbf{A}_n)$ and $(\mathbf{B}_n)$, $\mathbf{A}_n = o(\mathbf{B}_n)$ if $a_{ij} = o(b_{ij})$ for all elements a_{ij} of $\mathbf{A}_n$ and b_{ij} of $\mathbf{B}_n$. The definitions of O, o_p and O_p are similarly generalized to give an elementwise extension of the scalar order notation. The rationale for using $O_p(\mathbf{J}_{d'})$ rather than $O_p(\mathbf{I}_{d'})$ in (2) is to ensure that the relative rate in (2) remains well defined in cases where $\mathbf{H}_{\text{AMISE}}$ is diagonal but $\hat{\mathbf{H}}$ is not. For more details see Duong & Hazelton (2004). With this notation we can now state theorems 1 and 2, which provide the relative rates of convergence for UCV and BCV selectors of full bandwidth matrices.

Theorem 1

Assume that:

(A1) *All entries in $D^2 f(\boldsymbol{x})$ are bounded, continuous and square integrable.*
(A2) *All entries of $\mathbf{H} \to 0$ and $n^{-1}|\mathbf{H}|^{-1/2} \to 0$, as $n \to \infty$.*

Then the relative rate of convergence of the full bandwidth matrix selector $\hat{\mathbf{H}}_{\text{UCV}}$ to $\mathbf{H}_{\text{AMISE}}$ is $n^{-\min(d,4)/(2d+8)}$.

The proof of theorem 1 is collected together with other proofs in the appendix.

Theorem 2

Under the conditions of theorem 1 the relative rate of convergence of both full bandwidth matrix selectors $\hat{\mathbf{H}}_{\text{BCV1}}$ and $\hat{\mathbf{H}}_{\text{BCV2}}$ to $\mathbf{H}_{\text{AMISE}}$ is $n^{-\min(d,4)/(2d+8)}$.

This theorem is proved by Duong & Hazelton (2004).

The rates given in these theorems remain the same if $\mathbf{H}$ is constrained to be diagonal or even a scalar multiple of the identity matrix. Sain *et al.* (1994) give the rates diagonal bandwidth selectors (which agree with ours for $d \le 4$, but are incorrect for $d > 4$). The form of the rate changes after the fourth dimension because the squared bias of the BCV and UCV selectors then dominate the variance (see Duong & Hazelton (2004) for details). The rates in theorems 1 and 2 also apply when considering convergence to $\mathbf{H}_{\text{MISE}}$ (as opposed to $\mathbf{H}_{\text{AMISE}}$). This is because

$$\text{vech } (\mathbf{H}_{\text{AMISE}} - \mathbf{H}_{\text{MISE}}) = O(\mathbf{I}_{d'} n^{-2/(d+4)})\text{vech } \mathbf{H}_{\text{MISE}},$$

so that the discrepancy between $\mathbf{H}_{\text{AMISE}}$ and $\mathbf{H}_{\text{MISE}}$ is asymptotically negligible in comparison with the relative rate of convergence of $\hat{\mathbf{H}}_{\text{UCV}}, \hat{\mathbf{H}}_{\text{BCV1}}$ and $\hat{\mathbf{H}}_{\text{BCV2}}$ to $\mathbf{H}_{\text{AMISE}}$.

4.2. Asymptotic performance of SCV

The asymptotic properties of the SCV method depend crucially on the choice of pilot bandwidth matrix $\mathbf{G}$. Analysis and optimization of SCV for an unconstrained matrix $\mathbf{G}$ leads to substantial mathematical difficulties, mirroring those encountered by Wand & Jones (1994) in the context of plug-in selectors for bandwidth matrices. Like these authors, we circumvent the worst of the mathematical obstacles by imposing the constraint $\mathbf{G} = g^2\mathbf{I}$ for a scalar $g > 0$. Employment of such a restrictive pilot bandwidth matrix is only reasonable if the data are suitably transformed; we return to this matter when we discuss practical implementation in section 5. We noted in the introduction that the use of constrained bandwidth matrices can be markedly sub-optimal even if the data are pre-scaled or sphered, but experience from other kernel smoothing problems suggests that the deleterious effects of this type of restriction are less significant at the pilot stage than when applied to the final density estimate.

The problem of selecting a pilot bandwidth g for smoothed bootstrap bandwidth selection for univariate data has been studied by a number of authors, including Jones *et al.* (1991). For a univariate selector $\hat{h}$ these authors seek g so as to minimize the mean squared error

$$\mathrm{MSE}(\hat{h}) = \mathbb{E}[(\hat{h} - h_{\mathrm{AMISE}})]^2, \tag{13}$$

where h_{AMISE} is the minimizer of AMISE in the univariate case. Equation (13) could be extended to the multivariate setting in a number of ways. Our preference, based partly on mathematical convenience, is for the following performance criterion for g:

$$Q(g) = \sum_{i=1}^{d} \sum_{j=i}^{d} \mathbb{E}[(\hat{h}_{\mathrm{SCV},ij} - h_{\mathrm{AMISE},ij})^2],$$

where $\hat{h}_{\mathrm{SCV},ij}$ is the (i, j)th element of $\hat{\mathbf{H}}_{\mathrm{SCV}}$ and $h_{\mathrm{AMISE},ij}$ is the (i, j)th element of $\mathbf{H}_{\mathrm{AMISE}}$. We wish to find the minimizer g_0 of $Q(g)$. It turns out that the asymptotic form of $Q(g)$ is relatively simple, allowing us to write a formula for the leading term, g_1, in g_0.

Theorem 3

Suppose that the conditions for theorem 1 *hold, and that*

(S1) *f has bounded and continuous eighth-order partial derivatives*
(S2) *each element of $\Theta_6 = \int_{\mathbb{R}^d} (D^2)^3 f(\boldsymbol{x}) f(\boldsymbol{x}) d\boldsymbol{x}$ is finite*
(S3) *the sequence of pilot bandwidths $g = g_n$ satisfies $g^{-2}\mathbf{H} \to \mathbf{0}$* as $n \to \infty$.

Then $g_0 = g_1(1 + o(1))$ where

$$g_1 = \left\{ \frac{2(d+4)\boldsymbol{C}_{\mu_2}^{\mathrm{T}} \boldsymbol{C}_{\mu_2}}{\left[-(d+2)\boldsymbol{C}_{\mu_2}^{\mathrm{T}} \boldsymbol{C}_{\mu_1} + C_{\mu_0}^{1/2}\right] n} \right\}^{1/(d+6)}$$

in which

$$C_{\mu_0} = (d+2)^2 (\boldsymbol{C}_{\mu_2}^T \boldsymbol{C}_{\mu_1})^2 + 8(d+4)(\boldsymbol{C}_{\mu_1}^T \boldsymbol{C}_{\mu_1})(\boldsymbol{C}_{\mu_2}^T \boldsymbol{C}_{\mu_2})$$
$$\boldsymbol{C}_{\mu_1} = \frac{1}{2}\mathbf{D}_d^T \mathrm{vec}(\boldsymbol{\Theta}_6 \mathbf{C}_{\mathrm{AMISE}})$$
$$\boldsymbol{C}_{\mu_2} = \frac{1}{8}(4\pi)^{-d/2}[2\mathbf{D}_d^T \mathrm{vec}\ \mathbf{C}_{\mathrm{AMISE}} + (\mathrm{tr}\,\mathbf{C}_{\mathrm{AMISE}})\mathbf{D}_d^{\mathrm{T}}\ \mathrm{vec}\ \mathbf{I}_d]$$

with $\mathbf{C}_{\mathrm{AMISE}} = \mathbf{H}_{\mathrm{AMISE}} n^{2/(d+4)}$ and $\mathbf{D}_d$ being the duplication matrix of order d (*Magnus* & *Neudecker,* 1988).

We would like to apply SCV using the pilot bandwidth matrix $g_1^2\mathbf{I}$. Theorem 4 describes the asymptotic properties of $\hat{\mathbf{H}}_{\mathrm{SCV}}$ implemented in this fashion.

Theorem 4

Under the conditions of theorem 3, *the relative rate of convergence of $\hat{\mathbf{H}}_{\mathrm{SCV}}$ to $\mathbf{H}_{\mathrm{AMISE}}$ is $n^{-2/(d+6)}$ for $d \geq 2$.*

The proofs of theorems 3 and 4 are given in the appendix.

In practice, we do not know g_1 as it depends on functionals of f. Nonetheless, it can be shown that the rate of convergence theorem 4 continues to apply if these unknowns are estimated using a suitable, consistent plug-in estimator.

4.3. Comparison of bandwidth matrix selectors

In Table 1, we collect together the relative rates of convergence for our CV full bandwidth matrix selectors. The corresponding rates for Wand & Jones' (1994) and Duong & Hazelton's (2003) plug-in methods (labelled PI1 and PI2, respectively) are provided for comparison (see also Jones (1992). Of the full bandwidth matrix selectors, the plug-in method PI1 enjoys the best asymptotic performance, with PI2 and SCV only marginally worse for the practically important case of bivariate data. Comparable behaviour for these selectors has been found in the univariate case, see Wand & Jones (1995, pp. 81–84) and we expect this to carry over to the multivariate case because these selectors rely on similar bandwidth selection strategies: they both use asymptotic forms of the (A)MISE, use bias minimization rather than bias annihilation and use a single pilot bandwidth of order $n^{-1/(d+6)}$ and, as a result, have the same convergence rate. (Concerning the properties of plug-in selectors, see Duong & Hazelton (2004)). For the CV methods, UCV and BCV exhibit somewhat counter-intuitive behaviour in that their asymptotic properties improve with dimension up to $d = 4$. As a result, UCV and BCV have better large sample behaviour than SCV in the higher dimensions. While these comparisons are interesting from a theoretical standpoint, the practical implications are largely unclear. For instance, taking one of the most extreme comparisons, the relative rates for PI1 and UCV differ by a factor of only three for bivariate data when $n = 10{,}000$, a level of disparity that might easily be overturned by differences in the constant coefficients. We return to this in section 5.3.

