Journal of Multivariate Analysis **66**, 207–236 (1998)
Article No. MV981746

Error Process Indexed by Bandwidth Matrices in Multivariate Local Linear Smoothing*

J. A. Cristóbal and J. T. Alcalá

Zaragoza University, *Spain*

Received August 1, 1996; revised November 10, 1997

We focus on nonparametric multivariate regression function estimation by locally weighted least squares. The asymptotic behavior for a sequence of error processes indexed by bandwidth matrices is derived. We discuss feasible data-driven consistent estimators minimizing asymptotic mean squared error or efficient estimators reducing asymptotic bias at points where opposite sign curvatures of the regression function are present in different directions. © 1998 Academic Press

AMS 1991 subject classifications: primary 62G07; secondary 62G20.

Key words and phrases: bandwidth matrix; error process; local linear smoother; multiparameter stochastic process; nonparametric regression; tightness; weak convergence.

1. INTRODUCTION

Nonparametric regression (NPR) has become an almost indispensable tool in exploratory data analysis. In fact, a flexible estimation method is sought, which does not suppose any assumption on the form of the function to describe the association between covariates and response variables. With the nonparametric approach, such a function is determined only by the data. This feature and the current increasing availability of computer power and graphical tools, justify the great interest and popularity of these methods.

The monographs of Eubank [8], Fan and Gijbels [12], Härdle [16], Hastie and Tibshirani [17], Müller [21], Simonoff [30], Wahba [35], and Wand and Jones [37] offer a good introduction and a wide variety of specific applications of NPR to interesting examples with real data.

Let $(X_1, Y_1), ..., (X_n, Y_n)$ be a set of independent random copies of (X, Y), where $X \in \mathbb{R}^d$, $Y \in \mathbb{R}$ are the covariates and the response variables,

* Supported by CICYT Grant PB91-0794.

0047-259X/98 $25.00

having joint density $f_{XY}(.,.)$. We will denote by $f(.)$ the marginal density of X and the regression function by

$$m(\mathbf{x}) = E(Y \mid X = \mathbf{x}) = \int y \, \frac{f_{XY}(\mathbf{x}, y)}{f(\mathbf{x})} \, dy.$$

The most popular estimators of $m(\mathbf{x})$ are the multivariate kernel Nadaraya–Watson estimator (Nadaraya [24] and Watson [38]), and the multivariate Gasser–Müller kernel estimator (Gasser and Müller [15]), although other approaches, like smoothing splines, have been widely used too. In all cases, it is necessary to choose some parameters to determine the amount of smoothing to insert in the estimator. For the practical application, efficient choices of these parameters are needed.

We focus on multivariate local regression estimators (Cleveland [6], Stone [31]), which have desirable properties, such as optimal rates of convergence (Stone [31, 32]) and asymptotic minimax efficiency properties among all possible linear estimators (Fan [9], Fan *et al.* [10]). These procedures have advantages over other popular kernel methods (such as the Nadaraya–Watson and Gasser–Müller methods) because the asymptotic mean squared error (AMSE) is automatically adjusted and the "boundary effect" is hidden; so it does not require modifications at the boundary (Fan and Gijbels [11] showed that the asymptotic bias and variance near the boundary of the support of f have the same order as in the interior points). These benefits are even greater in the multivariate case, where the boundary problem is more severe. Moreover, this procedure has the ability of design adaptation, and it adapts to both fixed and random designs (see Fan and Gijbels [12] for an extended discussion about these points.)

Local regression smoothers estimate $m(\mathbf{x})$ using a weighted least-squares regression with weights based on a kernel function. We will study the problem of linear fit,

$$\text{Min } \varepsilon^T \mathbf{W} \varepsilon, \qquad \text{where} \quad \varepsilon_i = Y_i - \alpha - \beta^T (X_i - \mathbf{x}), \qquad i = 1, ..., n, \tag{1.1}$$

with a matrix $\mathbf{W} = \operatorname{diag}\{K_H(X_1 - \mathbf{x}), ..., K_H(X_n - \mathbf{x})\}$, where $K_H(u) = \det(H)^{-1} K(H^{-1}u)$, K is a d-variate nonnegative kernel function, and the bandwidth matrix H is a $d \times d$ positive definite matrix, but not necessarily symmetric, which depends on n and possibly on $\mathbf{x}$.

The above problem of optimization has a solution for α,

$$\hat{m}(\mathbf{x}) = \hat{\alpha} = \mathbf{e}_1^T (\mathbf{X}^T \mathbf{W} \mathbf{X})^{-1} \mathbf{X}^T \mathbf{W} \mathbf{Y}, \tag{1.2}$$

where $\mathbf{e}_1$ is the projection vector on the first component, the ith row of $\mathbf{X}$ is $(1, (X_i - \mathbf{x})^T)$, and $\mathbf{Y} = (Y_1, ..., Y_n)^T$.

