Available online at www.sciencedirect.com
SCIENCE DIRECT®
Journal of Multivariate Analysis
ELSEVIER
Journal of Multivariate Analysis 93 (2005) 417–433
http://www.elsevier.com/locate/jmva

Convergence rates for unconstrained bandwidth matrix selectors in multivariate kernel density estimation

Tarn Duong and Martin L. Hazelton*

School of Mathematics and Statistics M1019, University of Western Australia, 35 Stirling Hwy, Crawley, WA 6009, Australia

Received 7 March 2003

Available online 25 May 2004

Abstract

Progress in selection of smoothing parameters for kernel density estimation has been much slower in the multivariate than univariate setting. Within the context of multivariate density estimation attention has focused on diagonal bandwidth matrices. However, there is evidence to suggest that the use of full (or unconstrained) bandwidth matrices can be beneficial. This paper presents some results in the asymptotic analysis of data-driven selectors of full bandwidth matrices. In particular, we give relative rates of convergence for plug-in selectors and a biased cross-validation selector.

AMS 2000 subject classifications: 62G07; 62G20

Keywords: Asymptotic; Biased cross-validation; Gaussian kernel; MISE; Plug-in; Smoothing

1. Introduction

The choice of smoothing parameters is a problem of fundamental importance in kernel density estimation and related areas. Bandwidth selection for univariate kernel density estimation is the simplest form of this problem, and has been the subject of considerable research. Substantial advances been made leading to the

*Corresponding author. Fax: +61-8-6488-1028.
E-mail address: martin@maths.uwa.edu.au (M.L. Hazelton).

0047-259X/$ - see front matter
doi:10.1016/j.jmva.2004.04.004

development of bandwidth selectors which combine good practical performance with excellent asymptotic properties. See [4] for an overview. Progress in the case of multivariate case has been much slower. Nonetheless, the selection of bandwidth matrices is an important problem because of the utility of multivariate kernel density estimators in areas such as data visualization, nonparametric discriminant analysis and goodness of fit testing.

Successful approaches to univariate bandwidth selection, such as plug-in and cross-validation methods, can in principle be transferred to the multivariate setting. However, analysis of these techniques in more than one dimension is not entirely straightforward, at least in part because there is no univariate analogue to the multivariate issue of kernel orientation to the coordinate axes. It follows that multivariate bandwidth selection can be significantly simplified by constraining the bandwidth matrix to be diagonal. Several authors have studied data-driven choice of diagonal bandwidth matrices, and plug-in [12] and cross-validation [7] selectors have been developed. However, the lack of flexibility in this type of bandwidth matrix can have an adverse effect on the performance of the resulting density estimator, even if the data are pre-sphered. Consider, for example, the bivariate normal mixture density displayed in Fig. 1. We conducted a simulation study in which 400 data sets of size 100, and 400 of size 1000, were generated from this target density. Kernel density estimates were obtained using a 2-stage diagonal plug-in bandwidth matrix, and a 2-stage full plug-in bandwidth matrix. (See Section 2 for a description of this

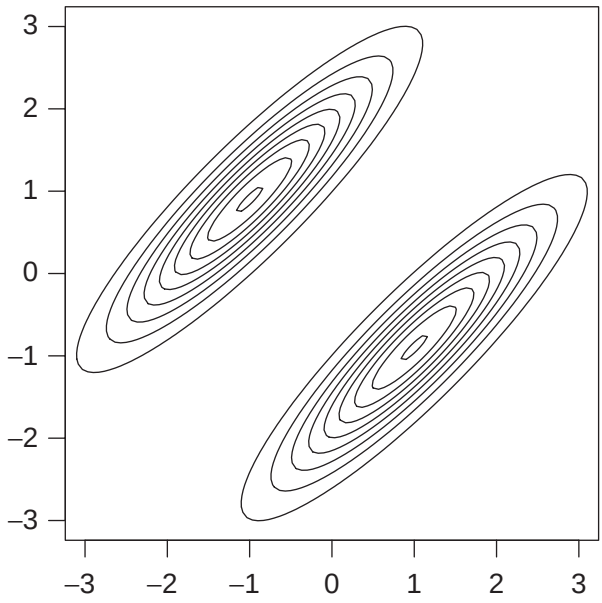

Fig. 1. Contour plot for the normal mixture

$$\frac{1}{2}N\left(\begin{bmatrix}1\\-0.9\end{bmatrix},\begin{bmatrix}1 & 0.9\\0.9 & 1\end{bmatrix}\right)+\frac{1}{2}N\left(\begin{bmatrix}-1\\0.9\end{bmatrix},\begin{bmatrix}1 & 0.9\\0.9 & 1\end{bmatrix}\right).$$

type of bandwidth selector.) The mean integrated squared error for the estimates using the diagonal bandwidth matrix was reduced by over 25% for the smaller sample size, and by more than 38% for the larger sample size, by using the full bandwidth matrix approach. Sphering the data does not help the diagonal bandwidth matrix estimator in this case because the overall covariance matrix of the target density is diagonal. Further examples of the advantages of using full bandwidth matrices with certain types of target density are supplied in the simulation study and real data analysis in [2]. The remarks of Wand and Jones [11] on the subject also deserve attention.

The purpose of this paper is to derive some results that are helpful in the asymptotic analysis of full (i.e. unconstrained) bandwidth matrix selectors in multivariate kernel density estimation. Our first result, Lemma 1, builds upon some heuristic arguments given by Wand and Jones [12] to give the relative rate of convergence of a bandwidth matrix selector in terms of asymptotic properties of estimates of mean integrated squared error. We demonstrate the application of this result in two specific cases. In the first instance we derive the convergence rates for the plug-in selectors recently investigated by Duong and Hazelton [1]. While the performance of these selectors has been assessed through a simulation study, the asymptotic behaviour has not been analysed in the literature. In the second case we consider biased cross-validation (BCV) selectors. Our results generalize those of Sain et al. [7], who focused on constrained bandwidth matrices, and correct the previously published convergence rate for this type of selector in high dimensions.

The remainder of the paper is organized as follows. In Section 2 we cover the necessary background material on bandwidth matrix selection and then give Lemma 1 and its proof. In Section 3 we turn our attention to plug-in selectors, convergence rates for which are given in Theorem 1. Convergence rates for BCV selectors are considered in Section 4. The main results are given in Theorem 2, the proof of which proceeds via Lemmas 2 and 3. The paper concludes with a discussion of some of the practical implications of our findings.

2. Bandwidth matrix selection

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

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

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

In common with most authors in the field, we measure the performance of $\hat{f}$ by mean integrated squared error (MISE),

$$\mathrm{MISE}(\mathbf{H}) \equiv \mathrm{MISE}\hat{f}(\cdot;\mathbf{H}) = \mathbb{E}\int (\hat{f}(\boldsymbol{x};\mathbf{H}) - f(\boldsymbol{x}))^2\, d\boldsymbol{x},$$

where it is understood here and hereafter that the integral is over $\mathbb{R}^d$ unless stated otherwise. MISE does not have a tractable closed form and so we resort to using an asymptotic approximation. The asymptotic mean integrated squared error (AMISE) is given by

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

where $R(K) = \int K(\boldsymbol{x})^2 d\boldsymbol{x} < \infty$, $\mu_2(K)\mathbf{I}_d = \int \boldsymbol{x}\boldsymbol{x}^T K(\boldsymbol{x})\, d\boldsymbol{x}$ with $\mu_2(K) < \infty$, $\mathbf{I}_d$ the $d \times d$ identity matrix, and vech is the vector half operator so that $\mathrm{vech}\,\mathbf{H}$ is the lower triangular half of $\mathbf{H}$ strung out columnwise into a vector. See [13, Chapter 4], for example. The $\boldsymbol{\Psi}_4$ matrix is the $\frac{1}{2}d(d+1) \times \frac{1}{2}d(d+1)$ matrix given by

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

where $D^2f(\boldsymbol{x})$ is the Hessian matrix of f and $\mathrm{dg}\,\mathbf{A}$ is matrix $\mathbf{A}$ with all of its non-diagonal elements set to zero. Sufficient conditions for the validity of the expansions defined by Eq. (1) 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 $n \to \infty$. With the introduction of some more notation we can derive an expression for individual elements of the matrix $\boldsymbol{\Psi}_4$. Let $\boldsymbol{r} = (r_1, r_2, , \ldots, r_d)$ where the $r_1, r_2, \ldots, r_d$ are non-negative integers. Let $|\boldsymbol{r}| = r_1 + r_2 + \cdots + r_d$ then the $\boldsymbol{r}$th partial derivative of f can be written as

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

Denote the integrated density derivative functional by

$$\psi_{\boldsymbol{r}} = \int f^{(\boldsymbol{r})}(\boldsymbol{x}) f(\boldsymbol{x})\, d\boldsymbol{x}$$

then the elements of $\boldsymbol{\Psi}_4$ are $\psi_{\boldsymbol{r}}$ functionals with $|\boldsymbol{r}| = 4$. In particular for the bivariate case

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

The bandwidth selectors described in this paper seek to estimate

$$\mathbf{H}_{\mathrm{AMISE}} = \operatorname*{argmin}_{\mathbf{H}} \mathrm{AMISE}\hat{f}(\cdot;\mathbf{H}),$$

which is a tractable surrogate for $\mathbf{H}_{\mathrm{MISE}}$, the minimizer of MISE. Both plug-in and BCV approaches work by obtaining estimates of the $\psi_{\boldsymbol{r}}$ functionals and hence of the

matrix $\mathbf{\Psi}_4$. Replacing $\mathbf{\Psi}_4$ by its estimate in (1) produces an estimate $\widehat{\text{AMISE}}$ of AMISE that can be minimized to give a data-driven bandwidth matrix. What distinguishes the plug-in from the BCV method is the manner in which $\psi_{\boldsymbol{r}}$ is estimated. In the plug-in case we denote the relevant estimators by $\check{\psi}_{\boldsymbol{r}}$ and $\check{\mathbf{\Psi}}_4$, and the estimated AMISE by

$$\text{PI}(\mathbf{H}) = n^{-1}R(K)|\mathbf{H}|^{-1/2} + \tfrac{1}{4}\mu_2(K)^2(\text{vech}^T\mathbf{H})\check{\mathbf{\Psi}}_4(\text{vech}\,\mathbf{H}). \tag{2}$$

The minimizer of (2) is the plug-in selector $\check{\mathbf{H}}$. For BCV the corresponding estimators are denoted $\tilde{\psi}_{\boldsymbol{r}}$ and $\tilde{\mathbf{\Psi}}_4$, and the estimated AMISE by

$$\text{BCV}(\mathbf{H}) = n^{-1}R(K)|\mathbf{H}|^{-1/2} + \tfrac{1}{4}\mu_2(K)^2(\text{vech}^T\mathbf{H})\tilde{\mathbf{\Psi}}_4(\text{vech}\,\mathbf{H}). \tag{3}$$

The minimizer of (3) is the BCV selector $\tilde{\mathbf{H}}$.