5. Numerical results

5.1. Practical implementation of selectors

Both UCV and the BCV methods are implemented in a straightforward method by minimizing the relevant objective functions. Implementation of SCV is more complicated because we need to obtain an estimate of the pilot bandwidth g_1 from theorem 3. This requires estimation of $\Theta_6 = \int_{\mathbb{R}^d} (D^2)^3 f(\mathbf{x}) f(\mathbf{x})\, \mathrm{d}\mathbf{x}$. The elements of this matrix are linear combinations of integrated density derivative functionals $\psi_{\mathbf{r}}$ for various $\mathbf{r}$. For example, in the bivariate case

$$\boldsymbol{\Theta}_6 = \begin{bmatrix} \psi_{60} + 2\psi_{42} + \psi_{24} & \psi_{51} + 2\psi_{33} + \psi_{15} \\ \psi_{51} + 2\psi_{33} + \psi_{15} & \psi_{42} + 2\psi_{24} + \psi_{06} \end{bmatrix}.$$

We may estimate Θ_6 by employing plug-in estimation of $\psi_{\mathbf{r}}$ as described by Duong & Hazelton (2003). Our preference, used in the numerical examples in this paper, is for the one-step sum of asymptotic MSE (SAMSE) version of this methodology. The bandwidth selector $\hat{\mathbf{H}}_{\mathrm{SCV}}$ is obtained by minimizing (9) with $\mathbf{G} = \hat{g}_1^2 \mathbf{I}$, where $\hat{g}_1$ is the plug-in estimate of g_1. Because of the

Table 1. *Comparison of relative rates of convergence for bandwidth matrix selectors to* $\mathbf{H}_{\mathrm{AMISE}}$. $\hat{\mathbf{H}}_{\mathrm{BCV}}$ *rates apply to both BCV1 and BCV2 methodologies*

		Convergence rate to $\mathbf{H}_{\mathrm{AMISE}}$					
Selector	d	$d=1$	$d=2$	$d=3$	$d=4$	$d=5$	$d=6$
$\hat{\mathbf{H}}_{\mathrm{PI1}}$	$n^{-4/(d+12)}$	$n^{-4/13}$	$n^{-2/7}$	$n^{-4/15}$	$n^{-1/4}$	$n^{-4/17}$	$n^{-2/9}$
$\hat{\mathbf{H}}_{\mathrm{PI2}}$	$n^{-2/(d+6)}$	$n^{-2/7}$	$n^{-1/4}$	$n^{-1/10}$	$n^{-1/5}$	$n^{-2/11}$	$n^{-1/6}$
$\hat{\mathbf{H}}_{\mathrm{UCV}}$	$n^{-\min(d,4)/(2d+8)}$	$n^{-1/10}$	$n^{-1/6}$	$n^{-3/14}$	$n^{-1/4}$	$n^{-2/9}$	$n^{-1/5}$
$\hat{\mathbf{H}}_{\mathrm{BCV}}$	$n^{-\min(d,4)/(2d+8)}$	$n^{-2/9}$	$n^{-1/6}$	$n^{-3/14}$	$n^{-1/4}$	$n^{-2/9}$	$n^{-1/5}$
$\hat{\mathbf{H}}_{\mathrm{SCV}}$	$\begin{cases} n^{-5/14} & d = 1 \\ n^{-2/(d+6)} & d > 1 \end{cases}$	$n^{-5/14}$	$n^{-1/4}$	$n^{-2/9}$	$n^{-1/5}$	$n^{-2/11}$	$n^{-1/6}$
$\mathbf{H}_{\mathrm{AMISE}} - \mathbf{H}_{\mathrm{MISE}}$	$n^{-2/(d+4)}$	$n^{-2/5}$	$n^{-1/3}$	$n^{-2/7}$	$n^{-1/4}$	$n^{-2/9}$	$n^{-1/5}$

constrained form of **G**, SCV should be applied to suitably transformed data. We considered two such transformations: presphering, when SCV is computed for data $\mathbf{S}^{-1/2}\boldsymbol{x}$, where **S** is the sample covariance matrix for the original data; and prescaling, in which SCV is computed for data $\mathbf{S}_{\mathcal{D}}^{-1/2}\boldsymbol{x}$, where $\mathbf{S}_{\mathcal{D}} = \mathrm{diag}(S_1^2, S_2^2, \ldots, S_d^2)$ and S_i^2 is *i*th diagonal element of **S**. If $\hat{\mathbf{H}}_{\mathrm{SCV}}$ is the bandwidth selected on the transformed scale, then the SCV bandwidth matrix on the original scale is $\mathbf{S}^{1/2}\hat{\mathbf{H}}_{\mathrm{SCV}}\mathbf{S}^{1/2}$ if presphering were used and $\mathbf{S}_{\mathcal{D}}^{1/2}\hat{\mathbf{H}}_{\mathrm{SCV}}\mathbf{S}_{\mathcal{D}}^{1/2}$ if prescaling were used. Henceforth, we write SCV1 and SCV2 for SCV implemented using prescaling and presphering, respectively.

For all the CV methods an objective function must be numerically minimized. We employed a quasi-Newton minimization algorithm in the statistical software package R (Ihaka & Gentleman, 1996). The computation of the bandwidth matrix is carried out by a numerical optimization. The initial bandwidth is the normal reference bandwidth

$$\left[\frac{4}{n(d+2)}\right]^{2/(d+4)}\mathbf{S},$$

where **S** is the sample covariance matrix. We choose this as our initial condition, because it is easy to compute and that it is oversmooth for most cases. Following from Terrell (1990), it is recommended that an oversmooth bandwidth (e.g. normal reference or Terrell's maximally smoothed) be used as a starting point in order to avoid introducing spurious features. We then optimize over Ξ where $\mathbf{H} = \Xi\Xi^T$, so that if $\hat{\Xi} = \mathrm{argmin}_{\Xi}\,\widehat{\mathrm{AMISE}}(\Xi\Xi^T)$ then $\hat{\mathbf{H}} = \hat{\Xi}\hat{\Xi}^T$ is guaranteed to be positive definite. The question of multiple minima is solved in our case by a simple procedure which is satisfactory in the large majority of cases. Starting from the normal reference bandwidth, the numerical minimization routine is allowed to search until it finds a minimum. This is analogous to the common approach of using the largest local minimum (e.g. Scott, 1992; Marron, 1993).

5.2. Bivariate simulation study

We conducted a simulation experiment to investigate the performance of the following methods of bandwidth matrix selection: DPI1 (the PI1 plug-in selector for diagonal bandwidth matrices), UCV for diagonal bandwidth matrices (labelled DUCV), BCV2 for diagonal bandwidth matrices (labelled DBCV2), and PI2, UCV, BCV2, and SCV2 for full bandwidth matrices. (We also considered BCV1 and SCV1 in our original studies, but the results were little different from those for BCV2 and SCV2, respectively, and so they are omitted for the sake of brevity.) Each selector was implemented for sample sizes $n = 100$ and $n = 1000$ for each of the four target densities. These target densities were all bivariate normal mixtures with (globally) diagonal covariance matrices. Their contour plots are displayed in Fig. 1. Let $N(\mu_1, \mu_2; \sigma_1^2, \sigma_2^2, \rho)$ denote a multivariate normal density with mean $[\mu_1 \quad \mu_2]^{\mathrm{T}}$, variances σ_1^2 and σ_2^2 and correlation ρ. Then the functional forms are defined as follows: target density *A* is *N*(0,0; 0.25,1,0), density *B* is an equal mixture of *N*(1, 0; 4/9, 4/9,0) and *N*(−1, 0; 4/9, 4/9,0), density *C* is an equal mixture of *N*(1, −0.9; 1, 1, 0.9) and *N*(−1, 0.9; 1, 1, 0.9) and density *D* is an equal mixture of *N*(73/64, −5/6; 25/64, 25/64, 4/5), *N*(7/32, −5/3; 25/64, 25/64, −1/4) and $N(87/64, -5/6; 15/32, 5/8, -1/(4\sqrt{3}))$.

Computation times for typical simulations runs of bivariate data carried out on Pentium 4 2.40 GHz machines are given in Table 2. We see that the DBCV2 and BCV are more computationally intensive than the others by a large margin. (These selectors for the larger sample size were run for only 100 trials because of their extreme computational burden and so their times in the table have already been multiplied by four for comparability.) This reflects the

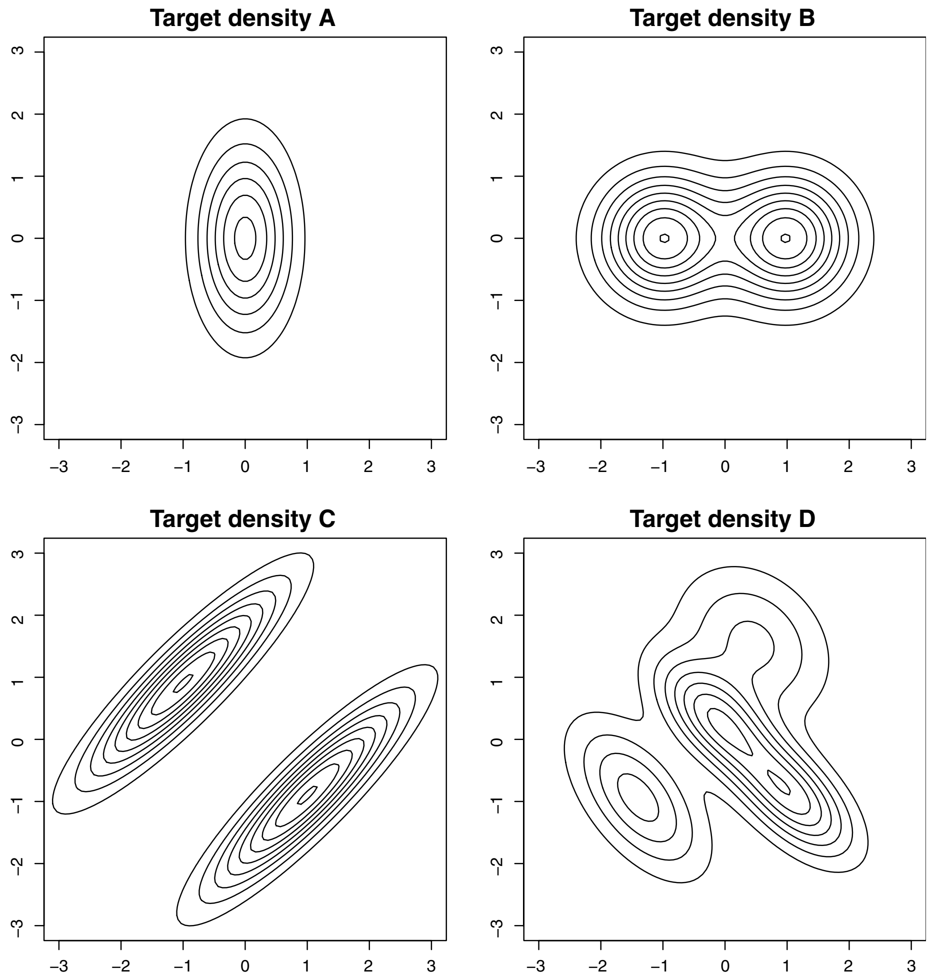

Fig. 1. Test densities.

Table 2. *Typical CPU times for 400 bivariate simulation trials (in seconds).*

	Selector						
	DPI1	DUCV	DBCV2	PI2	UCV	BCV2	SCV2
$n = 100$	628	964	2818	1612	436	10,788	1938
$n = 1000$	60,273	71,352	1,133,108	60,325	74,539	1,033,908	87,963

difficulty of finding the appropriate minimum of the BCV surface. The other selectors have fairly similar running times, with the diagonal ones usually requiring less time than their full matrix counterparts as the latter involve the selection of one extra parameter. From this table, we are able to infer that the full UCV, PI and SCV selectors do not impose computationally excessive burdens.

Bandwidth matrices were selected for 400 random samples generated from each combination of f and n. The integrated squared error (ISE) for each resulting density estimate was recorded. They are displayed in Fig. 2 using box plots with a log scale.

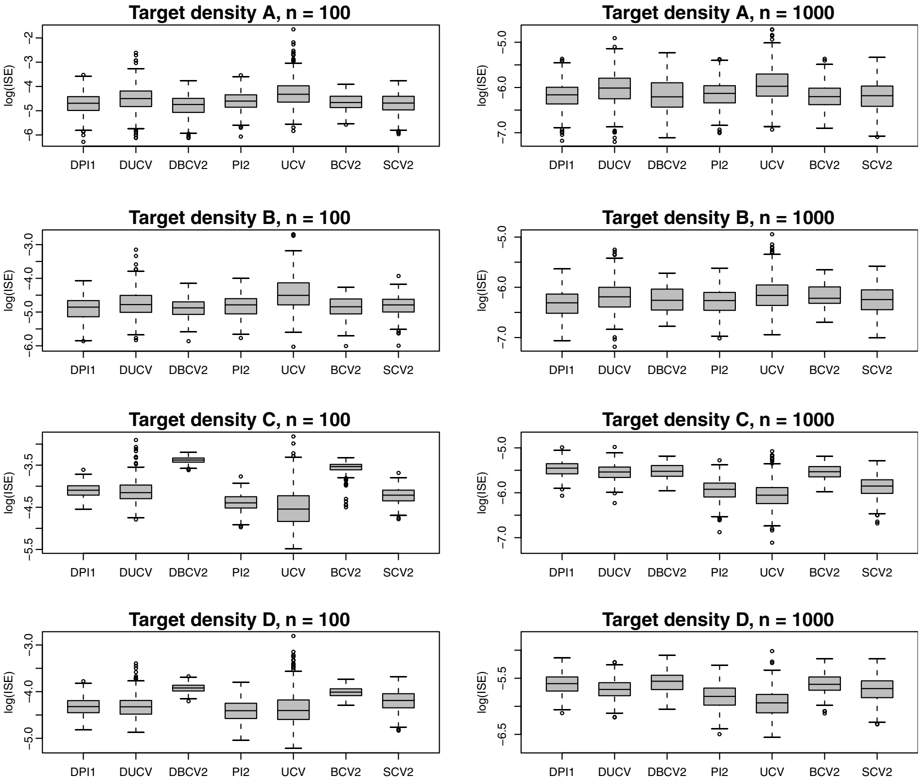

Fig. 2. Box plots of log(ISE) for bivariate bandwidth selectors.