The analysis with a matrix H_n of bandwidth leads not only to the introduction of a different amount of smoothing, but, moreover, it permits us to carry out rotations such that the weights determined by the kernel have ellipsoidal contours with a certain orientation (their axes are not necessarily in the same direction as the coordinate axes). The advantages of getting a full matrix of smoothing parameters have been emphasized in some contexts by Wand and Jones [36] and Ruppert and Wand [26]. In the former, a general exposition concerning the good behavior of this kind of estimators is given.

The main purpose is to show the asymptotic behavior of a sequence of processes with multivariate time, which are related to the choice of local smoothing parameters. As consequence, the asymptotic distribution of estimators with local bandwidth matrix constructed from the data are studied.

The proposed method is applicable whenever a bandwidth matrix minimizing the asymptotic mean squared error (AMSE) exists, although such a matrix might not exist, as we will discuss later. At points where opposite sign curvatures in different directions (not necessarily the axes directions) are present, we must change our attention to bandwidths which asymptotically cause the dominating bias to vanish due to the "curvature" of the regression function.

Results of weak convergence of bandwidth processes to a Gaussian limit process have been introduced in the literature with the aim of showing that there exist data-adaptive bandwidth choices in different contexts which are asymptotically efficient (see Abramson [1] and Krieger and Pickands [18] for the case of univariate density estimation, Müller and Stadtmüller [23] for the estimation of a regression function in a univariate fixed design, Mack and Müller [20] for the multivariate Nadaraya–Watson estimator with only one bandwidth parameter, Müller and Prewitt [22] for the multivariate convolution kernel estimator with a vector bandwidth in a fixed design). In this work, in the context of multivariate local regression smoothers, a functional limit theorem for matrix bandwidth processes is obtained, and its tightness is shown.

The above expression (1.2) for $\hat{m}(\mathbf{x})$ as a product of matrices, prevents a development of the error process, similar to that of Mack and Müller [20], based on a linearization of the ratio estimator. Therefore, we have found alternative expressions of such an estimator as a ratio of determinants and it allows us to split the error process into convergent processes.

This paper is organized as follows: In Section 2, we present specific representation and preliminary results, and then we discuss the optimal bandwidth matrix choice. Section 3 is devoted to the derivation of the weak convergence of matrix bandwidth processes. Auxiliary results and proofs are given in Section 4.

2. BANDWIDTH MATRIX CHOICE IN LOCAL LINEAR ESTIMATION

The expression (1.2) is adequate when analyzing the asymptotic conditional bias and variance of the estimator (Ruppert and Wand [26]), but other characteristics are awkwardly obtained from it.

In the next theorem, we give an alternative formulation to (1.2), which presents the estimator as a ratio of two statistics similar to the Nadaraya–Watson estimator.

Let $\mathbf{X}^{(i)}(u)$ be a matrix having vector u in the ith column and all others the same as $\mathbf{X}$. Let $\mathbf{j}$ be a subset of size $(d+1)$ taken from $\{1, ..., n\}$, and let $\mathbf{X_j}$ be the respective submatrix of $\mathbf{X}$ with row indices given by $\mathbf{j}$. By using the Binet–Cauchy expansion (Noble [25]) for the determinant of the product of matrices we obtain:

THEOREM 2.1. *Suppose that $\mathbf{X}_\mathbf{j}^T\mathbf{W}_\mathbf{j}\mathbf{X}_\mathbf{j}$ is not singular for a $\mathbf{j}$. Then*

$$\hat{\alpha} = \hat{m}_n(\mathbf{x}) = \frac{\sum_\mathbf{j} \det(\mathbf{X}_\mathbf{j}^T\mathbf{W}_\mathbf{j}\mathbf{X}_\mathbf{j}^{(1)}(\mathbf{Y}_\mathbf{j}))}{\sum_\mathbf{j} \det(\mathbf{X}_\mathbf{j}^T\mathbf{W}_\mathbf{j}\mathbf{X}_\mathbf{j})}, \tag{2.1}$$

where $\sum_\mathbf{j}$ means summation over all subsets of size $(d+1)$.

From (2.1) we can see $\hat{m}_n(\mathbf{x})$ as a weighted average of locally least squares estimator obtained only with $(d+1)$ observations. Let $\hat{m}_\mathbf{j}(\mathbf{x})$ be one of these. Then, it is straightforward from (2.1) that

$$\hat{\alpha} = \hat{m}_n(\mathbf{x}) = \sum_\mathbf{j} w_\mathbf{j}(\mathbf{x})\, \hat{m}_\mathbf{j}(\mathbf{x}),$$

with

$$w_\mathbf{j}(\mathbf{x}) = \frac{\det(\mathbf{X}_\mathbf{j}^T\mathbf{W}_\mathbf{j}\mathbf{X}_\mathbf{j})}{\det(\mathbf{X}^T\mathbf{W}\mathbf{X})}$$

and $\hat{m}_\mathbf{j}(\mathbf{x}) = 0$ if $\det(\mathbf{X}_\mathbf{j}^T\mathbf{W}_\mathbf{j}\mathbf{X}_\mathbf{j}) = 0$.