Following Wand [10], the plug-in estimator of $\psi_{\boldsymbol{r}}$ is

$$\check{\psi}_{\boldsymbol{r}}(\mathbf{G}) = n^{-1}\sum_{i=1}^{n}\hat{f}^{(\boldsymbol{r})}(\boldsymbol{X}_i;\mathbf{G}) = n^{-2}\sum_{i=1}^{n}\sum_{j=1}^{n}K_{\mathbf{G}}^{(\boldsymbol{r})}(\boldsymbol{X}_i - \boldsymbol{X}_j). \tag{4}$$

Here $\mathbf{G}$ is a pilot bandwidth matrix, crucial to the performance of the plug-in methodology. In line with Wand and Jones [12] and Duong and Hazelton [1] we constrain the pilot bandwidth matrix to be of the form $\mathbf{G} = g^2\mathbf{I}$. While this form of $\mathbf{G}$ may appear restrictive, the empirical work of Duong and Hazelton indicates that it can produce reasonable results when applied to pre-sphered data. Two data-driven methods for choosing g have been proposed. Wand and Jones [12] suggested employing a separate value of g for each functional $\check{\psi}_{\boldsymbol{r}}$ such that $|\boldsymbol{r}| = 4$. Specifically, for given $\boldsymbol{r}$ Wand and Jones suggested using the pilot bandwidth

$$g_{\boldsymbol{r},\text{AMSE}} = \operatorname*{argmin}_{g}\ \text{AMSE}\,\check{\psi}_{\boldsymbol{r}}(g),$$

where AMSE denotes asymptotic mean squared error. However, Duong and Hazelton noted that this approach could result in a matrix $\check{\mathbf{\Psi}}_4$ which is not positive-definite. These authors developed an alternative technique whereby a common g is applied in estimating all the $\psi_{\boldsymbol{r}}$ functionals which ensures that $\check{\mathbf{\Psi}}_4$ is positive-definite if K is multivariate normal. Duong and Hazelton proposed that the common g should be chosen to estimate $g_{4,\text{SAMSE}}$, the minimizer of the 'sum of AMSE' criterion

$$\text{SAMSE}_4(g) = \sum_{\boldsymbol{r}:|\boldsymbol{r}|=4}\text{AMSE}\,\check{\psi}_{\boldsymbol{r}}(g).$$

We shall refer to these different implementations of the plug-in bandwidth selector as the AMSE and SAMSE methods. In practice we do not know $g_{\boldsymbol{r},\text{AMSE}}$ or $g_{4,\text{SAMSE}}$ because they depend on functionals of f. Nonetheless, the convergence rates for the selectors do not suffer if the pilot bandwidths are replaced by any estimate of the correct order. Henceforth, we shall assume that the plug-in methods are executed using such pilot bandwidth estimates.

Biased cross-validation was introduced by Scott and Terrell [9] for univariate density estimation. In the multivariate setting Sain et al. [7] considered two slightly

different versions of BCV selector. We concentrate on the second (which these authors referred to as BCV2); cf. [3]. In this method the ψ_r functionals are estimated by

$$\tilde{\psi}_r(\mathbf{H}) = n^{-1} \sum_{i=1}^{n} \hat{f}_{-i}^{(r)}(\boldsymbol{X}_i; \mathbf{H}) = n^{-1}(n-1)^{-1} \sum_{i=1}^{n} \sum_{\substack{j=1 \\ j \neq 1}}^{n} K_{\mathbf{H}}^{(r)}(\boldsymbol{X}_i - \boldsymbol{X}_j). \tag{5}$$

In comparison with the plug-in method the pilot bandwidth has been set equal to $\mathbf{H}$, and the diagonal, non-stochastic, terms in (4) have been omitted. (It is also possible to implement the plug-in method with there terms removed although we do not pursue the matter here; cf. [5].)

The performance of a general bandwidth matrix selector can be assessed by its relative rate of convergence. We say that the selector $\hat{\mathbf{H}}$ converges to $\mathbf{H}_{\mathrm{AMISE}}$ with relative rate $n^{-\alpha}$ if

$$\mathrm{vech}(\hat{\mathbf{H}} - \mathbf{H}_{\mathrm{AMISE}}) = O_p(\mathbf{J}_{d^*} n^{-\alpha}) \mathrm{vech}\, \mathbf{H}_{\mathrm{AMISE}}, \tag{6}$$

where $\mathbf{J}_{d^*}$ is the $d^* \times d^*$ matrix of ones and $d^* = \frac{1}{2} d(d+1)$. Here we have extended the asymptotic order notation to matrix sequences. Specifically, let $\{\mathbf{A}_n\}$ and $\{\mathbf{B}_n\}$ be sequences of matrices with common dimensions. We write $\mathbf{A}_n = o(\mathbf{B}_n)$ if $a_{ij} = o(b_{ij})$ for all elements a_{ij} of $\mathbf{A}_n$ and b_{ij} of $\mathbf{B}_n$. We also have corresponding definitions for O, o_p and O_p. The rationale for using $O_p(\mathbf{J}_{d^*} n^{-\alpha})$, rather than $O_p(\mathbf{I}_{d^*} n^{-\alpha})$, in (6) is as follows. In the general case, when all elements of $\mathrm{vech}(\mathbf{H}_{\mathrm{AMISE}})$ are non-zero and $O(n^{-2/(d+4)})$, (6) remains valid if $O_p(\mathbf{J}_{d^*} n^{-\alpha})$ is replaced by $O_p(\mathbf{I}_{d^*} n^{-\alpha})$. However, some elements of $\mathrm{vech}(\mathbf{H}_{\mathrm{AMISE}})$ will be zero for certain types of target density. For example, the off-diagonal terms of $\mathbf{H}_{\mathrm{AMISE}}$ will be zero if f is a bivariate normal density with diagonal covariance matrix. In such circumstances it is natural to calculate the rate of convergence of the corresponding elements of $\hat{\mathbf{H}}$ relative to the general order of $\mathbf{H}_{\mathrm{AMISE}}$ (i.e. $O(n^{-2/(d+4)})$), since the relative rate of convergence to a zero element is undefined.

A problem in finding relative rates for plug-in and BCV selectors is that neither $\check{\mathbf{H}}$ nor $\tilde{\mathbf{H}}$ are available in closed form. Instead each must be found by numerical minimization of the appropriate estimate of AMISE. It is therefore useful to express convergence rates for a selector in terms of the asymptotic performance of the AMISE estimate. We can do so by means of Lemma 1.

Lemma 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* $n \to \infty$.
(A3) K *is a spherically symmetric probability density.*

Let $\hat{\mathbf{H}} = \mathrm{argmin}_{\mathbf{H}}\, \widehat{\mathrm{AMISE}}$ *be a bandwidth selector and define its mean squared error (MSE) by*

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

Then

$$\mathrm{MSE}(\mathrm{vech}\,\hat{\mathbf{H}}) = \mathrm{AMSE}(\mathrm{vech}\,\hat{\mathbf{H}})(\mathbf{I}_{d^*} + o(\mathbf{J}_{d^*})),$$

where the asymptotic MSE can be written as

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

in which

$$\mathrm{ABias}(\mathrm{vech}\,\hat{\mathbf{H}}) = [D_{\mathbf{H}}^2\mathrm{AMISE}(\mathbf{H}_{\mathrm{AMISE}})]^{-1}\mathbb{E}[D_{\mathbf{H}}(\widehat{\mathrm{AMISE}} - \mathrm{AMISE})(\mathbf{H}_{\mathrm{AMISE}})],$$

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

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

Proof. We may expand $D_{\mathbf{H}}\widehat{\mathrm{AMISE}}$ as follows:

$$\begin{aligned}D_{\mathbf{H}}\widehat{\mathrm{AMISE}}(\hat{\mathbf{H}}) &= D_{\mathbf{H}}(\widehat{\mathrm{AMISE}} - \mathrm{AMISE})(\hat{\mathbf{H}}) + D_{\mathbf{H}}\mathrm{AMISE}(\hat{\mathbf{H}}) \\ &= D_{\mathbf{H}}(\widehat{\mathrm{AMISE}} - \mathrm{AMISE})(\hat{\mathbf{H}}) + \{D_{\mathbf{H}}\mathrm{AMISE}(\mathbf{H}_{\mathrm{AMISE}}) \\ &\quad + [\mathbf{I}_{d^*} + o_p(\mathbf{J}_{d^*})]D_{\mathbf{H}}^2\mathrm{AMISE}(\mathbf{H}_{\mathrm{AMISE}})\mathrm{vech}(\hat{\mathbf{H}} - \mathbf{H}_{\mathrm{AMISE}})\}.\end{aligned}$$