As might be expected, no method is uniformly best. Nonetheless, a number of patterns do emerge. First, the diagonal matrix selectors perform adequately for densities *A* and *B*, but cannot compete with the full bandwidth methods on density *C* and to a lesser extent on density *D*. Prespheri ng the data does not help the diagonal matrix selectors since each test density is constructed to have zero correlation overall. The second noteworthy aspect of the results is the generally poor performance of the BCV methods. Not only are the results for BCV2 poor, but the BCV methods were difficult to implement. Numerical minimization of the BCV objective functions with respect to full bandwidth matrices was typically a slow process in comparison with the other methods, and sometimes required user intervention to overcome the convergence problems.

Turning to the other bandwidth selectors, UCV suffers in comparison with PI2 and SCV2 for densities *A* and *B*. For densities *C* and *D*, UCV produces rather variable results for $n = 100$ (as might be expected based on experience in the univariate setting) but performs surprisingly well at the larger sample size. The selectors PI2 and SCV2 are the most consistently reliable in the study, with the former slightly superior for density *D*. However, we conjecture that any small advantage that PI2 does enjoy over SCV2 is not an indication of any inherent superiority of plug-in over SCV methods. They are quite similar techniques, despite the nomenclature. Rather, it may be due to the fact that the current methods for choosing pilot bandwidths are slightly better developed for PI than SCV selectors.

5.3. *Multivariate simulation study*

Here, we look at a simulation study for data in two, four and six dimensions. For $d = 2, 4, 6$ the target density Ed is $N(\mathbf{0}_d, \Sigma_d)$ where $\mathbf{0}_d$ is a d-vector of zeros and Σ_d is a $d \times d$ matrix with ones along the diagonal and 0.9 for all off-diagonal elements. Notice that these densities are oriented at 45 degrees to the co-ordinate axes. Each simulation run consisted of 100 trials. The ISE plots are in Fig. 3. The performance of the full bandwidth matrices is better in all cases considered here (except for the six-variate DUCV selector, $n = 100$, which performs better than PI2). Moreover, the performance gain increases with increasing dimension: whereas, the gains obtained using a full bandwidth matrix in the bivariate case are relatively modest, the advantages are clearly more substantial for the four-variate and six-variate cases. Amongst the full bandwidth selectors, SCV2 has a clear advantage over PI2 in the higher dimensions. Also, the good asymptotic properties of UCV, as described in Table 1, translates into excellent practical performance despite the relatively modest sample sizes.

5.4. *Bivariate data analysis*

We now turn to the analysis of data on under-5 mortality (per 1000 live births) and average life expectancy (in years) for 73 countries with Gross National Income of < \$US 1000 per person per year. The data were obtained from Unicef (United Nations Children's Fund). Density estimates obtained using DUCV, DBCV2, UCV and SCV2 are displayed in Fig. 4. The selected bandwidth matrices are as follows:

$$\overset{\text{DUCV}}{\begin{bmatrix} 670.52 & 0 \\ 0 & 9.979 \end{bmatrix}} \quad \overset{\text{DBCV2}}{\begin{bmatrix} 1072.8 & 0 \\ 0 & 9.298 \end{bmatrix}} \quad \overset{\text{UCV}}{\begin{bmatrix} 388.2 & -83.34 \\ -83.34 & 25.13 \end{bmatrix}} \quad \overset{\text{SCV2}}{\begin{bmatrix} 1322.3 & -191.8 \\ -191.8 & 34.99 \end{bmatrix}}$$

The various bandwidth selectors provide a variety of interpretations of the structure of the data. In particular, DUCV and DBCV2 produce clear bimodality in the density estimate, while the SCV2 density estimate is unimodal. The density estimates for DUCV and DBCV2 do

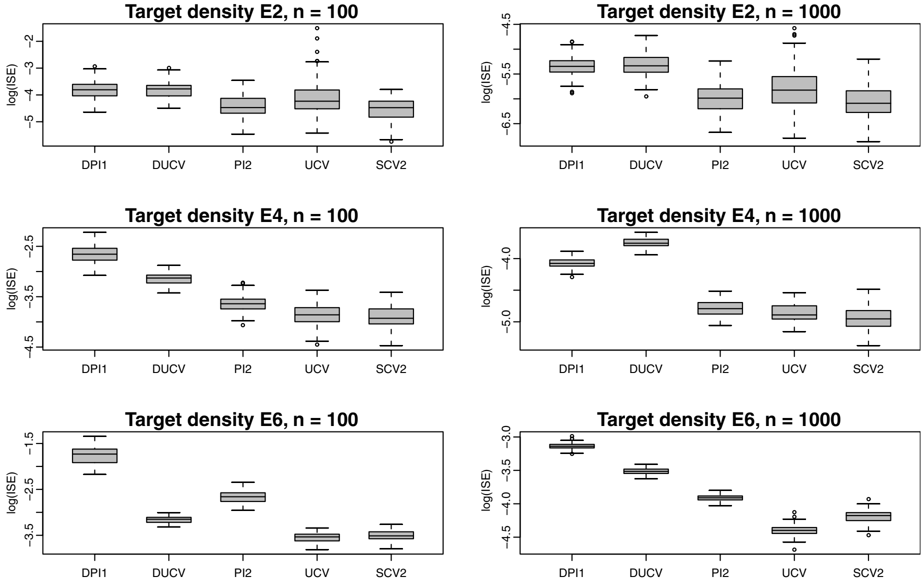

Fig. 3. Box plots of log (ISE) for multivariate bandwidth selectors.

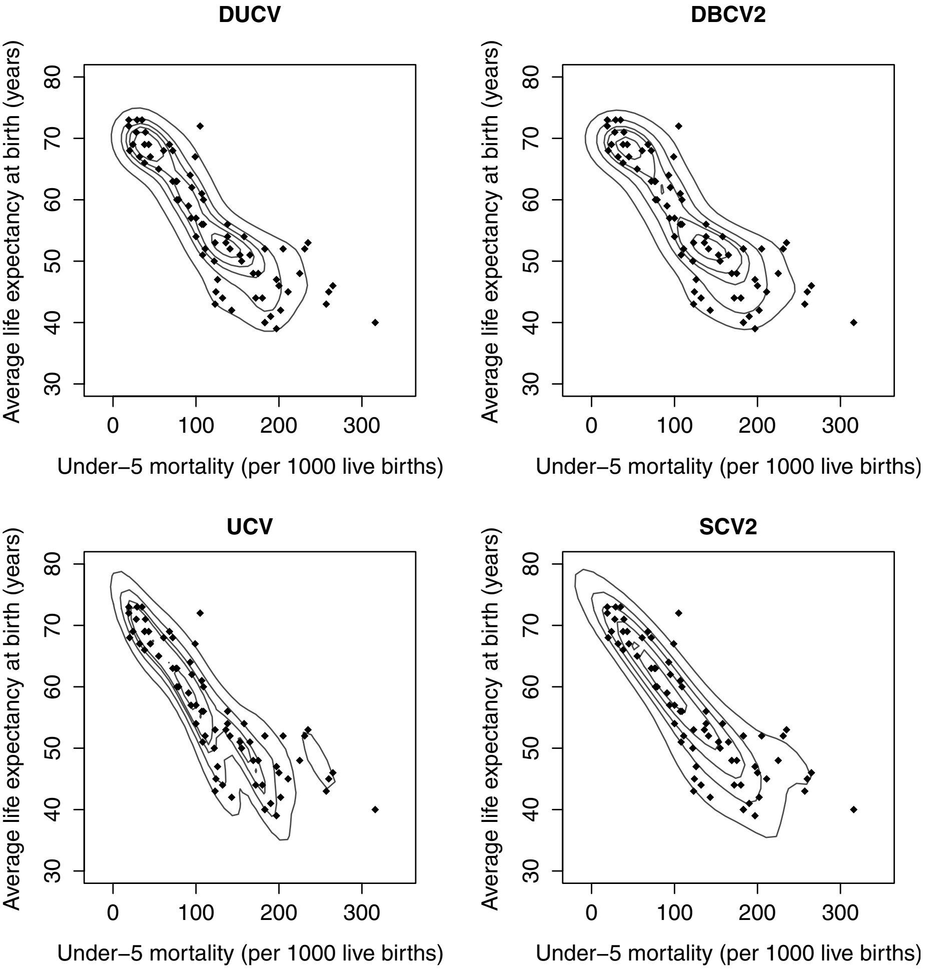

Fig. 4. Kernel density estimates from Unicef child mortality–life expectancy data, constructed using various cross-validation bandwidth matrices.

not display typical signs of undersmoothing (e.g. unwanted bumps in the tails and rough contours) and so experience from the univariate setting might suggest that the bimodal structure is no mere artefact, and that its absence from the SCV2 density estimate is a sign of oversmoothing in this latter case. However, on deeper reflection it is clear that for multivariate data oversmoothing in one coordinate direction can create additional modes. To illustrate this, we consider the artificial 'dumbbell' density given by the normal mixture $4/11N(-2, 2; 1, 1, 0) + 3/11N(0, 0; 0.8, 0.8, 0.9) + 4/11N(2, -2; 1, 1, 0)$, displayed in Fig. 5. This density has a single mode located at the 'bar' that connects the two flatter 'weights' (maintaining the dumbbell analogy). We generated a random sample of size 200 from this density and estimated densities using DBCV2 and SCV2 bandwidth matrices. The results are displayed in Fig. 6. For the DBCV2 density estimate the 'bar' is too wide and flat because of oversmoothing orthogonal to the orientation of the bar. It is this flattening of the bar that produces the bimodality. The SCV2 density estimate with its full bandwidth matrix is able to smooth the central, angled region appropriately and thus reproduce the unimodality of the target density.