Estimator (2.1) is linear in the observations $Y_1, ..., Y_n$ and the following expression is useful too. It follows from elementary properties of expansion of a determinant by rows and from $\mathbf{Y} = \sum_i Y_i\mathbf{e}_i$, where $\mathbf{e}_i$ is a n-vector with a 1 in its i coordinate and 0's elsewhere:

$$\hat{m}_n(\mathbf{x}) = \frac{\det(\mathbf{X}^T\mathbf{W}\mathbf{X}^{(1)}(\mathbf{Y}))}{\det(\mathbf{X}^T\mathbf{W}\mathbf{X})} = \sum_{i=1}^{n} \frac{Y_i \det(\mathbf{X}^T\mathbf{W}\mathbf{X}^{(1)}(\mathbf{e}_i))}{\det(\mathbf{X}^T\mathbf{W}\mathbf{X})} = \sum_{i=1}^{n} Y_i w_{in}(\mathbf{x}). \tag{2.2}$$

It is easy to see that

$$\sum_{i=1}^{n} w_{in}(\mathbf{x}) = 1$$

and

$$\sum_{i=1}^{n} \mathbf{a}^T (X_i - \mathbf{x})\, w_{in}(\mathbf{x}) = 0 \qquad \forall \mathbf{a} \in \mathbb{R}^d.$$

Both properties show that $w_{in}(\mathbf{x})$ is a sequence of conditional second-order weights, and involve estimators which are unbiased for linear functions $m(\mathbf{x})$ (see Ruppert and Wand [26]). This is desirable in estimating at points near the boundary of support of f and points with a very asymmetric design (see Fan and Gijbels [12]).

The estimation of regression function derivatives is also possible with this representation, if we want to estimate $m_{(j)}(\mathbf{x}) = \partial m(\mathbf{x})/\partial x_j$, $j = 1, ..., d$, then

$$\hat{m}_{(j)}(\mathbf{x}) = \frac{\det(\mathbf{X}^T \mathbf{W} \mathbf{X}^{(j+1)}(\mathbf{Y}))}{\det(\mathbf{X}^T \mathbf{W} \mathbf{X})}, \qquad j = 1, ..., d.$$

The main arguments could be adapted to cover similar results for these first derivative estimators.

Next, we give the following set of assumptions. Some results in Section 2 could be obtained with weaker conditions; however, they will be necessary in order to derive the main theorem in Section 3.

Concerning the kernel,

(W.1) $K(.)$ is a bounded, nonnegative, compactly supported kernel.

(W.2) $K(.)$ is a spherically symmetric kernel or a product of symmetric univariate kernels (in both situations, all odd order moments of K vanish).

(W.3) The first partial derivatives of $K(.)$ exist and are bounded.

Some formulas involving auxiliary results are simplified if we assume that K is a multivariate probability density function with mean equal to zero and covariance matrix $\mu_2 I_d$, with I_d the $d \times d$ identity matrix, and $\mu_2 = \int u_1^2 K(u)\, du$. Further, we denote $\mu_4 = \int u_i^4 K(u)\, du$, $i = 1, ..., d$, and $\mu_{22} = \int u_i^2 u_j^2 K(u)\, du$, $i, j = 1, ..., d$, and $i \neq j$.

Concerning the density function:

(D.1) f is bounded and continuous at $\mathbf{x}$, and $f(\mathbf{x}) > 0$.

(D.2) All mixed partial derivatives of f at $\mathbf{x}$ exist up to second order, with the second partial derivatives continuous at $\mathbf{x}$.

Concerning the regression function and other related functions:

(R.1) The function $g(.)=f(.)\,m(.)=\int y\, f_{XY}(., y)\, dy$ exists and is continuous at $\mathbf{x}$.

(R.2) $s^2(.)\, f(.)=\int y^2 f_{XY}(., y)\, dy$ exists and is continuous at $\mathbf{x}$.

(R.3) All mixed partial derivatives of $m(.)$ and $s^2(.)$ at $\mathbf{x}$ exist up to the second order, with the second partials continuous at $\mathbf{x}$.

(R.4) The functions $\alpha_k(.)\, f(.)=\int |y|^k f_{XY}(., y)\, dy$ exist and are continuous at $\mathbf{x}$ $(k\geqslant 1)$.