Now we have $D_{\mathbf{H}}\widehat{\mathrm{AMISE}}(\hat{\mathbf{H}}) = \mathbf{0}$ and $D_{\mathbf{H}}\mathrm{AMISE}(\mathbf{H}_{\mathrm{AMISE}}) = \mathbf{0}$ so that

$$\begin{aligned}\mathrm{vech}(\hat{\mathbf{H}} - \mathbf{H}_{\mathrm{AMISE}}) = {} & -[\mathbf{I}_{d^*} + o_p(\mathbf{J}_{d^*})][D_{\mathbf{H}}^2\mathrm{AMISE}(\mathbf{H}_{\mathrm{AMISE}})]^{-1} \\ & \times D_{\mathbf{H}}(\widehat{\mathrm{AMISE}} - \mathrm{AMISE})(\hat{\mathbf{H}}).\end{aligned}$$

Since $\widehat{\mathrm{AMISE}}(\mathbf{H}) \xrightarrow{p} \mathrm{AMISE}(\mathbf{H})$ then $\hat{\mathbf{H}} \xrightarrow{p} \mathbf{H}_{\mathrm{AMISE}}$ as $n \to \infty$ and so

$$D_{\mathbf{H}}(\widehat{\mathrm{AMISE}} - \mathrm{AMISE})(\hat{\mathbf{H}}) = [\mathbf{I}_{d^*} + o_p(\mathbf{J}_{d^*})]D_{\mathbf{H}}(\widehat{\mathrm{AMISE}} - \mathrm{AMISE})(\mathbf{H}_{\mathrm{AMISE}}).$$

This implies that

$$\begin{aligned}\mathrm{vech}(\hat{\mathbf{H}} - \mathbf{H}_{\mathrm{AMISE}}) = {} & -[\mathbf{I}_{d^*} + o_p(\mathbf{J}_{d^*})][D_{\mathbf{H}}^2\mathrm{AMISE}(\mathbf{H}_{\mathrm{AMISE}})]^{-1} \\ & \times D_{\mathbf{H}}(\widehat{\mathrm{AMISE}} - \mathrm{AMISE})(\mathbf{H}_{\mathrm{AMISE}}).\end{aligned}$$

Taking expectations and variances, respectively, completes the proof. □

If $\mathrm{MSE}(\mathrm{vech}\,\hat{\mathbf{H}}) = O(\mathbf{J}_{d^*}n^{-2\beta})(\mathrm{vech}\,\mathbf{H}_{\mathrm{AMISE}})(\mathrm{vech}^T\mathbf{H}_{\mathrm{AMISE}})$ then $\hat{\mathbf{H}}$ has relative rate $n^{-\beta}$. Hence Lemma 1 allows the relative rate for $\hat{\mathbf{H}}$ to $\mathbf{H}_{\mathrm{AMISE}}$ to be computed from knowledge of mean and covariance matrix of $D_{\mathbf{H}}(\widehat{\mathrm{AMISE}} - \mathrm{AMISE})(\mathbf{H}_{\mathrm{AMISE}})$. Naturally, this lemma can be adapted to consider convergence to $\mathbf{H}_{\mathrm{MISE}}$ by replacing all references to AMISE by MISE. Nonetheless, it is generally simpler to consider convergence to $\mathbf{H}_{\mathrm{AMISE}}$ and then examine whether the discrepancy between $\mathbf{H}_{\mathrm{MISE}}$ and its asymptotic form is significant.

3. Relative rates of convergence for plug-in selectors

Recall that for plug-in selectors, PI is the estimator of AMISE. Now

$$(\mathrm{PI}-\mathrm{AMISE})(\mathbf{H}) = \tfrac{1}{4}\mu_2(K)^2(\mathrm{vech}^T\mathbf{H})(\check{\mathbf{\Psi}}_4 - \mathbf{\Psi}_4)(\mathrm{vech}\,\mathbf{H})[1+o_p(1)]$$

so that

$$\mathbb{E}[D_{\mathbf{H}}(\mathrm{PI}-\mathrm{AMISE})(\mathbf{H})] = \tfrac{1}{2}\mu_2(K)^2[\mathbf{I}_{d^*}+o(\mathbf{J}_{d^*})](\mathbb{E}\check{\mathbf{\Psi}}_4 - \mathbf{\Psi}_4)(\mathrm{vech}\,\mathbf{H}),$$

$$\mathrm{Var}[D_{\mathbf{H}}(\mathrm{PI}-\mathrm{AMISE})(\mathbf{H})] = \tfrac{1}{4}\mu_2(K)^4[\mathbf{I}_{d^*}+o(\mathbf{J}_{d^*})]\mathrm{Var}[\check{\mathbf{\Psi}}_4(\mathrm{vech}\,\mathbf{H})].$$

Theorem 1. *Assume* (A1)–(A3) *from Lemma* 1. *Assume also that* $K^{(\mathbf{r})}$ *is square integrable, and that if* $|\mathbf{r}|=4$ *then* $K^{(\mathbf{r})}(\mathbf{0})=1$ *if all elements of* $\mathbf{r}$ *are even and* $K^{(\mathbf{r})}(\mathbf{0})=0$ *otherwise. If* $\check{\mathbf{H}}_{\mathrm{AMSE}}$ *and* $\check{\mathbf{H}}_{\mathrm{SAMSE}}$ *denote, respectively, the AMSE and SAMSE plug-in bandwidth selectors described in Section* 2, *then*:

(i) *The relative rate of convergence of* $\check{\mathbf{H}}_{\mathrm{AMSE}}$ *to* $\mathbf{H}_{\mathrm{AMISE}}$ *is* $n^{-4/(d+12)}$.
(ii) *The relative rate of convergence of* $\check{\mathbf{H}}_{\mathrm{SAMSE}}$ *to* $\mathbf{H}_{\mathrm{AMISE}}$ *is* $n^{-2/(d+6)}$.

Remark 1. The additional conditions on K are satisfied by most common kernels including the Gaussian.

Remark 2. Result (i) is implicit in the work of Wand and Jones [12].

Remark 3. The asymptotic properties of $\check{\mathbf{H}}_{\mathrm{AMSE}}$ are superior to those of $\check{\mathbf{H}}_{\mathrm{SAMSE}}$. Nonetheless, the difference in rates of convergence is not great. In particular, for the important bivariate case the relative rate of convergence to $\mathbf{H}_{\mathrm{AMISE}}$ for $\check{\mathbf{H}}_{\mathrm{AMSE}}$ is $n^{-2/7}$ and for $\check{\mathbf{H}}_{\mathrm{SAMSE}}$ is $n^{-1/4}$. Even for a sample of size $n=100,000$ the ratio of $n^{-2/7}$ to $n^{-1/4}$ is only about 1.5, so comparison of the convergence rates alone will provide little guidance as to whether AMSE or SAMSE approaches should be preferred in practice.

Remark 4. The relative rate of convergence for a plug-in selector of a diagonal bandwidth matrix $n^{-\min(8,d+4)/(2d+12)}$, as demonstrated by Wand and Jones [12]. This rate is faster than those for the full bandwidth selectors. Intuitively speaking, this indicates that choosing the orientation of the kernel functions is the most difficult aspect of the bandwidth selection problem for both AMSE and SAMSE plug-in methods.

Remark 5. It is straightforward to show that

$$\mathrm{vech}(\mathbf{H}_{\mathrm{AMISE}}-\mathbf{H}_{\mathrm{MISE}}) = O(\mathbf{J}_{d^*}n^{-2/(d+4)})\mathrm{vech}\,\mathbf{H}_{\mathrm{MISE}}$$

so that the discrepancy between $\mathbf{H}_{\mathrm{AMISE}}$ and $\mathbf{H}_{\mathrm{MISE}}$ is asymptotically negligible in comparison to the relative rate of convergence of $\check{\mathbf{H}}_{\mathrm{SAMSE}}$ to $\mathbf{H}_{\mathrm{AMISE}}$. However, the

discrepancy between $\mathbf{H}_{\mathrm{AMISE}}$ and $\mathbf{H}_{\mathrm{MISE}}$ dominates the AMSE rate from Theorem 1 for $d>4$.