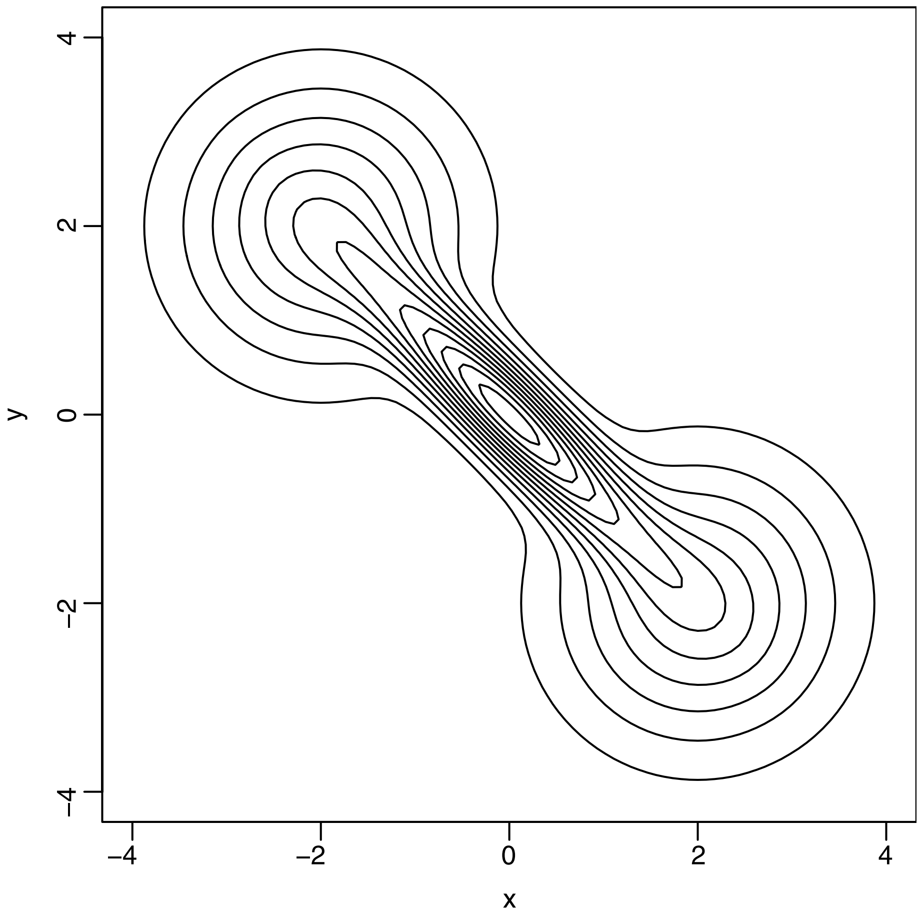

Fig. 5. Contour plot for ‘dumbbell’ density.

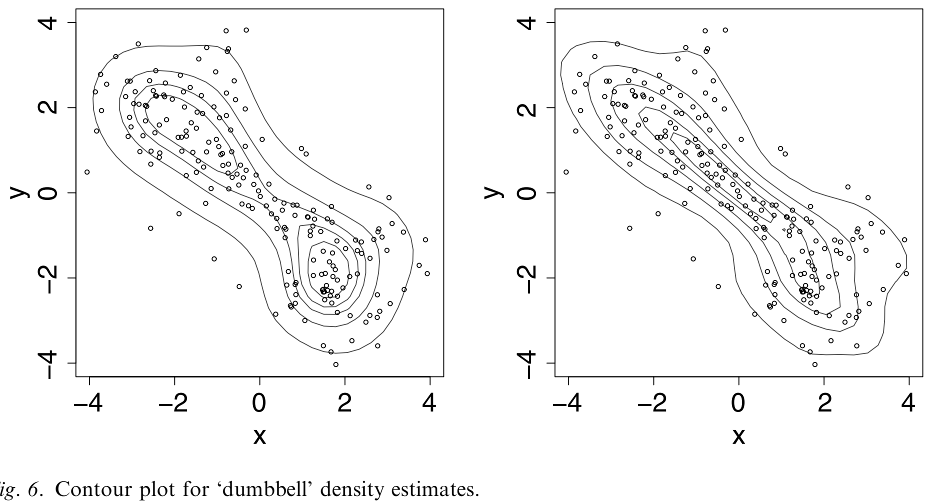

Fig. 6. Contour plot for ‘dumbbell’ density estimates.

6. Conclusions

Progress in bandwidth matrix selection for multivariate kernel density estimation has been relatively slow. CV selectors, for example, have received considerable attention in the univariate setting but very limited study in the multivariate case. In this paper, we have attempted to narrow this imbalance. Our contribution has been to generalize earlier work on UCV and BCV from diagonal to full bandwidth matrices, and to develop SCV methodology for multivariate data. Asymptotic analysis and simulation experiments suggest that SCV for full bandwidth matrices is the most reliable method amongst the CV selectors that we studied. For bivariate data our implementation of the SCV selector is reasonably comparable to the best plug-in methods currently available, while SCV enjoys an advantage for practical sample sizes in higher dimensions. Furthermore, it must be recognized that there is plenty of scope to refine our SCV methodology by developing more sophisticated techniques for selecting the pilot bandwidth matrix.

The selection of full bandwidth matrices for multivariate density estimation raises issues that have no univariate counterpart. In particular, the orientation of the kernel functions to the coordinate axes must be determined. Furthermore, intuition drawn from the univariate setting cannot necessarily be transferred to the multivariate case as we saw when smoothing the Unicef child mortality data. The additional difficulties involved in the multivariate case generate significant scope for further research in full bandwidth matrix selection. Looking further afield, the use of adaptive kernel density estimators has great potential for multivariate data. While some progress has been made, (e.g. Sain, 2002), this remains a challenging research direction.

Acknowledgements

The authors thank two anonymous referees for their helpful comments. Tarn Duong acknowledges the financial support of an Australian Postgraduate Award.

References

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

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

Duong, T. & Hazelton, M. L. (2004). Convergence rates for unconstrained bandwidth matrix selectors in multivariate kernel density estimation. *J. Multivariate Anal.* To appear.

Hall, P., Marron, J. & Park, B. (1992). Smoothed cross-validation. *Probab. Theory Related Fields* **92**, 1–20.

Hazelton, M. (1996). Bandwidth selection for local density estimators. *Scand. J. Statist.* **23**, 221–232.

Ihaka, R. & Gentleman, R. (1996). R: a language for data analysis & graphics. *J. Comput. Graph. Statist.* **5**, 299–314.

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

Jones, M. & Kappenman, R. (1992). On a class of kernel density estimate bandwidth selectors. *Scand. J. Statist.* **19**, 337–349.

Jones, M., Marron, J. & Park, B. (1991). A simple root *n* bandwidth selector. *Ann. Statist.* **19**, 1919–1932.

Jones, M., Marron, J. & Sheather, S. (1996). Progress in data-based bandwidth selection for kernel density estimation. *Comput. Statist.* **91**, 337–381.

Magnus, J. & Neudecker, H. (1988). *Matrix differential calculus with applications in statistics and econometrics*. John Wiley & Sons Ltd, Chichester.

Marron, J. (1993). Discussion of 'Practical performance of several data driven bandwidth selectors' by Park and Turlach. *Comput. Statist.* **8**, 17–19.

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

Sain, S. (2002). Multivariate locally adaptive density estimation. *Comput. Stat. Data Anal.* **39**, 165–186.

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

Scott, D. (1992). *Multivariate density estimation; theory, practice and visualization*. Wiley-Interscience, New York.

Scott, D. & Terrell, G. (1987). Biased and unbiased cross validation in density estimation. *J. Amer. Statist. Assoc.* **82**, 1131–1146.

Simonoff, J. (1996). *Smoothing methods in statistics*. Springer, New York.

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

Wand, M. & Jones, M. (1995). *Kernel smoothing*. Chapman & Hall, London.

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

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

Received March 2004, in final form November 2004

Corresponding author: Martin Hazelton, School of Mathematics and Statistics, University of Western Australia M019, 35 Stirling Highway, Crawley WA 6009, Australia. Email: martin@maths.uwa.edu.au

Appendix: proofs of theorems

Proof of theorem 1

A problem in finding relative rates for CV bandwidth matrix selectors is that these matrices are defined implicitly as the minimizers of various objective functions. The following result, proved by Duong and Hazelton (2004), is therefore useful.

Lemma 1

Let $\hat{\mathbf{H}} = \operatorname{argmin}_{\mathbf{H}} \widehat{\text{AMISE}}$ *be a bandwidth selector. Assume that:*

(A1) *All entries in* $D^2 f(\boldsymbol{X})$ *are bounded, continuous and square integrable.*

(A2) *All entries of* $\mathbf{H} \to 0$ *and* $n^{-1}|\mathbf{H}|^{-1/2} \to 0$, *as* $n \to \infty$.

(A3) *K is a spherically symmetric probability density.*

(A4) $(\widehat{\text{AMISE}}\,(\mathbf{H}) - \text{AMISE}\,(\mathbf{H}))/\text{AMISE}\,(\mathbf{H}) \xrightarrow{p} 0$ *as* $n \to \infty$.

Define the mean squared error (MSE) of $\hat{\mathbf{H}}$ by

$$\text{MSE}(\text{vech}\,\hat{\mathbf{H}}) = \mathbb{E}[\text{vech}\,(\hat{\mathbf{H}} - \mathbf{H}_{\text{AMISE}})\text{vech}^{\text{T}}(\hat{\mathbf{H}} - \mathbf{H}_{\text{AMISE}})].$$

Then

$$\text{MSE}\,(\text{vech}\,\hat{\mathbf{H}}) = \text{AMSE}\,(\text{vech}\,\hat{\mathbf{H}})(1 + o(1)),$$

where the asymptotic MSE can be written as

$$\text{AMSE}\,(\text{vech}\,\hat{\mathbf{H}}) = [\text{ABias}\,(\text{vech}\,\hat{\mathbf{H}})][\text{ABias}\,(\text{vech}\,\hat{\mathbf{H}})]^{\text{T}} + \text{AVar}\,(\text{vech}\,\hat{\mathbf{H}}),$$

in which

$$\begin{aligned}\text{ABias}(\text{vech}\,\hat{\mathbf{H}}) &= [D^2_{\mathbf{H}}\text{AMISE}(\mathbf{H}_{\text{AMISE}})]^{-1}\mathbb{E}[D_{\mathbf{H}}(\widehat{\text{AMISE}} - \text{AMISE})(\mathbf{H}_{\text{AMISE}})]\\ \text{AVar}(\text{vech}\,\hat{\mathbf{H}}) &= [D^2_{\mathbf{H}}\text{AMISE}(\mathbf{H}_{\text{AMISE}})]^{-1}\text{Var}[D_{\mathbf{H}}(\widehat{\text{AMISE}} - \text{AMISE})(\mathbf{H}_{\text{AMISE}})]\\ &\quad \times [D^2_{\mathbf{H}}\text{AMISE}(\mathbf{H}_{\text{AMISE}})]^{-1}.\end{aligned}$$

Here $D_{\mathbf{H}}$ *is the differential operator with respect to vech* $\mathbf{H}$ *and* $D^2_{\mathbf{H}}$ *is the corresponding Hessian operator.*

Lemma 2

*Assume A*1, *A*2 *and A*4 *of lemma* 1, *and* (*A*5)*: K is normal. Then*

$$\text{ABias}(\text{vech}\,\hat{\mathbf{H}}_{\text{UCV}}) = O(\mathbf{J}_{d'}n^{-2/(d+4)})\text{vech}\,\mathbf{H}_{\text{AMISE}}$$
$$\text{AVar}(\text{vech}\,\hat{\mathbf{H}}_{\text{UCV}}) = O(\mathbf{J}_{d'}n^{-d/(d+4)})(\text{vech}\,\mathbf{H}_{\text{AMISE}})(\text{vech}^T\mathbf{H}_{\text{AMISE}}).$$

Proof. A higher order expansion of the MISE is

$$\text{MISE}\,\hat{f}(\cdot;\mathbf{H}) = \text{AMISE}\,\hat{f}(\cdot;\mathbf{H}) + \frac{1}{8}\int_{\mathbb{R}^d}\text{tr}(\mathbf{H}D^2f(\boldsymbol{x}))\text{tr}(\mathbf{H}^2(\boldsymbol{D}^2)^2\boldsymbol{f}(\boldsymbol{x}))\text{d}\boldsymbol{x} + o(\|\text{vech}\,\mathbf{H}\|^3), \quad (14)$$

where D^2 is the Hessian operator with respect to the free variable x, so $(D^2)^2$ is obtained by 'multiplying' the Hessian operator with itself. This means that $(D^2)^2$ is a matrix of fourth-order partial differential operators.