Concerning the bandwidth matrix: The matrix sequence H_n is such that $(n \det H_n)^{-1}$ and each entry of H_n tend to zero as $n\to\infty$, with H_n nonsingular. Moreover,

(H.1) $H_n=h_n A$, with $h_n\to 0$ and $nh_n^d\to\infty$, as $n\to\infty$ and matrix A is such that $\det A=1$ for all n. (This condition allows us separate size and shape in H_n.)

(H.2) $H_n=n^{-1/(d+4)}H$, with H a nonsingular definite positive matrix.

Let $R(K)$ be $\int K^2(u)\, du$, $\sigma^2(\mathbf{x})=\mathrm{Var}(Y\mid X=\mathbf{x})>0$, and $\mathscr{H}_m(\mathbf{x})$ be the Hessian matrix of mixed second partials of $m(\mathbf{x})$. Using (2.2), expressions of conditional bias and variance of $\hat{m}_n(\mathbf{x})$ are derived and they are equal to those obtained by Ruppert and Wand [26]. Let $\mathbf{x}$ be a fixed element in the interior of the support of f. Assume that (W.1)–(W.2), (D.1)–(D.2), (R.1)–(R.3), and (H.1) hold. Then

$$\mathrm{E}\{\hat{m}_n(\mathbf{x})-m(\mathbf{x})\mid X_1, ..., X_n\}=\tfrac{1}{2}\,\mathrm{trace}\{H_n^T\mathscr{H}_m(\mathbf{x})\,H_n\}+\mathrm{o}_p(\mathrm{trace}(H_n^T H_n)) \tag{2.3}$$

and

$$\begin{aligned}&\mathrm{Var}\{\hat{m}_n(\mathbf{x})-m(\mathbf{x})\mid X_1, ..., X_n\}\\ &\quad=n^{-1}(\det H_n)^{-1}\,\{R(K)/f(\mathbf{x})\}\,\sigma^2(\mathbf{x})\{1+\mathrm{o}_p(1)\}.\end{aligned} \tag{2.4}$$

Both leading terms are combined to give the asymptotic conditional mean-squared error (AMSE):

$$\begin{aligned}&\mathrm{AMSE}(\hat{m}_n(\mathbf{x})\mid X_1, ..., X_n)\\ &\quad=n^{-1}(\det H_n)^{-1}\,\{R(K)/f(\mathbf{x})\}\,\sigma^2(\mathbf{x})+\tfrac{1}{4}\,\mathrm{trace}^2\,\{H_n^T\mathscr{H}_m(\mathbf{x})\,H_n\},\end{aligned} \tag{2.5}$$

and this value does not depend on the sample $X_1, ..., X_n$.

The choice of the bandwidth matrix is crucial for the behavior of the estimator. This matrix can be the same at every point when we take the estimation or otherwise we can adapt it to each point.

There are many ways of constructing adaptive estimates of the function $m(\mathbf{x})$. The most interesting consists of taking the matrix $H_n = H_n(\mathbf{x})$, a function of the point $\mathbf{x}$ where the regression function is to be estimated. Other possibilities can be to take a different matrix $H_n(X_i)$ for each point in the sample or to take a variable matrix H_n such that there are at least k sample observations falling into a neighborhood of $\mathbf{x}$.

On analyzing the AMSE, new situations can be seen because of the multivariate character of the problem. The leading bias term

$$B_n(\mathbf{x}) = \tfrac{1}{2}\,\mathrm{trace}(H_n^T \mathscr{H}_m(\mathbf{x})\, H_n) \tag{2.6}$$

has a geometric interpretation. It is a combination of some entries of $\mathscr{H}_m(\mathbf{x})$, i.e., of the curvature of $m(\mathbf{x})$ at $\mathbf{x}$ in the different directions of space. In the univariate setting, the bias is controlled entirely by the only bandwidth parameter, while in the multivariate setting we can sometimes compensate some curvatures with others, and so the leading term of the asymptotic bias can be cancelled.

This characteristic has been commented on in the context of density estimation by Terrell and Scott [34]. Müller and Prewitt [22] analyze its use in the regression estimation for the particular case of variable smoothing in directions parallel to coordinate axes.

In order to minimize the asymptotic bias, we will consider three different situations depending on the eigenvalues of the matrix $\mathscr{H}_m(\mathbf{x})$:

- *Case* I. $\mathscr{H}_m(\mathbf{x})$ is positive or negative definite. Then, there is not a matrix H_n such that (2.6) vanishes. The solution which minimizes (2.5) is given in the next lemma.