Proof of Theorem 1. From Wand and Jones [12] we know that the estimator $\check{\psi}_{\boldsymbol{r}}$ has slowest rate if at least one element of $\boldsymbol{r}$ is odd because it is impossible to annihilate the leading term in the bias in this instance. When $|\boldsymbol{r}|=4$ the optimal pilot bandwidth for these functional estimators is $g_{\boldsymbol{r},\mathrm{AMSE}}=O(n^{-2/(d+12)})$, giving

$$\operatorname{Bias}\check{\psi}_{\boldsymbol{r}}(g_{\boldsymbol{r},\mathrm{AMSE}})=O(g^2_{\boldsymbol{r},\mathrm{AMSE}})=O(n^{-4/(d+12)}),$$

$$\operatorname{Var}\check{\psi}_{\boldsymbol{r}}(g_{\boldsymbol{r},\mathrm{AMSE}})=O(n^{-2}g^{-d-8}_{\boldsymbol{r},\mathrm{AMSE}})=O(n^{-8/(d+12)}).$$

It follows that

$$\mathbb{E}[D_{\mathbf{H}}(\mathrm{PI}-\mathrm{AMISE})(\mathbf{H}_{\mathrm{AMISE}})]=O(\mathbf{J}_{d^*}n^{-4/(d+12)})\operatorname{vech}\mathbf{H}_{\mathrm{AMISE}} \tag{7}$$

and

$$\begin{aligned}&\operatorname{Var}[D_{\mathbf{H}}(\mathrm{PI}-\mathrm{AMISE})(\mathbf{H}_{\mathrm{AMISE}})]\\ &\quad=O(\mathbf{J}_{d^*}n^{-8/(d+12)})(\operatorname{vech}\mathbf{H}_{\mathrm{AMISE}})(\operatorname{vech}^T\mathbf{H}_{\mathrm{AMISE}}).\end{aligned} \tag{8}$$

The Hessian matrix

$$\begin{aligned}D^2_{\mathbf{H}}\mathrm{AMISE}\hat{f}(\cdot;\mathbf{H})=&\tfrac{1}{4}n^{-1}(4\pi)^{-d/2}|\mathbf{H}|^{-1/2}\mathbf{D}_d^T(\mathbf{H}^{-1}\otimes\mathbf{I}_d)\\ &\times[(\operatorname{vec}\mathbf{I}_d)(\operatorname{vec}^T\mathbf{I}_d)+2\mathbf{I}_{d^2}](\mathbf{I}_d\otimes\mathbf{H}^{-1})\mathbf{D}_d+\tfrac{1}{2}\boldsymbol{\Psi}_4\end{aligned}$$

converges to a constant, positive-definite matrix as $n\to\infty$. Here vec is the vector operator, so that vec $\mathbf{H}$ is concatenation of the columns of $\mathbf{H}$. The duplication matrix of order d is $\mathbf{D}_d$ and it relates the vec and vech operators in the following ways:

$$\operatorname{vec}\mathbf{H}=\mathbf{D}_d\operatorname{vech}\mathbf{H},$$

$$\mathbf{D}_d^T\operatorname{vec}\mathbf{H}=\operatorname{vech}(\mathbf{H}+\mathbf{H}^T-\operatorname{dg}\mathbf{H}).$$

Also $\otimes$ is the Kronecker (or tensor) product operator between two matrices. The proof of part (i) follows immediately by substituting (7) and (8) into the expansion of AMSE(vech $\check{\mathbf{H}}$) obtained from Lemma 1.

From Duong and Hazelton [1], the SAMSE pilot bandwidth is $g_{4,\mathrm{SAMSE}}=O(n^{-1/(d+6)})$. Straightforward calculations then give

$$\operatorname{Bias}\check{\psi}_{\boldsymbol{r}}(g_{j,\mathrm{SAMSE}})=O(g^2_{j,\mathrm{SAMSE}})=O(n^{-2/(d+6)})$$

when it follows that

$$\mathbb{E}[D_{\mathbf{H}}(\mathrm{PI}-\mathrm{AMISE})(\mathbf{H}_{\mathrm{AMISE}})]=O(\mathbf{J}_{d^*}n^{-2/(d+6)})\operatorname{vech}\mathbf{H}_{\mathrm{AMISE}}. \tag{9}$$

It can be shown that in the SAMSE plug-in method the variance of $\check{\psi}_{\boldsymbol{r}}$ is dominated by the leading term of the squared bias. Part (ii) then follows by substituting (9) into the result of Lemma 1, and noting the asymptotic constancy of the Hessian matrix as for part (i). □

4. Relative rates of convergence for BCV selectors

In this section we compute the relative rate of the BCV selector when K is Gaussian. The results can be extended to more general kernel functions at the expense of more complex proofs.

Theorem 2. *Assume* (A1)–(A2) *of Lemma* 1, *and that K is Gaussian. If* $\tilde{\mathbf{H}}$ *denotes the BCV selector, minimizing* (3), *then the relative rate of convergence of* $\tilde{\mathbf{H}}$ *to* $\mathbf{H}_{\mathrm{AMISE}}$ *is* $n^{-\min(d,4)/(2d+8)}$.

Remark 1. The BCV selector rate is slower than both AMSE and SAMSE plug-in rates.

Remark 2. The rate from Theorem 2 remains unchanged for the BCV selection when the bandwidth matrix is constrained to be diagonal, or even a constant multiple of the identity matrix. It follows that the relative rate of $n^{-d/(2d+8)}$ given by Sain et al. [7] for the BCV2 constrained matrices is incorrect for $d>4$. In particular, the rate does not tend to $n^{-1/2}$ as d becomes large. The form of the rate changes after the fourth dimension because the squared bias of the BCV selector then dominates; cf. Lemmas 2 and 3. The proof of Sain et al. does not keep proper track of second order bias terms which should lead to an additional term c_4h^5 in their equation 15.

Remark 3. The relative rate of convergence to $\mathbf{H}_{\mathrm{MISE}}$ is equal to that for convergence to $\mathbf{H}_{\mathrm{AMISE}}$ for the BCV selector.

The proof of Theorem 2 proceeds via a pair of lemmas.

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

$$\mathrm{ABias}(\mathrm{vech}\,\tilde{\mathbf{H}}) = O(\mathbf{J}_{d^*}n^{-2/(d+4)})\mathrm{vech}\,\mathbf{H}_{\mathrm{AMISE}}.$$

Proof. We start with

$$(\mathrm{BCV}-\mathrm{AMISE})(\mathbf{H}) = \tfrac{1}{4}(\mathrm{vech}^T\mathbf{H})(\tilde{\mathbf{\Psi}}_4(\mathbf{H})-\mathbf{\Psi}_4)(\mathrm{vech}\,\mathbf{H})[1+o_p(1)]$$

then

$$\mathbb{E}(\mathrm{BCV}-\mathrm{AMISE})(\mathbf{H}) = \tfrac{1}{4}(\mathrm{vech}^T\mathbf{H})(\mathbb{E}\tilde{\mathbf{\Psi}}_4(\mathbf{H})-\mathbf{\Psi}_4)(\mathrm{vech}\,\mathbf{H})[1+o_p(1)].$$

Now, $\mathbb{E}\tilde{\mathbf{\Psi}}_4(\mathbf{H})-\mathbf{\Psi}_4$ and is composed of elements of the type $\mathbb{E}\tilde{\psi}_{\boldsymbol{r}}(\mathbf{H})-\psi_{\boldsymbol{r}}$. As

$$\mathbb{E}\tilde{\psi}_{\boldsymbol{r}}(\mathbf{H})-\psi_{\boldsymbol{r}} = \frac{1}{2}\int \mathrm{tr}(\mathbf{H}D^2f(\boldsymbol{x}))f^{(\boldsymbol{r})}(\boldsymbol{x})\,d\boldsymbol{x}$$

(following Wand and Jones [13, pp. 67–70], for example) thus $\mathbb{E}(\mathrm{BCV}-\mathrm{AMISE})(\mathbf{H}) = O(||\mathrm{vech}\,\mathbf{H}||^3)$ and

$$\mathbb{E}[D_{\mathbf{H}}(\mathrm{BCV}-\mathrm{AMISE})(\mathbf{H}_{\mathrm{AMISE}})] = O(\mathbf{J}_{d^*}n^{-2/(d+4)})\mathrm{vech}\,\mathbf{H}_{\mathrm{AMISE}},$$

as $\mathbf{H}_{\mathrm{AMISE}} = O(\mathbf{J}_d n^{-2/(d+4)})$. □

Lemma 3. *Under the conditions of Theorem* 2,

$$\mathrm{AVar}(\mathrm{vech}\,\tilde{\mathbf{H}}) = O(\mathbf{J}_{d^*} n^{-d/(d+4)})(\mathrm{vech}\,\mathbf{H}_{\mathrm{AMISE}})(\mathrm{vech}^T \mathbf{H}_{\mathrm{AMISE}}).$$

Proof. Let $\boldsymbol{y} = \mathrm{vech}\,\mathbf{H}$ and $\mathbf{A}(\boldsymbol{y}) = \tilde{\boldsymbol{\Psi}}_4(\mathbf{H})$. We have

$$\begin{aligned} d(\boldsymbol{y}^T \mathbf{A}(\boldsymbol{y})\boldsymbol{y}) &= d(\boldsymbol{y}^T \mathbf{A}(\boldsymbol{y}))\boldsymbol{y} + \boldsymbol{y}^T \mathbf{A}(\boldsymbol{y})\, d\boldsymbol{y} \\ &= [(d\boldsymbol{y}^T)\mathbf{A}(\boldsymbol{y}) + \boldsymbol{y}^T d\mathbf{A}(\boldsymbol{y})]\boldsymbol{y} + \boldsymbol{y}^T \mathbf{A}(\boldsymbol{y})\, d\boldsymbol{y} \\ &= 2\boldsymbol{y}^T \mathbf{A}(\boldsymbol{y})\, d\boldsymbol{y} + \mathrm{vec}^T(\boldsymbol{y}\boldsymbol{y}^T)\, d\mathrm{vec}\,\mathbf{A}(\boldsymbol{y}) \end{aligned}$$

as $(d\boldsymbol{y}^T)\mathbf{A}(\boldsymbol{y})\boldsymbol{y} = \boldsymbol{y}^T \mathbf{A}(\boldsymbol{y})\, d\boldsymbol{y}$ and $\boldsymbol{y}^T d\mathbf{A}(\boldsymbol{y})\boldsymbol{y} = \mathrm{tr}(\boldsymbol{y}\boldsymbol{y}^T d\mathbf{A}(\boldsymbol{y})) = \mathrm{vec}^T(\boldsymbol{y}\boldsymbol{y}^T)\, d\mathrm{vec}\,\mathbf{A}(\boldsymbol{y})$. Then using the first identification table of Magnus and Neudecker [6, p. 176] the derivative is

$$\begin{aligned} D_{\boldsymbol{y}}(\boldsymbol{y}^T \mathbf{A}(\boldsymbol{y})\boldsymbol{y}) &= 2\mathbf{A}(\boldsymbol{y})\boldsymbol{y} + [D_{\boldsymbol{y}}\mathbf{A}(\boldsymbol{y})]^T \mathrm{vec}(\boldsymbol{y}\boldsymbol{y}^T) \\ &= 2\mathbf{A}(\boldsymbol{y})\boldsymbol{y} + [D_{\boldsymbol{y}}\mathbf{A}(\boldsymbol{y})]^T (\boldsymbol{y} \otimes \mathbf{I}_{d^*})\boldsymbol{y}. \end{aligned}$$