As $\mathbb{E}\text{UCV}(\mathbf{H}) = \text{MISE}\hat{f}(\cdot;\mathbf{H}) - R(f)$ then

$$\begin{aligned}&\mathbb{E}[D_{\mathbf{H}}(\text{UCV} - \text{AMISE})(\mathbf{H})]\\ &= D_{\mathbf{H}}\left[-R(f) - \frac{1}{8}\int_{\mathbb{R}^d}\text{tr}(\mathbf{H}D^2f(\boldsymbol{x}))\text{tr}(\mathbf{H}^2(D^2)^2f(\boldsymbol{x}))\text{d}\boldsymbol{x} + o(\|\text{vech}\,\mathbf{H}\|^3)\right]\\ &= -\frac{1}{8}\int_{\mathbb{R}^d}\text{tr}(\mathbf{H}^2(D^2)^2f(\boldsymbol{x}))\mathbf{D}_d^{\text{T}}\text{vec}\,D^2f(\boldsymbol{x})\text{d}\boldsymbol{x}\\ &\quad -\frac{1}{4}\int_{\mathbb{R}^d}\text{tr}(\mathbf{H}D^2f(\boldsymbol{x}))\mathbf{D}_d^{\text{T}}\text{vec}(\mathbf{H}(D^2)^2f(\boldsymbol{x}))\text{d}\boldsymbol{x} + o(\|\text{vech}\,\mathbf{H}\|\text{vech}\,\mathbf{H})\end{aligned}$$

as $D_{\mathbf{H}}\text{tr}(\mathbf{AH}) = \mathbf{D}_d^{\text{T}}\text{vec}\,\mathbf{A}$ and $D_{\mathbf{H}}\text{tr}(\mathbf{AH}^2) = \mathbf{D}_d^{\text{T}}\text{vec}(\mathbf{HA})$ for a matrix $\mathbf{A}$ of appropriate dimensions. So ABias(vech $\hat{\mathbf{H}}_{\text{UCV}}$) is $O(\mathbf{J}_{d'}n^{-2/(d+4)})$ vech $\mathbf{H}_{\text{AMISE}}$.

For the asymptotic variance,

$$\begin{aligned}&\text{Var}\,[D_{\mathbf{H}}(\text{UCV} - \text{AMISE})(\mathbf{H}_{\text{AMISE}})]\\ &= \text{Var}\,[D_{\mathbf{H}}\text{UCV}\,(\mathbf{H}_{\text{AMISE}})]\\ &= \text{Var}\left[n^{-1}(n-1)^{-1}\sum_{i=1}^{n}\sum_{\substack{j=1\\ j\neq i}}^{n}D_{\mathbf{H}}(\phi_{2\mathbf{H}} - 2\phi_{\mathbf{H}})(\boldsymbol{X}_i - \boldsymbol{X}_j)\right]\\ &= \text{Var}\left[n^{-2}\sum_{i=1}^{n}\sum_{\substack{j=1\\ j\neq i}}^{n}(\boldsymbol{\varphi}_{2\mathbf{H}} - \boldsymbol{\varphi}_{\mathbf{H}})(\boldsymbol{X}_i - \boldsymbol{X}_j)\right][1 + o(n^{-1})]\\ &= 2n^{-2}\text{Var}[(\boldsymbol{\varphi}_{2\mathbf{H}} - \boldsymbol{\varphi}_{\mathbf{H}})(\boldsymbol{X}_1 - \boldsymbol{X}_2)]\\ &\quad + 4n^{-1}\text{Cov}[(\boldsymbol{\varphi}_{2\mathbf{H}} - \boldsymbol{\varphi}_{\mathbf{H}})(\boldsymbol{X}_1 - \boldsymbol{X}_2), (\boldsymbol{\varphi}_{2\mathbf{H}} - \boldsymbol{\varphi}_{\mathbf{H}})(\boldsymbol{X}_2 - \boldsymbol{X}_3)],\end{aligned}$$

where

$$\boldsymbol{\varphi}_{a\mathbf{H}}(\boldsymbol{X}) = \phi_{a\mathbf{H}}(\boldsymbol{X})\mathbf{D}_d^{\text{T}}\text{vec}[(a\mathbf{H})^{-1}\boldsymbol{X}\boldsymbol{X}^{\text{T}}(a\mathbf{H})^{-1} - (a\mathbf{H})^{-1}]$$

and $D_{\mathbf{H}}\phi_{a\mathbf{H}}(\boldsymbol{x}) = \frac{1}{2}a\boldsymbol{\varphi}_{a\mathbf{H}}(\boldsymbol{x})$.

The first term of Var[$D_{\mathbf{H}}$(UCV −AMISE)(**H**)] is

$$\begin{aligned}
&\mathrm{Var}[(\boldsymbol{\varphi}_{2\mathbf{H}} - \boldsymbol{\varphi}_{\mathbf{H}})(\boldsymbol{X}_1 - \boldsymbol{X}_2)] \\
&= \mathbb{E}\Big\{[(\boldsymbol{\varphi}_{2\mathbf{H}} - \boldsymbol{\varphi}_{\mathbf{H}})(\boldsymbol{X}_1 - \boldsymbol{X}_2)][(\boldsymbol{\varphi}_{2\mathbf{H}} - \boldsymbol{\varphi}_{\mathbf{H}})(\boldsymbol{X}_1 - \boldsymbol{X}_2)]^{\mathrm{T}}\Big\} \\
&\quad - [\mathbb{E}(\boldsymbol{\varphi}_{2\mathbf{H}} - \boldsymbol{\varphi}_{\mathbf{H}})(\boldsymbol{X}_1 - \boldsymbol{X}_2)][\mathbb{E}(\boldsymbol{\varphi}_{2\mathbf{H}} - \boldsymbol{\varphi}_{\mathbf{H}})(\boldsymbol{X}_1 - \boldsymbol{X}_2)]^{\mathrm{T}}.
\end{aligned}$$

Also,

$$\begin{aligned}
&\mathbb{E}(\boldsymbol{\varphi}_{2\mathbf{H}} - \boldsymbol{\varphi}_{\mathbf{H}})(\boldsymbol{X}_1 - \boldsymbol{X}_2) \\
&= D_{\mathbf{H}}[\mathbb{E}(\phi_{2\mathbf{H}} - 2\phi_{\mathbf{H}})(\boldsymbol{X}_1 - \boldsymbol{X}_2)] \\
&= D_{\mathbf{H}}\left[- R(f) + \frac{1}{4}\int_{\mathbb{R}^d} \mathrm{tr}(\mathbf{H}^2(D^2)^2 f(\boldsymbol{y}))f(\boldsymbol{y})\mathrm{d}\boldsymbol{y} + o(\|\mathrm{vech}\mathbf{H}\|^2)\right] \\
&= \frac{1}{2}\int_{\mathbb{R}^d} \mathbf{D}_d^{\mathrm{T}}\mathrm{vec}(\mathbf{H}(D^2)^2 f(\boldsymbol{y}))f(\boldsymbol{y})\mathrm{d}\boldsymbol{y} + o(\mathrm{vech}\ \mathbf{H}).
\end{aligned}$$

Next, $\mathbb{E}\{(\varphi_{2\mathbf{H}} - \varphi_{\mathbf{H}})(\boldsymbol{X}_1 - \boldsymbol{X}_2)[(\varphi_{\mathbf{H}} - \varphi_{\mathbf{H}})(\boldsymbol{X}_1 - \boldsymbol{X}_2)]^{\mathrm{T}}\}$ contains expressions of the type

$$\begin{aligned}
&\mathbb{E}\{\phi_{a\mathbf{H}}(\boldsymbol{X}_1 - \boldsymbol{X}_2)\mathbf{D}_d^{\mathrm{T}}\mathrm{vec}\ [(a\mathbf{H})^{-1}(\boldsymbol{X}_1 - \boldsymbol{X}_2)(\boldsymbol{X}_1 - \boldsymbol{X}_2)^{\mathrm{T}}(a\mathbf{H})^{-1} - (a\mathbf{H})^{-1}] \\
&\quad \times \phi_{b\mathbf{H}}(\boldsymbol{X}_1 - \boldsymbol{X}_2)\mathrm{vec}^{\mathrm{T}}[(b\mathbf{H})^{-1}(\boldsymbol{X}_1 - \boldsymbol{X}_2)(\boldsymbol{X}_1 - \boldsymbol{X}_2)^{\mathrm{T}}(b\mathbf{H})^{-1} - (b\mathbf{H})^{-1}]\mathbf{D}_d\}. \qquad (15)
\end{aligned}$$

As $\phi_{a\mathbf{H}}(\boldsymbol{x})\phi_{b\mathbf{H}}(\boldsymbol{x}) = (2\pi)^{-d/2}|(a + b)\mathbf{H}|^{-1/2}\phi_{a'\mathbf{H}}(\boldsymbol{x})$, where $a' = ab/(a + b)$ then we can simplify (15) in the following manner:

$$\begin{aligned}
&\mathbb{E}\{\phi_{a\mathbf{H}}(\boldsymbol{X}_1 - \boldsymbol{X}_2)\mathbf{D}_d^{\mathrm{T}}\mathrm{vec}[(a\mathbf{H})^{-1}(\boldsymbol{X}_1 - \boldsymbol{X}_2)(\boldsymbol{X}_1 - \boldsymbol{X}_2)^{\mathrm{T}}(a\mathbf{H})^{-1} - (a\mathbf{H})^{-1}] \\
&\quad \times \phi_{b\mathbf{H}}(\boldsymbol{X}_1 - \boldsymbol{X}_2)\mathrm{vec}^{\mathrm{T}}[(b\mathbf{H})^{-1}(\boldsymbol{X}_1 - \boldsymbol{X}_2)(\boldsymbol{X}_1 - \boldsymbol{X}_2)^{\mathrm{T}}(b\mathbf{H})^{-1} - (b\mathbf{H})^{-1}]\mathbf{D}_d\} \\
&= O(\mathbf{J}_{d'}|\mathbf{H}|^{-1/2})\int_{\mathbb{R}^{2d}} \phi_{a/\mathbf{H}}(\boldsymbol{x} - \boldsymbol{y})\mathbf{D}_d^{\mathrm{T}}\mathrm{vec}[(a\mathbf{H})^{-1}(\boldsymbol{x} - \boldsymbol{y})(\boldsymbol{x} - \boldsymbol{y})^{\mathrm{T}}(a\mathbf{H})^{-1} - (a\mathbf{H})^{-1}] \\
&\quad \times \mathrm{vec}^{\mathrm{T}}[(b\mathbf{H})^{-1}(\boldsymbol{x} - \boldsymbol{y})(\boldsymbol{x} - \boldsymbol{y})^{\mathrm{T}}(b\mathbf{H})^{-1} - (b\mathbf{H})^{-1}]\mathbf{D}_d f(\boldsymbol{x})f(\boldsymbol{y})\,\mathrm{d}\boldsymbol{x}\mathrm{d}\boldsymbol{y} \\
&= O(\mathbf{J}_{d'}|\mathbf{H}|^{-1/2})(\mathrm{vech}\ \mathbf{H}^{-1})(\mathrm{vech}^{\mathrm{T}}\mathbf{H}^{-1}).
\end{aligned}$$

This term is dominant over $[\mathbb{E}(\varphi_{2\mathbf{H}} - \varphi_{\mathbf{H}})(\boldsymbol{X}_1 - \boldsymbol{X}_2)][\mathbb{E}(\varphi_{2\mathbf{H}} - \varphi_{\mathbf{H}})(\boldsymbol{X}_1 - \boldsymbol{X}_2)]^{\mathrm{T}}$, which is $O(\mathbf{J}_{d'})(\mathrm{vech}\ \mathbf{H})(\mathrm{vech}^{\mathrm{T}}\mathbf{H})$ so

$$\mathrm{Var}[(\boldsymbol{\varphi}_{2\mathbf{H}} - \boldsymbol{\varphi}_{\mathbf{H}})(\boldsymbol{X}_1 - \boldsymbol{X}_2)] = O(\mathbf{J}_{d'}|\mathbf{H}|^{-1/2})(\mathrm{vech}\ \mathbf{H}^{-1})(\mathrm{vech}^{\mathrm{T}}\ \mathbf{H}^{-1}). \qquad (16)$$