LEMMA 2.1. *Let* $\mathbf{x}$ *be a point in the interior of the support of* $f(\,.\,)$, *such that* $\mathscr{H}_m(\mathbf{x})$ *is positive or negative definite. Assume that* (W.1)–(W.2), (D.1)–(D.2), (R.1)–(R.3), *and* (H.1) *hold. Then, the matrix* H_n *minimizing AMSE in* (2.5) *is*

$$H_n(\mathbf{x}) = n^{-1/d+4} \left\{\frac{\sigma^2(\mathbf{x})\, R(K)}{df(\mathbf{x})}\right\}^{1/d+4} (\det \mathscr{H}_m(\mathbf{x}))^{1/2(4+d)}\, O(\mathbf{x})\, \Gamma^+(\mathbf{x})^{-1/2}, \tag{2.7}$$

where $\Gamma^+(\mathbf{x})$ *is a diagonal matrix of eigenvalues* (*in absolute value*) *and* $O(\mathbf{x})$ *is the matrix of eigenvectors of* $\mathscr{H}_m(\mathbf{x})$. *The optimal value of AMSE is*

$$\mathrm{AMSE}^*(\mathbf{x}) = n^{-4/(d+4)} (\det \mathscr{H}_m(\mathbf{x}))^{2/(4+d)} \times \left\{\frac{\sigma^2(\mathbf{x})\, R(K)}{f(\mathbf{x})}\right\}^{4/d+4} \left\{\frac{d}{4}+1\right\} d^{d/d+4}.$$

- *Case* II. $\mathscr{H}_m(\mathbf{x})$ has both positive and negative eigenvalues. Then, the regression function has positive curvatures in some directions and negative in others; i.e., $m(\,.\,)$ is saddle-shaped at $\mathbf{x}$. We can choose a matrix H_n such that (2.6) vanishes.

LEMMA 2.2. *Assume that $\mathscr{H}_m(\mathbf{x})$ has both positive and negative eigenvalues and the other conditions of Lemma* 2.1 *hold. It is possible to construct a matrix H_n so that the asymptotic conditional bias is $\mathrm{o}_P(n^{-2/(d+4)})$. One such matrix is*

$$H_n(\mathbf{x}) = n^{-1/d+4} \left\{ \frac{\sigma^2(\mathbf{x})\, R(K)}{d f(\mathbf{x})} \right\}^{1/d+4} (\det \mathscr{H}_m(\mathbf{x}))^{1/2(4+d)}\, O(\mathbf{x})\, \Gamma^*(\mathbf{x})^{1/2}, \tag{2.8}$$

where $\Gamma^*(\mathbf{x})$ is a full rank diagonal matrix with trace $\Gamma^*(\mathbf{x})\,\Gamma(\mathbf{x}) = 0$. *Here, $\Gamma(\mathbf{x})$ is the diagonal matrix of eigenvalues of $\mathscr{H}_m(\mathbf{x})$.*

We can see the action of H_n first as a rotation of coordinate axes around $\mathbf{x}$, so that the new axes agree with the directions of curvatures of a different sign, and then, a suitable scaling, so that the new parametrization of the regression function verifies a Laplacian equation.

- *Case* III. $\mathscr{H}_m(\mathbf{x})$ is semidefinite with at least one zero eigenvalue. Then, $\mathscr{H}_m(\mathbf{x})$ is nonfull rank. The asymptotic bias corresponds with one of a problem with a smaller dimension than d, and the contribution of these points to the bias is of a lower order than in Case I.

In density estimation the situation is similar to this, but it is less interesting because usually one needs a more accurate density estimation near the modes, i.e., in points falling in Case I. In these areas, the adapted smoothing does not improve the order of convergence of MSE in relation to the use of an optimal matrix H_n in a global sense. However, saddle points of the regression function are generally quite interesting for estimating $m(\mathbf{x})$.

The Nadaraya–Watson estimator allows us to insert a matrix H_n in a similar analysis, too. But the expressions of the asymptotic bias make the discussion of the feasible cases complicated and confusing.

We must note that the matrices proposed in (2.7) and (2.8) could be right-handedly multiplied by an orthogonal matrix G without a change in the AMSE stated by Lemmas 2.1 and 2.2. Then, we can simply consider a symmetric bandwidth matrix (e.g., we take $G = O(\mathbf{x})^T$), and this choice may reduce the number of parameters to be estimated in practice. The rotation produced by G means a new change in the coordinate system in order to present the estimation, so it usually seems more convenient to return to the original coordinate system with $G = O(\mathbf{x})^T$.