Using this, the derivative of $\mathrm{BCV} - \mathrm{AMISE}$ is

$$\begin{aligned} & D_{\mathbf{H}}(\mathrm{BCV} - \mathrm{AMISE})(\mathbf{H}) \\ &= D_{\mathbf{H}}[\tfrac{1}{4}(\mathrm{vech}^T \mathbf{H})(\tilde{\boldsymbol{\Psi}}_4(\mathbf{H}) - \boldsymbol{\Psi}_4)(\mathrm{vech}\,\mathbf{H})] \\ &= \tfrac{1}{2}(\tilde{\boldsymbol{\Psi}}_4(\mathbf{H}) - \boldsymbol{\Psi}_4)(\mathrm{vech}\,\mathbf{H}) + \tfrac{1}{4}[D_{\mathbf{H}}\tilde{\boldsymbol{\Psi}}_4(\mathbf{H})]^T (\mathrm{vech}\,\mathbf{H} \otimes \mathbf{I}_{d^*})(\mathrm{vech}\,\mathbf{H}). \end{aligned}$$

Then the variance of $D_{\mathbf{H}}(\mathrm{BCV} - \mathrm{AMISE})(\mathbf{H})$ will be of the same rate as the minimum rate of $\mathrm{Var}[\tilde{\boldsymbol{\Psi}}_4(\mathbf{H})(\mathrm{vech}\,\mathbf{H})]$ and $\mathrm{Var}\{[D_{\mathbf{H}}\tilde{\boldsymbol{\Psi}}_4(\mathbf{H})]^T (\mathrm{vech}\,\mathbf{H} \otimes \mathbf{I}_{d^*})\}$.

The first of these is

$$\begin{aligned} \mathrm{Var}[\tilde{\boldsymbol{\Psi}}_4(\mathbf{H})(\mathrm{vech}\,\mathbf{H})] &= \mathbb{E}[\tilde{\boldsymbol{\Psi}}_4(\mathbf{H})(\mathrm{vech}\,\mathbf{H})(\mathrm{vech}^T \mathbf{H})\tilde{\boldsymbol{\Psi}}_4(\mathbf{H})] \\ &\quad - [\mathbb{E}\tilde{\boldsymbol{\Psi}}_4(\mathbf{H})(\mathrm{vech}\,\mathbf{H})][(\mathrm{vech}^T \mathbf{H})\mathbb{E}\tilde{\boldsymbol{\Psi}}_4(\mathbf{H})]. \end{aligned}$$

Now $\mathbb{E}[\tilde{\boldsymbol{\Psi}}_4(\mathbf{H})\tilde{\boldsymbol{\Psi}}_4(\mathbf{H})] - [\mathbb{E}\tilde{\boldsymbol{\Psi}}_4(\mathbf{H})][\mathbb{E}\tilde{\boldsymbol{\Psi}}_4(\mathbf{H})]$ contains elements of the type

$$\begin{aligned} \mathbb{E}[\tilde{\psi}_{\boldsymbol{r}_1}(\mathbf{H})\tilde{\psi}_{\boldsymbol{r}_2}(\mathbf{H})] - [\mathbb{E}\tilde{\psi}_{\boldsymbol{r}_1}(\mathbf{H})][\mathbb{E}\tilde{\psi}_{\boldsymbol{r}_2}(\mathbf{H})] &= \mathrm{Cov}[\tilde{\psi}_{\boldsymbol{r}_1}(\mathbf{H}), \tilde{\psi}_{\boldsymbol{r}_2}(\mathbf{H})] \\ &= O(\min\{\mathrm{Var}\,\tilde{\psi}_{\boldsymbol{r}_1}(\mathbf{H}), \mathrm{Var}\,\tilde{\psi}_{\boldsymbol{r}_2}(\mathbf{H})\}). \end{aligned}$$

Following Wand and Jones [13, pp. 67–70], for example, we know that $\mathrm{Var}\,\tilde{\psi}_{\boldsymbol{r}}(\mathbf{H}) = O(n^{-2}|\mathbf{H}|^{1/2}||\mathrm{vech}\,\mathbf{H}||^{-|\boldsymbol{r}|})$ provided that $n^{-2}|\mathbf{H}|^{-1/2}||\mathrm{vech}\,\mathbf{H}||^{-|\boldsymbol{r}|} \to 0$ as $n \to \infty$. This is true for $\mathbf{H} = O(\mathbf{J}_d n^{-2/(d+4)})$ and $|\boldsymbol{r}| = 4$. Thus it yields

$$\begin{aligned} & \mathrm{Var}[\tilde{\boldsymbol{\Psi}}_4(\mathbf{H}_{\mathrm{AMISE}})(\mathrm{vech}\,\mathbf{H}_{\mathrm{AMISE}})] \\ &= O(\mathbf{J}_{d^*} n^{-d/(d+4)})(\mathrm{vech}\,\mathbf{H}_{\mathrm{AMISE}})(\mathrm{vech}^T \mathbf{H}_{\mathrm{AMISE}}). \end{aligned}$$

The second term is

$$\begin{aligned}
&\mathrm{Var}\{[D_{\mathbf{H}}\tilde{\mathbf{\Psi}}_4(\mathbf{H})]^T(\mathrm{vech}\,\mathbf{H}\otimes\mathbf{I}_{d^*})(\mathrm{vech}\,\mathbf{H})\}\\
&\quad=\ \mathbb{E}\{[D_{\mathbf{H}}\tilde{\mathbf{\Psi}}_4(\mathbf{H})]^T(\mathrm{vech}\,\mathbf{H}\otimes\mathbf{I}_{d^*})(\mathrm{vech}\,\mathbf{H})(\mathrm{vech}^T\mathbf{H})\\
&\qquad\times(\mathrm{vech}^T\mathbf{H}\otimes\mathbf{I}_{d^*})D_{\mathbf{H}}\tilde{\mathbf{\Psi}}_4(\mathbf{H})\}\\
&\qquad-\ \mathbb{E}[D_{\mathbf{H}}\tilde{\mathbf{\Psi}}_4(\mathbf{H})]^T(\mathrm{vech}\,\mathbf{H}\otimes\mathbf{I}_{d^*})(\mathrm{vech}\,\mathbf{H})(\mathrm{vech}^T\mathbf{H})\\
&\qquad\times(\mathrm{vech}^T\mathbf{H}\otimes\mathbf{I}_{d^*})\mathbb{E}[D_{\mathbf{H}}\tilde{\mathbf{\Psi}}_4(\mathbf{H})].
\end{aligned}$$

Now $\mathbb{E}[D_{\mathbf{H}}\tilde{\mathbf{\Psi}}_4(\mathbf{H})]^T[D_{\mathbf{H}}\tilde{\mathbf{\Psi}}_4(\mathbf{H})]-\mathbb{E}[D_{\mathbf{H}}\tilde{\mathbf{\Psi}}_4(\mathbf{H})]^T\mathbb{E}[D_{\mathbf{H}}\tilde{\mathbf{\Psi}}_4(\mathbf{H})]$ contains blocks of elements of the type

$$\begin{aligned}
&\sum_{\boldsymbol{r}}\mathbb{E}\{[D_{\mathbf{H}}\tilde{\psi}_{\boldsymbol{r}}(\mathbf{H})][D_{\mathbf{H}}\tilde{\psi}_{\boldsymbol{r}}(\mathbf{H})]^T\}-\mathbb{E}[D_{\mathbf{H}}\tilde{\psi}_{\boldsymbol{r}}(\mathbf{H})]\mathbb{E}[D_{\mathbf{H}}\tilde{\psi}_{\boldsymbol{r}}(\mathbf{H})]^T\\
&\quad=\ \sum_{\boldsymbol{r}}\mathrm{Var}\,D_{\mathbf{H}}\tilde{\psi}_{\boldsymbol{r}}(\mathbf{H})\\
&\quad=\ \sum_{\boldsymbol{r}}\mathrm{Var}\left[n^{-1}(n-1)^{-1}\sum_{i=1}^{n}\sum_{\substack{j=1\\ j\neq 1}}^{n}D_{\mathbf{H}}\phi_{\mathbf{H}}^{(\boldsymbol{r})}(\boldsymbol{X}_i-\boldsymbol{X}_j)\right],
\end{aligned}\tag{10}$$

where $\phi_{\mathbf{\Sigma}}(\cdot)$ is the multivariate normal density with mean vector $\mathbf{0}$ and covariance matrix $\mathbf{\Sigma}$. Using the normal kernel allows us to compute the derivative of $\phi_{\mathbf{H}}^{(\boldsymbol{r})}$ more easily:

$$\begin{aligned}
D_{\mathbf{H}}\phi_{\mathbf{H}}^{(\boldsymbol{r})}(\boldsymbol{x})&=\frac{\partial^{|\boldsymbol{r}|}}{\partial x_1^{r_1}\ldots\partial x_d^{r_d}}D_{\mathbf{H}}\phi_{\mathbf{H}}(\boldsymbol{x})\\
&=\frac{\partial^{|\boldsymbol{r}|}}{\partial x_1^{r_1}\ldots\partial x_d^{r_d}}\tfrac{1}{2}\phi_{\mathbf{H}}(\boldsymbol{x})\mathbf{D}_d^T\mathrm{vec}[\mathbf{H}^{-1}\boldsymbol{x}\boldsymbol{x}^T\mathbf{H}^{-1}-\mathbf{H}^{-1}]\\
&=\tfrac{1}{2}\phi_{\mathbf{H}}^{(\boldsymbol{r})}(\boldsymbol{x})\mathbf{D}_d^T\mathrm{vec}[\mathbf{H}^{-1}\boldsymbol{x}\boldsymbol{x}^T\mathbf{H}^{-1}]\\
&\quad+\tfrac{1}{2}\phi_{\mathbf{H}}(\boldsymbol{x})\mathbf{D}_d^T\mathrm{vec}\left[\mathbf{H}^{-1}\frac{\partial^{|\boldsymbol{r}|}}{\partial x_1^{r_1}\ldots\partial x_d^{r_d}}(\boldsymbol{x}\boldsymbol{x}^T)\mathbf{H}^{-1}\right]\\
&\quad-\tfrac{1}{2}\phi_{\mathbf{H}}^{(\boldsymbol{r})}(\boldsymbol{x})\mathbf{D}_d^T\mathrm{vec}\,\mathbf{H}^{-1}.
\end{aligned}$$

If we look at $|\boldsymbol{r}|=4$ then

$$\begin{aligned}
&\sum_{i=1}^{n}\sum_{\substack{j=1\\ j\neq 1}}^{n}\phi_{\mathbf{H}}(\boldsymbol{X}_i-\boldsymbol{X}_j)\frac{\partial^{|\boldsymbol{r}|}}{\partial x_1^{r_1}\ldots\partial x_d^{r_d}}[(\boldsymbol{X}_i-\boldsymbol{X}_j)(\boldsymbol{X}_i-\boldsymbol{X}_j)^T]\\
&\quad=\ \sum_{i=1}^{n}\sum_{\substack{j=1\\ j\neq 1}}^{n}\phi_{\mathbf{H}}(\boldsymbol{X}_i-\boldsymbol{X}_j)\mathbf{C}_0,
\end{aligned}$$

where

$$\mathbf{C}_0 = \begin{cases} 2\mathbf{E}_{kk} + 2\mathbf{E}_{\ell\ell} & \text{if } \boldsymbol{r} = 2\mathbf{e}_k + 2\mathbf{e}_\ell,\ k,\ell = 1,2\ldots,d, \\ \mathbf{0} & \text{otherwise} \end{cases}$$

and $\mathbf{E}_{ij}$ is a $d^* \times d^*$ elementary matrix which has 1 as its (i,j)-th element and 0 elsewhere. So then

$$\begin{aligned} & n^{-1}(n-1)^{-1} \sum_{i=1}^{n} \sum_{\substack{j=1 \\ j\neq 1}}^{n} D_{\mathbf{H}} \phi_{\mathbf{H}}^{(\boldsymbol{r})}(\boldsymbol{X}_i - \boldsymbol{X}_j) \\ &\quad = \tfrac{1}{2}\mathbf{D}_d^T(\mathbf{H}^{-1} \otimes \mathbf{H}^{-1}) \operatorname{vec} \tilde{\boldsymbol{\psi}}_{\boldsymbol{r}}^{[2]}(\mathbf{H}) + \tfrac{1}{2}\tilde{\psi}_{\mathbf{0}}(\mathbf{H})\mathbf{D}_d^T(\mathbf{H}^{-1} \otimes \mathbf{H}^{-1}) \operatorname{vec} \mathbf{C}_0 \\ &\qquad - \tfrac{1}{2}\tilde{\psi}_{\boldsymbol{r}}(\mathbf{H})\mathbf{D}_d^T \operatorname{vec} \mathbf{H}^{-1} \end{aligned} \tag{11}$$

using $\operatorname{vec}(\mathbf{ABC}) = (\mathbf{C}^T \otimes \mathbf{A})\operatorname{vec} \mathbf{B}$ and where

$$\operatorname{vec} \tilde{\boldsymbol{\psi}}_{\boldsymbol{r}}^{[2]}(\mathbf{H}) = n^{-1}(n-1)^{-1} \sum_{i=1}^{n} \sum_{\substack{j=1 \\ j\neq 1}}^{n} \phi_{\mathbf{H}}^{(\boldsymbol{r})}(\boldsymbol{X}_i - \boldsymbol{X}_j) \operatorname{vec}[(\boldsymbol{X}_i - \boldsymbol{X}_j)(\boldsymbol{X}_i - \boldsymbol{X}_j)^T].$$

Now the order of the variance of the left-hand side of Eq. (11) is the minimum order of the three terms on the right-hand side. We know that $\operatorname{Var} \tilde{\psi}_{\boldsymbol{r}}(\mathbf{H}) = O(n^{-2}|\mathbf{H}|^{-1/2}||\operatorname{vech} \mathbf{H}||^{-|\boldsymbol{r}|})$ so the second term of the right-hand side is

$$\begin{aligned} & \operatorname{Var}[\tilde{\psi}_{\boldsymbol{r}}(\mathbf{H}_{\mathrm{AMISE}})\mathbf{D}_d^T \operatorname{vec} \mathbf{H}_{\mathrm{AMISE}}^{-1}] \\ &\quad = O(\mathbf{J}_{d^*} n^{-2} |\mathbf{H}_{\mathrm{AMISE}}|^{-1/2} ||\operatorname{vech} \mathbf{H}_{\mathrm{AMISE}}||^{-4})(\operatorname{vech} \mathbf{H}_{\mathrm{AMISE}})(\operatorname{vech}^T \mathbf{H}_{\mathrm{AMISE}}) \\ &\quad = O(\mathbf{J}_{d^*} n^{(-d+4)/(d+4)}) \end{aligned} \tag{12}$$

and the third term is

$$\begin{aligned} & \operatorname{Var}[\tilde{\psi}_{\mathbf{0}}(\mathbf{H}_{\mathrm{AMISE}})\mathbf{D}_d^T(\mathbf{H}_{\mathrm{AMISE}}^{-1} \otimes \mathbf{H}_{\mathrm{AMISE}}^{-1}) \operatorname{vec} \mathbf{C}_0] \\ &\quad = O(\mathbf{J}_{d^*} n^{-2} |\mathbf{H}_{\mathrm{AMISE}}|^{-1/2})(\operatorname{vech} \mathbf{H}_{\mathrm{AMISE}}^{-2})(\operatorname{vech}^T \mathbf{H}_{\mathrm{AMISE}}^{-2}) \\ &\quad = O(\mathbf{J}_{d^*} n^{-d/(d+4)}). \end{aligned} \tag{13}$$

We will now examine the first term of the right-hand side of Eq. (11). As the summand of the double sum of $\operatorname{vec} \tilde{\boldsymbol{\psi}}_{\boldsymbol{r}}^{[2]}(\mathbf{H})$ is a symmetric function so

$$\begin{aligned} \operatorname{Var} \operatorname{vec} \tilde{\boldsymbol{\psi}}_{\boldsymbol{r}}^{[2]}(\mathbf{H}) = {} & 2n^{-2} \operatorname{Var} \phi_{\mathbf{H}}^{(\boldsymbol{r})}(\boldsymbol{X}_1 - \boldsymbol{X}_2) \operatorname{vec}[(\boldsymbol{X}_1 - \boldsymbol{X}_2)(\boldsymbol{X}_1 - \boldsymbol{X}_2)^T] \\ & + 4n^{-1} \operatorname{Cov}\{\phi_{\mathbf{H}}^{(\boldsymbol{r})}(\boldsymbol{X}_1 - \boldsymbol{X}_2) \operatorname{vec}[(\boldsymbol{X}_1 - \boldsymbol{X}_2)(\boldsymbol{X}_1 - \boldsymbol{X}_2)^T], \\ & \phi_{\mathbf{H}}^{(\boldsymbol{r})}(\boldsymbol{X}_2 - \boldsymbol{X}_3) \operatorname{vec}^T[(\boldsymbol{X}_2 - \boldsymbol{X}_3)(\boldsymbol{X}_2 - \boldsymbol{X}_3)^T]\}. \end{aligned}$$