The second term of Var[$D_{\mathbf{H}}$(UCV − AMISE)(**H**)] is

$$\begin{aligned}
&\mathrm{Cov}[(\boldsymbol{\varphi}_{2\mathbf{H}} - \boldsymbol{\varphi}_{\mathbf{H}})(\boldsymbol{X}_1 - \boldsymbol{X}_2), (\boldsymbol{\varphi}_{2\mathbf{H}} - \boldsymbol{\varphi}_{\mathbf{H}})(\boldsymbol{X}_2 - \boldsymbol{X}_3)] \\
&= \mathbb{E}\Big\{[(\boldsymbol{\varphi}_{2\mathbf{H}} - \boldsymbol{\varphi}_{\mathbf{H}})(\boldsymbol{X}_1 - \boldsymbol{X}_2)][(\boldsymbol{\varphi}_{2\mathbf{H}} - \boldsymbol{\varphi}_{\mathbf{H}})(\boldsymbol{X}_2 - \boldsymbol{X}_3)]^{\mathrm{T}}\Big\} \\
&\quad - [\mathbb{E}(\boldsymbol{\varphi}_{2\mathbf{H}} - \boldsymbol{\varphi}_{\mathbf{H}})(\boldsymbol{X}_1 - \boldsymbol{X}_2)][\mathbb{E}(\boldsymbol{\varphi}_{2\mathbf{H}} - \boldsymbol{\varphi}_{\mathbf{H}})(\boldsymbol{X}_2 - \boldsymbol{X}_3)]^{\mathrm{T}}.
\end{aligned}$$

We have already derived an order expression for the latter term in this covariance. The former term $\mathbb{E}\{(\varphi_{2\mathbf{H}} - \varphi_{\mathbf{H}})(\boldsymbol{X}_1 - \boldsymbol{X}_2)[(\varphi_{2\mathbf{H}} - \varphi_{\mathbf{H}})(\boldsymbol{X}_2 - \boldsymbol{X}_3)]^{\mathrm{T}}\}$ contains expressions of the type

$$\begin{aligned}
&\mathbb{E}\{\phi_{a\mathbf{H}}(\boldsymbol{X}_1 - \boldsymbol{X}_2)\mathbf{D}_d^{\mathrm{T}}\mathrm{vec}[(a\mathbf{H})^{-1}(\boldsymbol{X}_1 - \boldsymbol{X}_2)(\boldsymbol{X}_1 - \boldsymbol{X}_2)^{\mathrm{T}}(a\mathbf{H})^{-1} - (a\mathbf{H})^{-1}] \\
&\quad \times \phi_{b\mathbf{H}}(\boldsymbol{X}_2 - \boldsymbol{X}_3)\mathrm{vec}^{\mathrm{T}}[(b\mathbf{H})^{-1}(\boldsymbol{X}_2 - \boldsymbol{X}_3)(\boldsymbol{X}_2 - \boldsymbol{X}_3)^{\mathrm{T}}(b\mathbf{H})^{-1} - (b\mathbf{H})^{-1}]\mathbf{D}_d\}.
\end{aligned}$$

We can simplify this expression:

$$
\begin{aligned}
&\int_{\mathbb{R}^{3d}} \phi_{a\mathbf{H}}(\boldsymbol{x}-\boldsymbol{y})\mathbf{D}_d^{\mathrm{T}}\mathrm{vec}[(a\mathbf{H})^{-1}(\boldsymbol{x}-\boldsymbol{y})(\boldsymbol{x}-\boldsymbol{y})^{\mathrm{T}}(a\mathbf{H})^{-1}-(a\mathbf{H})^{-1}] \\
&\quad \times \phi_{b\mathbf{H}}(\boldsymbol{x}-\boldsymbol{z})\mathrm{vec}^{\mathrm{T}}[(b\mathbf{H})^{-1}(\boldsymbol{y}-\boldsymbol{z})(\boldsymbol{y}-\boldsymbol{z})^{\mathrm{T}}(b\mathbf{H})^{-1}-(b\mathbf{H})^{-1}]\mathbf{D}_d f(\boldsymbol{x})f(\boldsymbol{y})f(\mathbf{z})\,\mathrm{d}\boldsymbol{x}\,\mathrm{d}\boldsymbol{y}\,\mathrm{d}\boldsymbol{z} \\
&= \int_{\mathbb{R}^{3d}} \phi_{\mathbf{I}}(\boldsymbol{v})\mathbf{D}_d^{\mathrm{T}}\mathrm{vec}[(a\mathbf{H})^{-1/2}\boldsymbol{v}\boldsymbol{v}^{\mathrm{T}}(a\mathbf{H})^{-1/2}-(a\mathbf{H})^{-1}] \\
&\quad \times \phi_{\mathbf{I}}(\boldsymbol{w})\mathrm{vec}^{\mathrm{T}}[(b\mathbf{H})^{-1/2}\boldsymbol{w}\boldsymbol{w}^{\mathrm{T}}(b\mathbf{H})^{-1/2}-(b\mathbf{H})^{-1}] \\
&\quad \times f(\boldsymbol{y}+(a\mathbf{H})^{1/2}\boldsymbol{v})f(\mathbf{y})f(\boldsymbol{y}-(b\mathbf{H})^{1/2}\boldsymbol{w})\,\mathrm{d}\boldsymbol{v}\,\mathrm{d}\boldsymbol{w}\,\mathrm{d}\boldsymbol{y} \\
&= O(\mathbf{J}_{d'})(\mathrm{vech}\,\mathbf{H})(\mathrm{vech}^{\mathrm{T}}\mathbf{H}),
\end{aligned}
$$

which means that

$$
\mathrm{Cov}[(\varphi_{2\mathbf{H}}-\varphi_{\mathbf{H}})(\mathbf{X}_1-\mathbf{X}_2),(\varphi_{2\mathbf{H}}-\varphi_{\mathbf{H}})(\mathbf{X}_2-\mathbf{X}_3)] = O(\mathbf{J}_{d'})(\mathrm{vech}\,\mathbf{H})(\mathrm{vech}^{\mathrm{T}}\mathbf{H}).
$$

Combining the expression for this covariance with (16), as $\mathbf{H}_{\mathrm{AMISE}} = O(\mathbf{J}_{d'}n^{-2/(d+4)})$, yields

$$
\begin{aligned}
&\mathrm{Var}[D_{\mathbf{H}}(\mathrm{UCV}-\mathrm{AMISE})(\mathbf{H}_{\mathrm{AMISE}})] \\
&= O(\mathbf{J}_{d'}n^{-2}|\mathbf{H}_{\mathrm{AMISE}}|^{-1/2})(\mathrm{vech}\,\mathbf{H}_{\mathrm{AMISE}}^{-1})(\mathrm{vech}^{\mathrm{T}}\mathbf{H}_{\mathrm{AMISE}}^{-1}) \\
&\qquad + O(\mathbf{J}_{d'}n^{-1})(\mathrm{vech}\,\mathbf{H}_{\mathrm{AMISE}})(\mathrm{vech}^{\mathrm{T}}\mathbf{H}_{\mathrm{AMISE}}) \\
&= O(\mathbf{J}_{d'}n^{-d/(d+4)})(\mathrm{vech}\,\mathbf{H}_{\mathrm{AMISE}})(\mathrm{vech}^{\mathrm{T}}\mathbf{H}_{\mathrm{AMISE}}).
\end{aligned}
$$

As $D^2_{\mathbf{H}}\mathrm{AMISE}(\mathbf{H}_{\mathrm{AMISE}}) = O(\mathbf{J}_d)$ then

$$
\mathrm{AVar}(\mathrm{vec}\,\hat{\mathbf{H}}_{\mathrm{UCV}}) = O(\mathbf{J}_{d'}n^{-d/(d+4)})(\mathrm{vech}\,\mathbf{H}_{\mathrm{AMISE}})(\mathrm{vech}^{\mathrm{T}}\mathbf{H}_{\mathrm{AMISE}}).
$$

Thus the proof is complete.

Theorem 1 is proved by applying lemma 1 to the result of lemma 2.

Proof of theorems 3 and 4

We need to keep track of additional terms in the asymptotic expansions for theorem 3, so we define

$$
\mathrm{AMISE}'(\mathbf{H}) = \mathrm{AMISE}(\mathbf{H}) + \frac{1}{8}\int_{\mathbb{R}^d} \mathrm{tr}(\mathbf{H}D^2 f(\boldsymbol{x}))\mathrm{tr}(\mathbf{H}^2(D^2)^2 f(\boldsymbol{x}))\mathrm{d}\boldsymbol{x},
$$

which is correct for AMISE up to second-order. We write $\mathbf{H}_{\mathrm{AMISE}'}$ for the minimizer of AMISE′. It is straightforward to show that lemma 1 applies when AMISE and $\mathbf{H}_{\mathrm{AMISE}}$ are replaced by AMISE′ and $\mathbf{H}_{\mathrm{AMISE}'}$, respectively. The next two lemmas make use of this fact in order to compute asymptotic expansions for the bias and variance of vech $\hat{\mathbf{H}}_{\mathrm{SCV}}$.

Lemma 3

Under the conditions of theorem 3,

$$
\begin{aligned}
\mathrm{Bias}(\mathrm{vech}\,\hat{\mathbf{H}}_{\mathrm{SCV}}) &= g^2 n^{-2/(d+4)}\mathbf{C}_{\mu_1} + n^{-1}g^{-d-4}n^{-2/(d+4)}\mathbf{C}_{\mu_2} + O(\mathbf{J}_{d'}g^4)\mathrm{vech}\,\mathbf{H}_{\mathrm{AMISE}} \\
&\quad + o(g^2 + n^{-1}g^{-d-4}),
\end{aligned}
$$

where

$$
\mathbf{C}_{\mu_1} = \frac{1}{2}\mathbf{D}_d^{\mathrm{T}}\mathrm{vec}(\boldsymbol{\Theta}_6\mathbf{C}_{\mathrm{AMISE}})\mathbf{C}_{\mu_2} = \frac{1}{8}(4\pi)^{-d/2}[2\mathbf{D}_d^{\mathrm{T}}\mathrm{vec}\,\mathbf{C}_{\mathrm{AMISE}} + (\mathrm{tr}\,\mathbf{C}_{\mathrm{AMISE}})\mathbf{D}_d^{\mathrm{T}}\mathrm{vec}\,\mathbf{I}_d].
$$

Proof. Straightforward manipulations give

$$\begin{aligned}\mathrm{SCV}(\mathbf{H}) &= n^{-1}(4\pi)^{-d/2}|\mathbf{H}|^{-1/2} + n^{-1}(\phi_{2\mathbf{H}+2\mathbf{G}} - 2\phi_{\mathbf{H}+2\mathbf{G}} + \phi_{2\mathbf{G}})(\mathbf{0})\\ &\quad + n^{-2}\sum_{i=1}^{n}\sum_{\substack{j=1\\ j\neq i}}^{n}(\phi_{2\mathbf{H}+2\mathbf{G}} - 2\phi_{\mathbf{H}+2\mathbf{G}} + \phi_{2\mathbf{G}})(\boldsymbol{X}_i - \boldsymbol{X}_j).\end{aligned}$$