Before finishing this section, we give a result for the asymptotic distribution of error $\hat{m}_n(\mathbf{x}) - m(\mathbf{x})$. Let the following functions be:

$$B(\mathbf{x}) = \frac{1}{2} \operatorname{trace}\{H^T \mathcal{H}_m(\mathbf{x})\, H\},$$

$$S(\mathbf{x}) = R(K) \frac{\sigma^2(\mathbf{x})}{(\det H)\, f(\mathbf{x})}.$$

THEOREM 2.2. *Let* $\mathbf{x}$ *be a point under conditions of Lemma* 2.1 *and suppose that* $H_n \xrightarrow{n} n^{-1/d+4} H$, *with* H *a full rank and definite positive matrix. Then,*

$$\sup_{z \in \mathbb{R}} \left| P(n^{2/(d+4)}(\hat{m}_n(\mathbf{x}) - m(\mathbf{x})) \leqslant z) - \Phi\left(\frac{z - B(\mathbf{x})}{S^{1/2}(\mathbf{x})}\right)\right| \xrightarrow[n]{} 0,$$

where $\Phi(\,.\,)$ *is the standard normal distribution function. So*

$$n^{2/(d+4)}(\hat{m}_n(\mathbf{x}) - m(\mathbf{x})) \xrightarrow{\mathcal{L}} N(B(\mathbf{x}), S(\mathbf{x})).$$

This result is a consequence of a more general result established by Battacharya and Müller [2] for asymptotic behavior of functionals of data averages.

3. CONVERGENCE OF THE ERROR PROCESS

For the study of convergence of the error process we will only consider bandwidth matrices of the form $H_n = n^{-1/(d+4)} H$, according to (H.2). Matrix H belongs to $\mathfrak{H}$, a compact of $Gl(\mathbb{R}, d)$, the space of $d \times d$ regular matrices. There is a bijection between such a space and an open set on $\mathbb{R}^{d^2}$ (in relation to the corresponding vector space topology), through the vectorization operator vec(.), and so we can see $\mathfrak{H}$ as a compact on $\mathbb{R}^{d^2}$. The space of matrices is endowed with the Frobenius norm ($\|H\|^2 = \operatorname{trace}(H^T H) = \operatorname{vec}(H)^T \operatorname{vec}(H)$) which becomes the Euclidean norm on the corresponding vectorizations.

The numerator and the denominator in (2.1) are symmetric statistics in the observations, and they are associated with some functionals of the density and regression functions, as is later shown. We will often use the property of symmetry in the observations to simplify some proofs in a similar way to what happens with U-statistics (Serfling [29]). In fact, they are U-statistics with a varying kernel depending on n. If we denote

$$U_n(\mathbf{x}; H_n) \equiv \left(\binom{n}{d+1} (d+1)! \, (\det H_n)^2 \right)^{-1} \sum_{\mathbf{j}} (\det(\mathbf{X}_{\mathbf{j}}^T \mathbf{X}_{\mathbf{j}}^{(1)}(\mathbf{Y}_{\mathbf{j}})) \det \mathbf{W}_{\mathbf{j}}),$$

$$V_n(\mathbf{x}; H_n) \equiv \left(\binom{n}{d+1} (d+1)! \, (\det H_n)^2 \right)^{-1} \sum_{\mathbf{j}} (\det(\mathbf{X}_{\mathbf{j}}^T \mathbf{X}_{\mathbf{j}}) \det \mathbf{W}_{\mathbf{j}}),$$

the estimator (2.1) becomes

$$\hat{m}_n(\mathbf{x}) = \frac{U_n(\mathbf{x}; H_n)}{V_n(\mathbf{x}; H_n)}.$$

Until further notice, to abbreviate, we will write $\hat{m}_n(H)$, $U_n(H)$, $V_n(H)$, instead of $\hat{m}_n(\mathbf{x}; H_n)$, $U_n(\mathbf{x}; H_n)$, $V_n(\mathbf{x}; H_n)$ and A, B will denote matrices in $\mathfrak{H}$.

The error process is

$$R_n(H) = n^{2/(d+4)}(\hat{m}_n(H) - m(\mathbf{x})), \qquad H \in \mathfrak{H}. \tag{3.1}$$

The main theorem is the weak convergence of this stochastic process indexed by the bandwidth matrix.

THEOREM 3.1. *Assume* (W.1)–(W.3), (D.1)–(D.2), *and* (R.1)–(R.4). *Then*

$$R_n(H) \Rightarrow R(H),$$

where $R(.)$ *is a multivariate Gaussian process with multivariate index, which is characterized by*

$$\mathrm{E}(R(H)) = \tfrac{1}{2} \operatorname{trace}\{H^T \mathcal{H}_m(\mathbf{x}) \, H\}$$

and

$$\operatorname{Cov}(R(A), R(B)) = \frac{\sigma^2(x)}{f(\mathbf{x}) \det A \det B} \int K(A^{-1}u) \, K(B^{-1}u) \, du.$$