The first term of $\operatorname{Var}\operatorname{vec}\tilde{\boldsymbol{\psi}}_{\boldsymbol{r}}^{[2]}(\mathbf{H})$ is

$$\begin{aligned}&\operatorname{Var}\{\phi_{\mathbf{H}}^{(\boldsymbol{r})}(\boldsymbol{X}_1-\boldsymbol{X}_2)\operatorname{vec}[(\boldsymbol{X}_1-\boldsymbol{X}_2)(\boldsymbol{X}_1-\boldsymbol{X}_2)^T]\}\\&\quad= O(\mathbf{J}_{d^2}|\mathbf{H}|^{-1/2}||\operatorname{vech}\mathbf{H}||^{-|\boldsymbol{r}|})(\operatorname{vec}\mathbf{H})(\operatorname{vec}^T\mathbf{H})\end{aligned}$$

as

$$\begin{aligned}&\mathbb{E}\{\phi_{\mathbf{H}}^{(\boldsymbol{r})}(\boldsymbol{X}_1-\boldsymbol{X}_2)\operatorname{vec}[(\boldsymbol{X}_1-\boldsymbol{X}_2)(\boldsymbol{X}_1-\boldsymbol{X}_2)^T]\}\\&\quad=\int_{\mathbb{R}^{2d}}\phi_{\mathbf{H}}^{(\boldsymbol{r})}(\boldsymbol{x}-\boldsymbol{y})\operatorname{vec}[(\boldsymbol{x}-\boldsymbol{y})(\boldsymbol{x}-\boldsymbol{y})^T]f(\boldsymbol{x})f(\boldsymbol{y})\,d\boldsymbol{x}\,d\boldsymbol{y}\\&\quad=\int_{\mathbb{R}^{2d}}\phi_{\mathbf{H}}(\boldsymbol{x}-\boldsymbol{y})\operatorname{vec}[(\boldsymbol{x}-\boldsymbol{y})(\boldsymbol{x}-\boldsymbol{y})^T]f(\boldsymbol{x})f^{(\boldsymbol{r})}(\boldsymbol{y})\,d\boldsymbol{x}\,d\boldsymbol{y}\\&\quad=\int_{\mathbb{R}^{2d}}\phi_{\mathbf{I}}(\boldsymbol{w})\operatorname{vec}(\mathbf{H}^{1/2}\boldsymbol{w}\boldsymbol{w}^T\mathbf{H}^{1/2})f(\boldsymbol{y}+\mathbf{H}^{1/2}\boldsymbol{w})\,d\boldsymbol{w}\,d\boldsymbol{y}\\&\quad=\int_{\mathbb{R}^{2d}}\phi_{\mathbf{I}}(\boldsymbol{w})\operatorname{vec}(\mathbf{H}^{1/2}\boldsymbol{w}\boldsymbol{w}^T\mathbf{H}^{1/2})[f(\boldsymbol{y})+O(||\operatorname{vech}\mathbf{H}||)]\,d\boldsymbol{w}\,d\boldsymbol{y}\\&\quad=\boldsymbol{\psi}_{\boldsymbol{r}}\operatorname{vec}\mathbf{H}+O(||\operatorname{vech}\mathbf{H}||)\operatorname{vec}\mathbf{H}\end{aligned}$$

and

$$\begin{aligned}&\mathbb{E}\{\phi_{\mathbf{H}}^{(\boldsymbol{r})}(\boldsymbol{X}_1-\boldsymbol{X}_2)^2\operatorname{vec}[(\boldsymbol{X}_1-\boldsymbol{X}_2)(\boldsymbol{X}_1-\boldsymbol{X}_2)^T]\operatorname{vec}^T[(\boldsymbol{X}_1-\boldsymbol{X}_2)(\boldsymbol{X}_1-\boldsymbol{X}_2)]\}\\&\quad=\int_{\mathbb{R}^{2d}}\phi_{\mathbf{H}}^{(\boldsymbol{r})}(\boldsymbol{x}-\boldsymbol{y})^2\operatorname{vec}[(\boldsymbol{x}-\boldsymbol{y})(\boldsymbol{x}-\boldsymbol{y})^T]\operatorname{vec}^T[(\boldsymbol{x}-\boldsymbol{y})(\boldsymbol{x}-\boldsymbol{y})^T]f(\boldsymbol{x})f(\boldsymbol{y})\,d\boldsymbol{x}\,d\boldsymbol{y}\\&\quad=\int_{\mathbb{R}^{2d}}[|\mathbf{H}|^{-1/2}\phi_{\mathbf{I}}^{(\boldsymbol{r})}(\mathbf{H}^{-1/2}(\boldsymbol{x}-\boldsymbol{y}))O(\mathbf{J}_{d^2}||\operatorname{vech}\mathbf{H}||^{-|\boldsymbol{r}|/2})]^2\operatorname{vec}[(\boldsymbol{x}-\boldsymbol{y})(\boldsymbol{x}-\boldsymbol{y})^T]\\&\qquad\times\operatorname{vec}^T[(\boldsymbol{x}-\boldsymbol{y})(\boldsymbol{x}-\boldsymbol{y})^T]f(\boldsymbol{x})f(\boldsymbol{y})\,d\boldsymbol{x}\,d\boldsymbol{y}\\&\quad=O(\mathbf{J}_{d^2}|\mathbf{H}|^{-1/2}||\operatorname{vech}\mathbf{H}||^{-|\boldsymbol{r}|})\int_{\mathbb{R}^{2d}}\phi_{\mathbf{I}}^{(\boldsymbol{r})}(\boldsymbol{w})^2\operatorname{vec}(\mathbf{H}^{1/2}\boldsymbol{w}\boldsymbol{w}^T\mathbf{H}^{1/2})\\&\qquad\times\operatorname{vec}^T(\mathbf{H}^{1/2}\boldsymbol{w}\boldsymbol{w}^T\mathbf{H}^{1/2})f(\boldsymbol{y}+\mathbf{H}^{1/2}\boldsymbol{w})f(\boldsymbol{y})\,d\boldsymbol{w}\,d\boldsymbol{y}\\&\quad=O(\mathbf{J}_{d^2}|\mathbf{H}|^{-1/2}||\operatorname{vech}\mathbf{H}||^{-|\boldsymbol{r}|})\int_{\mathbb{R}^{2d}}\phi_{\mathbf{I}}^{(\boldsymbol{r})}(\boldsymbol{w})^2(\mathbf{H}^{1/2}\otimes\mathbf{H}^{1/2})\operatorname{vec}(\boldsymbol{w}\boldsymbol{w}^T)\\&\qquad\times\operatorname{vec}^T(\boldsymbol{w}\boldsymbol{w}^T)(\mathbf{H}^{1/2}\otimes\mathbf{H}^{1/2})[f(\boldsymbol{y})+o(1)]f(\boldsymbol{y})\,d\boldsymbol{w}\,d\boldsymbol{y}\\&\quad=O(\mathbf{J}_{d^2}|\mathbf{H}|^{-1/2}||\operatorname{vech}\mathbf{H}||^{-|\boldsymbol{r}|})(\operatorname{vec}\mathbf{H})(\operatorname{vec}^T\mathbf{H}).\end{aligned}$$

The second term of $\operatorname{Var}\operatorname{vec}\tilde{\boldsymbol{\psi}}_{\boldsymbol{r}}^{[2]}(\mathbf{H})$ is

$$\begin{aligned}&\operatorname{Cov}\{\phi_{\mathbf{H}}^{(\boldsymbol{r})}(\boldsymbol{X}_1-\boldsymbol{X}_2)\operatorname{vec}[(\boldsymbol{X}_1-\boldsymbol{X}_2)(\boldsymbol{X}_1-\boldsymbol{X}_2)^T],\\&\quad\phi_{\mathbf{H}}^{(\boldsymbol{r})}(\boldsymbol{X}_2-\boldsymbol{X}_3)\operatorname{vec}^T[(\boldsymbol{X}_2-\boldsymbol{X}_3)(\boldsymbol{X}_2-\boldsymbol{X}_3)^T]\}=O(\mathbf{J}_{d^2})(\operatorname{vec}\mathbf{H})(\operatorname{vec}^T\mathbf{H})\end{aligned}$$

as

$$\begin{aligned}
&\mathbb{E}\{\phi_{\mathbf{H}}^{(r)}(\boldsymbol{X}_1-\boldsymbol{X}_2)\mathrm{vec}[(\boldsymbol{X}_1-\boldsymbol{X}_2)(\boldsymbol{X}_1-\boldsymbol{X}_2)^T]\\
&\quad\times\phi_{\mathbf{H}}^{(r)}(\boldsymbol{X}_2-\boldsymbol{X}_3)\mathrm{vec}^T[(\boldsymbol{X}_2-\boldsymbol{X}_3)(\boldsymbol{X}_2-\boldsymbol{X}_3)^T]\}\\
&\quad=\int_{\mathbb{R}^{3d}}\phi_{\mathbf{H}}^{(r)}(\boldsymbol{x}-\boldsymbol{y})\mathrm{vec}[(\boldsymbol{x}-\boldsymbol{y})(\boldsymbol{x}-\boldsymbol{y})^T]\phi_{\mathbf{H}}^{(r)}(\boldsymbol{y}-\boldsymbol{z})\mathrm{vec}^T[(\boldsymbol{y}-\boldsymbol{z})(\boldsymbol{y}-\boldsymbol{z})^T]\\
&\qquad\times f(\boldsymbol{x})f(\boldsymbol{y})f(\boldsymbol{z})\,d\boldsymbol{x}\,d\boldsymbol{y}\,d\boldsymbol{z}\\
&\quad=\int_{\mathbb{R}^{3d}}\phi_{\mathbf{H}}(\boldsymbol{x}-\boldsymbol{y})\mathrm{vec}[(\boldsymbol{x}-\boldsymbol{y})(\boldsymbol{x}-\boldsymbol{y})^T]\phi_{\mathbf{H}}(\boldsymbol{y}-\boldsymbol{z})\mathrm{vec}^T[(\boldsymbol{y}-\boldsymbol{z})(\boldsymbol{y}-\boldsymbol{z})^T]\\
&\qquad\times f^{(r)}(\boldsymbol{x})f^{(r)}(\boldsymbol{y})f(\boldsymbol{z})\,d\boldsymbol{x}\,d\boldsymbol{y}\,d\boldsymbol{z}\\
&\quad=\int_{\mathbb{R}^{3d}}\phi_{\mathbf{I}}(\boldsymbol{v})\phi_{\mathbf{I}}(\boldsymbol{w})\mathrm{vec}(\mathbf{H}^{1/2}\boldsymbol{v}\boldsymbol{v}^T\mathbf{H}^{1/2})\mathrm{vec}^T(\mathbf{H}^{1/2}\boldsymbol{w}\boldsymbol{w}^T\mathbf{H}^{1/2})\\
&\qquad\times f^{(r)}(\boldsymbol{y}+\mathbf{H}^{1/2}\boldsymbol{w})f^{(r)}(\boldsymbol{y})f(\boldsymbol{y}-\mathbf{H}^{1/2}\boldsymbol{w})\,d\boldsymbol{v}\,d\boldsymbol{w}\,d\boldsymbol{y}\\
&\quad=\ O(\mathbf{J}_{d^2})(\mathrm{vec}\,\mathbf{H})(\mathrm{vec}^T\mathbf{H})
\end{aligned}$$