The expected value of this is

$$\mathbb{E}\mathrm{SCV}(\mathbf{H}) = n^{-1}[(4\pi)^{-d/2}|\mathbf{H}|^{-1/2} + C_1] + \mathbb{E}(\phi_{2\mathbf{H}+2\mathbf{G}} - 2\phi_{\mathbf{H}+2\mathbf{G}} + \phi_{2\mathbf{G}})(\boldsymbol{X}_1 - \boldsymbol{X}_2),$$

where $C_1 = (2\pi)^{-d/2}|2\mathbf{H} + 2\mathbf{G}|^{-1/2}-2(2\pi)^{-d/2}|\mathbf{H} + 2\mathbf{G}|^{-1/2} + (2\pi)^{-d/2}|2\mathbf{G}|^{-1/2}$. For $\mathbf{A} = a\mathbf{H} + b\mathbf{G}$,

$$\mathbb{E}\phi_{\mathbf{A}}(\boldsymbol{X}_1 - \boldsymbol{X}_2) = \int_{\mathbb{R}^{2d}} \phi_{\mathbf{A}}(\boldsymbol{x} - \boldsymbol{y})f(\boldsymbol{x})f(\boldsymbol{y})\mathrm{d}\boldsymbol{x}\mathrm{d}\boldsymbol{y} = \int_{\mathbb{R}^{2d}} \phi_{\mathbf{I}}(\boldsymbol{w})f(\boldsymbol{y} + \mathbf{A}^{1/2}\boldsymbol{w})f(\boldsymbol{y})\mathrm{d}\boldsymbol{w}\mathrm{d}\boldsymbol{y}.$$

The eighth-order Taylor series expansion of $f(\boldsymbol{y} + \mathbf{A}^{1/2}\boldsymbol{w})$ is

$$\begin{aligned}f(\boldsymbol{y} + \mathbf{A}^{1/2}\boldsymbol{w}) =& f(\boldsymbol{y}) + \mathrm{tr}(\mathbf{A}^{1/2}D\boldsymbol{w}^{\mathrm{T}})f(\boldsymbol{y}) + \frac{1}{2!}\mathrm{tr}(\mathbf{A}D^2\boldsymbol{w}\boldsymbol{w}^{\mathrm{T}})f(\boldsymbol{y})\\ &+ \frac{1}{3!}[\mathrm{tr}(\mathbf{A}^{1/2}D\boldsymbol{w}^{\mathrm{T}})\mathrm{tr}(\mathbf{A}D^2\boldsymbol{w}\boldsymbol{w}^{\mathrm{T}})]f(\boldsymbol{y}) + \frac{1}{4!}\mathrm{tr}^2(\mathbf{A}D^2\boldsymbol{w}\boldsymbol{w}^{\mathrm{T}})f(\boldsymbol{y})\\ &+ \frac{1}{5!}[\mathrm{tr}(\mathbf{A}^{1/2}D\boldsymbol{w}^{\mathrm{T}})\mathrm{tr}^2(\mathbf{A}D^2\boldsymbol{w}\boldsymbol{w}^{\mathrm{T}})]f(\boldsymbol{y}) + \frac{1}{6!}\mathrm{tr}^2(\mathbf{A}D^2\boldsymbol{w}\boldsymbol{w}^{\mathrm{T}})f(\boldsymbol{y})\\ &+ \frac{1}{7!}[\mathrm{tr}(\mathbf{A}^{1/2}D\boldsymbol{w}^{\mathrm{T}})\mathrm{tr}^3(\mathbf{A}D^2\boldsymbol{w}\boldsymbol{w}^{\mathrm{T}})]f(\boldsymbol{y}) + \frac{1}{8!}\mathrm{tr}^4(\mathbf{A}D^2\boldsymbol{w}\boldsymbol{w}^{\mathrm{T}})f(\boldsymbol{y})\\ &+ o(\|\mathrm{vech}\mathbf{A}\|^4).\end{aligned}$$

For $i = 0, 1, 2, \ldots$, let

$$\begin{aligned}m_{2i} = m_{2i}(\phi_{\mathbf{I}}; \mathbf{A}) &= \int_{\mathbb{R}^d} \phi_{\mathbf{I}}(\boldsymbol{w})\mathrm{tr}^i(\mathbf{A}D^2\boldsymbol{w}\boldsymbol{w}^{\mathrm{T}})\,\mathrm{d}\boldsymbol{w}\\ m_{2i+1} = m_{2i+1}(\phi_{\mathbf{I}}; \mathbf{A}) &= \int_{\mathbb{R}^d} \phi_{\mathbf{I}}(\boldsymbol{w})\mathrm{tr}^i(\mathbf{A}D^2\boldsymbol{w}\boldsymbol{w}^{\mathrm{T}})\mathrm{tr}(\mathbf{A}^{1/2}D\boldsymbol{w}^{\mathrm{T}})\,\mathrm{d}\boldsymbol{w}\end{aligned}$$

then we have $m_0 = 1$, $m_2 = \mathrm{tr}(\mathbf{A}D^2)$, $m_4 = 3\mathrm{tr}(\mathbf{A}^2(D^2)^2)$, $m_6 = 15tr(\mathbf{A}^3(D^2)^3)$ and $m_8 = 105tr(\mathbf{A}^4(D^2)^4)$; and $m_1 = m_3 = m_5 = m_7 = 0$. Thus

$$\begin{aligned}&\mathbb{E}\phi_{\mathbf{A}}(\boldsymbol{X}_1 - \boldsymbol{X}_2)\\ &= \int_{\mathbb{R}^d}[m_0 f(\boldsymbol{y}) + \frac{1}{2}m_2 f(\boldsymbol{y}) + \frac{1}{4!}m_4 f(\boldsymbol{y}) + \frac{1}{6!}m_6 f(\boldsymbol{y}) + \frac{1}{8!}m_8 f(\boldsymbol{y})]f(\boldsymbol{y})\mathrm{d}\boldsymbol{y}\\ &\quad + o(\|\mathrm{vech}\mathbf{A}\|^4)\\ &= \int_{\mathbb{R}^d}[f(\boldsymbol{y}) + \frac{1}{2}\mathrm{tr}(\mathbf{A}D^2 f(\boldsymbol{y})) + \frac{1}{8}\mathrm{tr}(\mathbf{A}^2(D^2)^2 f(\boldsymbol{y})) + \frac{5}{240}\mathrm{tr}(\mathbf{A}^3(D^2)^3 f(\boldsymbol{y}))\\ &\quad + \frac{1}{384}\mathrm{tr}(\mathbf{A}^4(D^2)^4 f(\boldsymbol{y}))]f(\boldsymbol{y})\mathrm{d}\boldsymbol{y} + o(\|\mathrm{vech}\mathbf{A}\|^4).\end{aligned}$$

Now

$$\begin{aligned}&\mathbb{E}(\phi_{2\mathbf{H}+2\mathbf{G}} - 2\phi_{\mathbf{H}+2\mathbf{G}} + \phi_{2\mathbf{G}})(\boldsymbol{X}_1 - \boldsymbol{X}_2)\\ &= \frac{1}{4}\int_{\mathbb{R}^d}\mathrm{tr}(\mathbf{H}^2(D^2)^2 f(\boldsymbol{y}))f(\boldsymbol{y})\mathrm{d}\boldsymbol{y} + \frac{1}{4}\int_{\mathbb{R}^d}\mathrm{tr}(\mathbf{H}^2\mathbf{G}(D^2)^3 f(\boldsymbol{y}))f(\boldsymbol{y})\mathrm{d}\boldsymbol{y}\\ &\quad + \frac{1}{8}\int_{\mathbb{R}^d}\mathrm{tr}(\mathbf{H}^3(D^2)^3 f(\boldsymbol{y}))f(\boldsymbol{y})\mathrm{d}\boldsymbol{y} + O(\|\mathrm{vech}\,\mathbf{H}^2\mathbf{G}^2\|).\end{aligned}$$

so

$$\mathbb{E}\mathrm{SCV}(\mathbf{H}) = n^{-1}C_1 + \mathrm{AMISE}'(\mathbf{H}) + \frac{1}{4}\int_{\mathbb{R}^d} \mathrm{tr}(\mathbf{H}^2\mathbf{G}(D^2)^3 f(\boldsymbol{y}))f(\boldsymbol{y})\mathrm{d}\boldsymbol{y} + O(\|\mathrm{vech}\,\mathbf{H}^2\mathbf{G}^2\|)$$

and

$$\mathbb{E}[(\mathrm{SCV} - \mathrm{AMISE}')(\mathbf{H})] = n^{-1}C_1 + \frac{1}{4}\mathrm{tr}(\mathbf{H}^2\mathbf{G}\boldsymbol{\Theta}_6) + O(\|\mathrm{vech}\,\mathbf{H}^2\mathbf{G}^2\|),$$

where $\Theta_6 = \int_{\mathbb{R}^d}(D^2)^3 f(\boldsymbol{y})f(\boldsymbol{y})\mathrm{d}\boldsymbol{y}$.

We now have $\mathbb{E}(\mathrm{SCV}{-}\mathrm{AMISE}')(\mathbf{H})$. The next step is to find the derivative of this. The derivative of C_1 is

$$\begin{aligned}
D_{\mathbf{H}}C_1 &= -(2\pi)^{-d/2}|2\mathbf{H} + 2\mathbf{G}|^{-1/2}\mathbf{D}_d^{\mathrm{T}}\mathrm{vec}(2\mathbf{H} + 2\mathbf{G})^{-1} \\
&\quad + (2\pi)^{-d/2}|\mathbf{H} + 2\mathbf{G}|^{-1/2}\mathbf{D}_d^{\mathrm{T}}\mathrm{vec}(\mathbf{H} + 2\mathbf{G})^{-1} \\
&= -(4\pi)^{-d/2}\left[\frac{1}{2}g^{-d-2}\mathbf{D}_d^{\mathrm{T}}\mathrm{vec}\,\mathbf{I}_d - \frac{1}{4}g^{-d-4}(2\mathbf{D}_d^{\mathrm{T}}\mathrm{vec}\,\mathbf{H} + (\mathrm{tr}\mathbf{H})\mathbf{D}_d^{\mathrm{T}}\mathrm{vec}\,\mathbf{I}_d)\right] \\
&\quad + (4\pi)^{-d/2}\left[\frac{1}{2}g^{-d-2}\mathbf{D}_d^{\mathrm{T}}\mathrm{vec}\,\mathbf{I}_d - \frac{1}{8}g^{-d-4}(2\mathbf{D}_d^{\mathrm{T}}\mathrm{vec}\,\mathbf{H} + (\mathrm{tr}\mathbf{H})\mathbf{D}_d^{\mathrm{T}}\mathrm{vec}\,\mathbf{I}_d)\right] \\
&\quad + O(g^{-d-6}\|\mathrm{vech}\,\mathbf{H}\|^2) \\
&= \frac{1}{8}(4\pi)^{-d/2}g^{-d-4}[2\mathbf{D}_d^{\mathrm{T}}\mathrm{vec}\,\mathbf{H} + (\mathrm{tr}\mathbf{H})\mathbf{D}_d^{\mathrm{T}}\mathrm{vec}\,\mathbf{I}_d] + O(g^{-d-6}\|\mathrm{vech}\,\mathbf{H}\|^2)
\end{aligned}$$

after some lengthy manipulations.