These results are of greater interest when we want to work with a bandwidth matrix $H^* \in \mathfrak{H}$, optimal as (2.7) or (2.8), since H^* depends upon unknown parameters as $\mathcal{H}_m(\mathbf{x})$, $\sigma^2(\mathbf{x})$, or $f(\mathbf{x})$, and it cannot be used However, we can get consistent estimates of these parameters and construct a matrix $\hat{H}^*(\mathbf{x})$ (e.g., by plug-in methods) such that $\hat{m}_n(\mathbf{x}; \hat{H}_n^*)$ is as efficient as $\hat{m}_n(\mathbf{x}; H_n^*)$; that is,

$$n^{2/(d+4)}[\hat{m}_n(\mathbf{x}; \hat{H}_n^*) - \hat{m}_n(\mathbf{x}; H_n^*)] \xrightarrow{P} 0.$$

Similar results using weak convergence are given in Krieger and Pickards [18] and Abramson [1] in estimating a density and in Mack and Müller [20] and Müller and Prewitt [22] in estimating a regression function.

If **x** belongs to Case I and with the various hypotheses carried out on $f(.)$, $m(.)$, and the kernel function, it holds that the optimum rate for the AMSE is of $n^{-4/(d+4)}$, and this rate can be reached by using matrices of the $H_n = n^{-1/(d+4)}H$ form. So Lemma 2.2 indicates how H should be taken so that the AMSE constant can be mimimized.

If **x** is in Case II and we do not add stronger conditions on $m(.)$ and $f(.)$, there will exist no H that mimimizes this AMSE constant, but it still makes sense to consider matrices of the $H_n = n^{-1/(d+4)}H$ form that verify $B(\mathbf{x}) = 0$.

If we demand that the bandwidth matrix should be symmetric, there exists only one matrix that will mimimize the ANISE constant in points **x** of Case I. However, there are multiple choices of H which cancel the dominant term of asymptotic bias for points x of Case II.

The tightness of the process given in (3.1) and the existence of consistent estimators of the unknown values allows us to obtain efficient estimators of $m_n(\mathbf{x}; n^{-1/(d+4)}H^*)$. As corollary of the main result, we have

Corollary 3.1. *Let H^* be in $\mathfrak{H}$ and suppose that a sequence $\{\hat{H}_n^*\}_n \subset \mathfrak{H}$ of data-driven bandwidth matrices exists such that $\hat{H}_n^* \xrightarrow{P} H^*$, as $n \to \infty$, then*

$$n^{2/(d+4)}(\hat{m}_n(\mathbf{x}; n^{-1/(d+4)}\hat{H}_n^*) - m(\mathbf{x}; n^{-1/(d+4)}H^*)) \xrightarrow{\mathscr{L}} \mathscr{N}(B^*(\mathbf{x}), S^*(\mathbf{x}))$$

with

$$B^*(\mathbf{x}) = \frac{1}{2}\,\text{trace}\{H^{*T}\mathscr{H}_m(\mathbf{x})\,H^*\},$$

$$S^*(\mathbf{x}) = R(K)\,\frac{\sigma^2(\mathbf{x})}{(\det H^*)\,f(\mathbf{x})}.$$

Consistent estimators for $\mathscr{H}_m(\mathbf{x})$ are obtained using a local quadratic or cubic smoother (see Fan and Gijbels [12] or Ruppert and Wand [26]). For density estimation in a multivariate context see Scott [27] and for consistent variance estimation we can use estimators similar to those proposed by Müller and Prewitt [22] or an estimator based on a normalized weighted residual sum of squares (see Fan and Gijbels [12] and Fan and Yao [13]).

We start the analysis by decomposing the error process,

$$R_n(H) = \frac{1}{f^{d+1}(\mathbf{x})}\,[\psi_n(H) + Y_n(H)] - \frac{\hat{m}_n(H)}{f^{d+1}(\mathbf{x})}\,[\zeta_n(H) + Z_n(H)],$$

where

$$\begin{aligned}\psi_n(H) &= n^{2/(d+4)}[\mathrm{E}U_n(H) - m(\mathbf{x})\, f^{d+1}(\mathbf{x})],\\ \zeta_n(H) &= n^{2/(d+4)}[\mathrm{E}V_n(H) - f^{d+1}(\mathbf{x})],\\ Y_n(H) &= n^{2/(d+4)}[U_n(H) - \mathrm{E}U_n(H)],\\ Z_n(H) &= n^{2/(d+4)}[V_n(H) - \mathrm{E}V_n(H)].\end{aligned}$$

We will obtain first, the Gaussian limit process of $Y_n(H)$, $Z_n(H)$, and later, the Gaussian limit process of $R_n(H)$. The mode of convergence is the weak convergence in $\mathscr{C}(\mathfrak{H})$, the space of continuous functions on $\mathfrak{H}$, with the supremum norm.