which is the same order as $\mathbb{E}\{\phi_{\mathbf{H}}^{(r)}(\boldsymbol{X}_1-\boldsymbol{X}_2)\mathrm{vec}[(\boldsymbol{X}_1-\boldsymbol{X}_2)(\boldsymbol{X}_1-\boldsymbol{X}_2)^T]\}\ \mathbb{E}\{\phi_{\mathbf{H}}^{(r)}(\boldsymbol{X}_2-\boldsymbol{X}_3)\mathrm{vec}^T[(\boldsymbol{X}_2-\boldsymbol{X}_3)(\boldsymbol{X}_2-\boldsymbol{X}_3)^T]\}$. Putting these together yields

$$\begin{aligned}
&\mathrm{Var}[\mathbf{D}_d^T(\mathbf{H}_{\mathrm{AMISE}}^{-1}\otimes\mathbf{H}_{\mathrm{AMISE}}^{-1})\mathrm{vec}\,\tilde{\boldsymbol{\psi}}_r^{[2]}(\mathbf{H}_{\mathrm{AMISE}})]\\
&\quad=\ O(\mathbf{J}_{d^*}n^{-2}|\mathbf{H}_{\mathrm{AMISE}}|^{-1/2}||\mathrm{vech}\,\mathbf{H}_{\mathrm{AMISE}}||^{-4})(\mathrm{vech}\,\mathbf{H}_{\mathrm{AMISE}}^{-2})(\mathrm{vech}^T\mathbf{H}_{\mathrm{AMISE}}^{-2})\\
&\qquad+\ O(\mathbf{J}_{d^*}n^{-1})(\mathrm{vech}\,\mathbf{H}_{\mathrm{AMISE}})(\mathrm{vech}^T\mathbf{H}_{\mathrm{AMISE}})\\
&\quad=\ O(\mathbf{J}_{d^*}n^{(-d+4)/(d+4)}).
\end{aligned}\tag{14}$$

Eqs. (12)–(14) are the variances of the individual terms of the right-hand side of Eq. (11) so the variance of Eq. (11) is

$$\mathrm{Var}\left[n^{-1}(n-1)^{-1}\sum_{i=1}^{n}\sum_{\substack{j=1\\ j\neq 1}}^{n}D_{\mathbf{H}}\phi_{\mathbf{H}}^{(r)}(\boldsymbol{X}_i-\boldsymbol{X}_j)\right]=O(\mathbf{J}_{d^*}n^{(-d+4)/(d+4)})$$

which in turn implies that

$$\begin{aligned}
&\mathrm{Var}\{[D_{\mathbf{H}}\tilde{\boldsymbol{\Psi}}_4(\mathbf{H}_{\mathrm{AMISE}})]^T(\mathrm{vech}\,\mathbf{H}_{\mathrm{AMISE}}\otimes\mathbf{I}_{d^*})(\mathrm{vech}\,\mathbf{H}_{\mathrm{AMISE}})\}\\
&\quad=\ O(\mathbf{J}_{d^*}n^{(-d+4)/(d+4)})[(\mathrm{vech}\,\mathbf{H}_{\mathrm{AMISE}})(\mathrm{vech}^T\mathbf{H}_{\mathrm{AMISE}})]^2\\
&\quad=\ O(\mathbf{J}_{d^*}n^{-d/(d+4)})(\mathrm{vech}\,\mathbf{H}_{\mathrm{AMISE}})(\mathrm{vech}^T\mathbf{H}_{\mathrm{AMISE}}).
\end{aligned}$$

This is the same order as $\mathrm{Var}[\tilde{\boldsymbol{\Psi}}_4(\mathbf{H}_{\mathrm{AMISE}})(\mathrm{vech}\,\mathbf{H}_{\mathrm{AMISE}})]$, which is the other term in the variance of $D_{\mathbf{H}}(\mathrm{BCV}-\mathrm{AMISE})(\mathbf{H})$ i.e.

$$\begin{aligned}&\mathrm{Var}[D_{\mathbf{H}}(\mathrm{BCV}-\mathrm{AMISE})(\mathbf{H})]\\ &\quad= O(\mathbf{J}_{d^*}n^{-d/(d+4)})(\mathrm{vech}\,\mathbf{H}_{\mathrm{AMISE}})(\mathrm{vech}^T\mathbf{H}_{\mathrm{AMISE}}).\end{aligned}$$ □

Combining Lemmas 1–3, we have proved Theorem 2.

5. Discussion

In this paper we have described a general method for deriving relative rates of convergence for full bandwidth matrix selectors. This methodology has been applied to compute rates for the plug-in selector of Wand and Jones [12], the plug-in selector of Duong and Hazelton [1], and for a generalized form of the biased cross-validation selector of Sain et al. [7]. While these rates provide a guide towards the comparative performance of the bandwidth selectors in question, the usual caveats regarding the interpretation of asymptotic results within a finite sample setting apply. Simulation experiments and analyses of real data sets can provide insight into the behaviour of bandwidth matrix selectors for moderate sample sizes. Both [1] and [2] describe the results from such studies for bivariate data from a range of types of target density. The BCV bandwidth matrix selectors tended to perform less well than both AMSE and SAMSE plug-in selectors in these studies, which is in keeping with the theoretical results in this paper. See the first remark after Theorem 2.

Bandwidth matrix selection for data in more than two dimensions has not received much attention by way of numerical studies in the literature. This is largely a reflection of the decreased utility of kernel density estimation in high dimensions. For example, bivariate density estimates are useful for exploratory data analysis because they can be displayed using familiar contour or perspective ('wire frame') plots. Visualization of density estimates in higher dimensions is more difficult, although Scott [8] offers some ingenious approaches to the problem. Furthermore, the well-known 'curse of dimensionality' makes it more or less impossible to obtain reliable kernel density estimates in dimensions much higher than four without gigantic sample sizes. The full bandwidth matrix selection methods on which we have focused are most practicable for bivariate data, when one needs estimate only one additional smoothing parameter in comparison to a diagonal matrix approach. As a consequence, the algorithms for full matrix plug-in selection developed by Duong and Hazelton [1] require only a modest increase in computational cost in comparison to algorithms for diagonal bandwidth matrices when the data are bivariate. In the general d-dimensional case, full bandwidth matrices require specification of $d(d+1)/2$ parameters as opposed to just d for a diagonal bandwidth matrix. This casts doubt upon the utility of full bandwidth matrices for d larger than three or four, although this is of limited importance given the overall problems in high dimensional

density estimation discussed above. The practicability of full bandwidth matrices for $d = 3$ is less clear, and further analysis of this case is an avenue for future research.

Acknowledgments

T.D. acknowledges the financial support of an Australian Postgraduate Award at the University of Western Australia.

The authors thank two anonymous referees for their helpful comments.

References

[1] T. Duong, M.L. Hazelton, Plug-in bandwidth matrices for bivariate kernel density estimation, J. Nonparametric Statist. 1 (2003) 17–30.
[2] T. Duong, M.L. Hazelton, Cross-validation bandwidth matrices for multivariate kernel density estimation (2004) submitted for publication.
[3] M.C. Jones, R.F. Kappenman, On a class of kernel density estimate bandwidth selectors, Scand. J. Statist. Theory Appl. 19 (1992) 337–349.
[4] M.C. Jones, J. Marron, S. Sheather, A brief survey of bandwidth selection for density estimation, J. Amer. Statist. Assoc. 91 (1996) 401–407.
[5] M.C. Jones, S.J. Sheather, Using nonstochastic terms to advantage in kernel-based estimation of integrated squared density derivatives, Statist. Probab. Lett. 11 (1991) 511–514.
[6] J.R. Magnus, H. Neudecker, Matrix Differential Calculus with Applications in Statistics and Econometrics, Wiley, Chichester, 1988.
[7] S.R. Sain, K.A. Baggerly, D.W. Scott, Cross-validation of multivariate densities, J. Amer. Statist. Assoc. 89 (1994) 807–817.
[8] D.W. Scott, Multivariate Density Estimation: Theory, Practice and Visualization, Wiley, New York, 1992.
[9] D. Scott, G. Terrell, Biased and unbiased cross-validation in density estimation, J. Amer. Statist. Assoc. 82 (1987) 1131–1146.
[10] M.P. Wand, Error analysis for general multivariate kernel estimators, J. Nonparametric Statist. 2 (1992) 1–15.
[11] M.P. Wand, M.C. Jones, Comparison of smoothing parameterizations in bivariate kernel density estimation, J. Amer. Statist. Assoc. 88 (1993) 520–528.
[12] M.P. Wand, M.C. Jones, Multivariate plug-in bandwidth selection, Comput. Statist. 9 (1994) 97–116.
[13] M.P. Wand, M.C. Jones, Kernel Smoothing, Chapman & Hall, London, 1995.