The derivative of $\frac{1}{4}g^2\mathrm{tr}(\mathbf{H}^2\boldsymbol{\Theta}_6) + O(g^4\|\mathrm{vech}\,\mathbf{H}\|^2)$ is $\frac{1}{2}g^2\mathbf{D}_d^{\mathrm{T}}\mathrm{vec}(\boldsymbol{\Theta}_6\mathbf{H}) + O(g^4\mathrm{vech}\,\mathbf{H})$. Combining these two derivatives and then interchanging the expectation and derivative operators, we have

$$\begin{aligned}
&\mathbb{E}[D_{\mathbf{H}}(\mathrm{SCV} - \mathrm{AMISE}')(\mathbf{H}_{\mathrm{AMISE}})] \\
&= \frac{1}{2}g^2\mathbf{D}_d^{\mathrm{T}}\mathrm{vec}(\boldsymbol{\Theta}_6\mathbf{H}_{\mathrm{AMISE}}) + \frac{1}{8}(4\pi)^{-d/2}n^{-1}g^{-d-4}[2\mathbf{D}_d^{\mathrm{T}}\ \mathrm{vec}\ \mathbf{H}_{\mathrm{AMISE}} + (\mathrm{tr}\ \mathbf{H}_{\mathrm{AMISE}})\mathbf{D}_d^{\mathrm{T}}\mathrm{vec}\,\mathbf{I}_d] \\
&+ o(\|\mathrm{vech}\,\mathbf{H}_{\mathrm{AMISE}}\|(g^2 + n^{-1}g^{-d-4})).
\end{aligned}$$

As $D^2_{\mathbf{H}}\mathrm{AMISE}(\mathbf{H}_{\mathrm{AMISE}}) = O(\mathbf{J}_{d'})$ the result follows.

Lemma 4

Under the conditions of theorem 3

$$\mathrm{Var}(\mathrm{vech}\,\hat{\mathbf{H}}_{\mathrm{SCV}}; g) = O(\mathbf{J}_{d'}(n^{-2}g^{-d-8} + n^{-1}))(\mathrm{vech}\,\mathbf{H}_{\mathrm{AMISE}})(\mathrm{vech}^{\mathrm{T}}\,\mathbf{H}_{\mathrm{AMISE}}).$$

Proof.

$$\begin{aligned}
&\mathrm{Var}[D_{\mathbf{H}}(\mathrm{SCV} - \mathrm{AMISE}')(\mathbf{H})] \\
&= \mathrm{Var}[D_{\mathbf{H}}\mathrm{SCV}(\mathbf{H})] \\
&= n^{-4}\mathrm{Var}\left[\sum_{i=1}^{n}\sum_{\substack{j=1\\ j\neq i}}^{n} D_{\mathbf{H}}(\phi_{2\mathbf{H}+2\mathbf{G}} - 2\phi_{\mathbf{H}+2\mathbf{G}} + \phi_{2\mathbf{G}})(\boldsymbol{X}_i - \boldsymbol{X}_j)\right] \\
&= 2n^{-2}\mathrm{Var}[(\boldsymbol{\varphi}_{2\mathbf{H}+2\mathbf{G}} - \boldsymbol{\varphi}_{\mathbf{H}+2\mathbf{G}})(\boldsymbol{X}_1 - \boldsymbol{X}_2)] \\
&\quad + 4n^{-1}\mathrm{Cov}[(\boldsymbol{\varphi}_{2\mathbf{H}+2\mathbf{G}} - \boldsymbol{\varphi}_{\mathbf{H}+2\mathbf{G}})(\boldsymbol{X}_1 - \boldsymbol{X}_2), (\boldsymbol{\varphi}_{2\mathbf{H}+2\mathbf{G}} - \boldsymbol{\varphi}_{\mathbf{H}+2\mathbf{G}})(\boldsymbol{X}_2 - \boldsymbol{X}_3)], \qquad (17)
\end{aligned}$$

where

$$\boldsymbol{\varphi}_{a\mathbf{H}+b\mathbf{G}}(\boldsymbol{X}) = \phi_{a\mathbf{H}+b\mathbf{G}}(\boldsymbol{X})\mathbf{D}_d^{\mathrm{T}}\mathrm{vec}[(a\mathbf{H}+b\mathbf{G})^{-1}\boldsymbol{X}\boldsymbol{X}^{\mathrm{T}}(a\mathbf{H}+b\mathbf{G})^{-1} - (a\mathbf{H}+b\mathbf{G})^{-1}]$$

and $D_{\mathbf{H}}\phi_{a\mathbf{H}+b\mathbf{G}}(\boldsymbol{x}) = \frac{1}{2}a\boldsymbol{\varphi}_{a\mathbf{H}+b\mathbf{G}}(\boldsymbol{x})$. The first term of Var[$D_{\mathbf{H}}$(SCV–AMISE′)(**H**)] is

$$\begin{aligned}
&\mathrm{Var}[(\boldsymbol{\varphi}_{2\mathbf{H}+2\mathbf{G}} - \boldsymbol{\varphi}_{\mathbf{H}+2\mathbf{G}})(\boldsymbol{X}_1 - \boldsymbol{X}_2)]\\
&= \mathbb{E}\Big\{\big[(\boldsymbol{\varphi}_{2\mathbf{H}+2\mathbf{G}} - \boldsymbol{\varphi}_{\mathbf{H}+2\mathbf{G}})(\boldsymbol{X}_1 - \boldsymbol{X}_2)\big]\big[(\boldsymbol{\varphi}_{2\mathbf{H}+2\mathbf{G}} - \boldsymbol{\varphi}_{\mathbf{H}+2\mathbf{G}})(\boldsymbol{X}_1 - \boldsymbol{X}_2)\big]^{\mathrm{T}}\Big\}\\
&\quad - \big[\mathbb{E}(\boldsymbol{\varphi}_{2\mathbf{H}+2\mathbf{G}} - \boldsymbol{\varphi}_{\mathbf{H}+2\mathbf{G}})(\boldsymbol{X}_1 - \boldsymbol{X}_2)\big]\big[\mathbb{E}(\boldsymbol{\varphi}_{2\mathbf{H}+2\mathbf{G}} - \boldsymbol{\varphi}_{\mathbf{H}+2\mathbf{G}})(\boldsymbol{X}_1 - \boldsymbol{X}_2)\big]^{\mathrm{T}}\\
&= O(\mathbf{J}_{d'}g^{-d-8})(\mathrm{vech}\,\mathbf{H})(\mathrm{vech}^{\mathrm{T}}\mathbf{H}), \qquad (18)
\end{aligned}$$

after manipulations similar to those in the proof of lemma 3.

We now turn our attention to the second term of Var[$D_{\mathbf{H}}$(SCV–AMISE′)(**H**)]:

$$\begin{aligned}
&\mathrm{Cov}[(\boldsymbol{\varphi}_{2\mathbf{H}+2\mathbf{G}} - \boldsymbol{\varphi}_{\mathbf{H}+2\mathbf{G}})(\boldsymbol{X}_1 - \boldsymbol{X}_2), (\boldsymbol{\varphi}_{2\mathbf{H}+2\mathbf{G}} - \boldsymbol{\varphi}_{\mathbf{H}+2\mathbf{G}})(\boldsymbol{X}_2 - \boldsymbol{X}_3)]\\
&= \mathbb{E}\Big\{\big[(\boldsymbol{\varphi}_{2\mathbf{H}+2\mathbf{G}} - \boldsymbol{\varphi}_{\mathbf{H}+2\mathbf{G}})(\boldsymbol{X}_1 - \boldsymbol{X}_2)\big]\big[(\boldsymbol{\varphi}_{2\mathbf{H}+2\mathbf{G}} - \boldsymbol{\varphi}_{\mathbf{H}+2\mathbf{G}})(\boldsymbol{X}_2 - \boldsymbol{X}_3)\big]^{\mathrm{T}}\Big\}\\
&\quad - \big[\mathbb{E}(\boldsymbol{\varphi}_{2\mathbf{H}+2\mathbf{G}} - \boldsymbol{\varphi}_{\mathbf{H}+2\mathbf{G}})(\boldsymbol{X}_1 - \boldsymbol{X}_2)\big]\big[\mathbb{E}(\boldsymbol{\varphi}_{2\mathbf{H}+2\mathbf{G}} - \boldsymbol{\varphi}_{\mathbf{H}+2\mathbf{G}})(\boldsymbol{X}_2 - \boldsymbol{X}_3)\big]^{\mathrm{T}}.
\end{aligned}$$

We already have values for the second part of this expression. For the first part, we obtain

$$\begin{aligned}
&\mathrm{Cov}[(\boldsymbol{\varphi}_{2\mathbf{H}+2\mathbf{G}} - \boldsymbol{\varphi}_{\mathbf{H}+2\mathbf{G}})(\boldsymbol{X}_1 - \boldsymbol{X}_2), (\boldsymbol{\varphi}_{2\mathbf{H}+2\mathbf{G}} - \boldsymbol{\varphi}_{\mathbf{H}+2\mathbf{G}})(\boldsymbol{X}_2 - \boldsymbol{X}_3)]\\
&= O(\mathbf{J}_{d'})(\mathrm{vech}\,\mathbf{H})(\mathrm{vech}^{\mathrm{T}}\,\mathbf{H}). \qquad (19)
\end{aligned}$$

If we substitute (18) and (19) into (17):

$$\begin{aligned}
&\mathrm{Var}[D_{\mathbf{H}}(\mathrm{SCV} - \mathrm{AMISE}')(\hat{\mathbf{H}}_{\mathrm{SCV}}; g)]\\
&= O(\mathbf{J}_{d'}(n^{-2}g^{-d-8} + n^{-1}))(\mathrm{vech}\,\mathbf{H}_{\mathrm{AMISE}})(\mathrm{vech}^{\mathrm{T}}\,\mathbf{H}_{\mathrm{AMISE}})
\end{aligned}$$

and the result is proved.

Now, g_1 is the minimizer of the asymptotic version of $Q(g)$, and

$$\begin{aligned}
Q(g) &= \mathrm{tr}\,\mathrm{MSE}(\mathrm{vech}\,\hat{\mathbf{H}})\\
&= [\mathrm{Bias}(\mathrm{vech}\,\hat{\mathbf{H}}_{\mathrm{SCV}}; g)]^{\mathrm{T}}[\mathrm{Bias}(\mathrm{vech}\,\hat{\mathbf{H}}_{\mathrm{SCV}}; g)] + O\Big(n^{-(2d+12)/(d+4)}g^{-d-8} + n^{-(d+8)/(d+4)}\Big)\\
&= \Big(g^2 n^{-2/(d+4)}\mathbf{C}_{\mu_1} + n^{-(d+6)/(d+4)}g^{-d-4}\mathbf{C}_{\mu_2}\Big)^{\mathrm{T}}(g^2 n^{-2/(d+4)}\mathbf{C}_{\mu_1}\\
&\quad + n^{-(d+6)/(d+4)}g^{-d-4}\mathbf{C}_{\mu_2}) + O\Big(n^{-(2d+12)/(d+4)}g^{-d-8} + n^{-(d+8)/(d+4)}\Big)\\
&= Q^*(g), \qquad (20)
\end{aligned}$$

where

$$\begin{aligned}
Q^*(g) &= g^4 n^{-4/(d+4)}\mathbf{C}_{\mu_1}^{\mathrm{T}}\mathbf{C}_{\mu_1} + 2n^{-(d+8)/(d+4)}g^{-d-2}\mathbf{C}_{\mu_2}^{\mathrm{T}}\mathbf{C}_{\mu_1} + n^{-(2d+12)/(d+4)}g^{-2d-8}\mathbf{C}_{\mu_2}^{\mathrm{T}}\mathbf{C}_{\mu_2}\\
&\quad + O(n^{-(2d+12)/(d+4)}g^{-d-8} + n^{-(d+8)/(d+4)}).
\end{aligned}$$

Here we have used the previous pair of lemmas, incorporating the variance into the ultimately negligible last term. Ignoring this term, the function $Q^*(g)n^{4/(d+4)}/g^4$ is a quadratic in $n^{-1}g^{-d-6}$, which can be minimized by elementary calculus to give the result in theorem 3.

Theorem 4 follows by application of lemmas 3 and 4 with $g = g_1$.