Uniform Convergence of Nonstochastic Elements

LEMMA 3.1. *Assume that* (W.1)–(W.3) *hold.*

(i) *If* (D.2) *is satisfied, then*

$$\sup_{H \in \mathfrak{H}} |\zeta_n(H) - \zeta(H)| \xrightarrow{n} 0, \qquad as \quad n \to \infty,$$

with

$$\begin{aligned}\zeta(H) = {} & \tfrac{1}{2} f(\mathbf{x})^d\, (1 + \mu_4 + (d-1)\, \mu_{22}) \operatorname{trace}(H^T \mathscr{H}_f(\mathbf{x})\, H)\\ & - f(\mathbf{x})^{d-1} \operatorname{trace}(H^T D_f(\mathbf{x})\, D_f^T(\mathbf{x})\, H),\end{aligned}$$

where $D_f(\mathbf{x})$ is the vector of first-order partial derivatives of f at point $\mathbf{x}$.

(ii) *Moreover, if* (R.3) *is satisfied, then*

$$\sup_{H \in \mathfrak{H}} |\psi_n(H) - \psi(H)| \xrightarrow{n} 0, \qquad as \quad n \to \infty,$$

with

$$\begin{aligned}\psi(H) = {} & \tfrac{1}{2} f^{d+1}(\mathbf{x}) \operatorname{trace}(H^T \mathscr{H}_m(\mathbf{x})\, H)\\ & + m(\mathbf{x})\, f^d(\mathbf{x}) \{\tfrac{1}{2}(1 + \mu_4 + (d-1)\, \mu_{22})\} \operatorname{trace}(H^T \mathscr{H}_f(\mathbf{x})\, H)\\ & - m(\mathbf{x})\, f(\mathbf{x})^{d-1} \operatorname{trace}(H^T D_f(\mathbf{x})\, D_f^T(\mathbf{x})\, H).\end{aligned}$$

Note. It is straightforward that

$$\psi(H) - m(\mathbf{x})\, \zeta(H) = \tfrac{1}{2} f^{d+1}(\mathbf{x}) \operatorname{trace}\{H^T \mathscr{H}_m(\mathbf{x})\, H\}.$$

Asymptotic Normality of $Y_n(H_0)$ and $Z_n(H_0)$ at a Single Matrix $H_0 \in \mathfrak{H}$

The functional weak convergence of these processes is shown if we prove the convergence of the finite-dimensional distributions and the tightness of the processes. First, we will establish auxiliary results to obtain these finite-dimensional distributions. Then, we will check the tightness by means of results derived from Bickel and Wichura [3].

Let us define the following constants depending upon the kernel,

$$d_{1K}=\int K^2(u)\,du,$$

$$d_{2K}=\sum_{i=1}^{d}\int u_i^2 K^2(u)\,du+\sum_{i,\,j=1}^{d}\int u_i^2 u_j^2 K^2(u)\,du,$$

$$d_K=d_{1K}+d_{2K}.$$

Lemma 3.2. *Assume that* (W.1)–(W.2) *and* (D.1)–(D.2) *hold. Then*

(i) $$Z_n(H_0)\xrightarrow{\mathscr{L}}\mathscr{N}\left(0,\frac{f^{2d+1}(\mathbf{x})}{\det H_0}\,d_K\right).$$

(ii) *Moreover, if* (R.1)–(R.4) *are satisfied, then*

$$Y_n(H_0)\xrightarrow{\mathscr{L}}\mathscr{N}\left(0,\frac{f^{2d+1}(\mathbf{x})}{\det H_0}\,[s^2(\mathbf{x})\,d_{1K}+m^2(\mathbf{x})\,d_{2K}]\right).$$

Note. (i) The asymptotic covariance between $Z_n(H_0)$ and $Y_n(H_0)$ is

$$\operatorname{Cov}(Z_n(H_0),\,Y_n(H_0))\xrightarrow{n}\frac{f^{2d+1}(\mathbf{x})\,m(\mathbf{x})}{\det H_0}\,d_K,\qquad \text{as}\quad n\to\infty.$$

(ii) If $\hat{m}_n(H_0)\xrightarrow{P}m(\mathbf{x})$, it is straightforward that

$$R_n(H_0)\xrightarrow{\mathscr{L}}\mathscr{N}\left(\frac{1}{2}\operatorname{trace}(H_0^T\mathscr{H}_m(\mathbf{x})\,H_0);\frac{\sigma^2(\mathbf{x})}{f(\mathbf{x})\det H_0}\,d_{1K}\right),$$

with $\sigma^2(\mathbf{x})=s^2(\mathbf{x})-m^2(\mathbf{x})=\operatorname{Var}(Y\mid X=\mathbf{x})$.

Covariances of the Auxiliary Processes

To avoid a more complicated notation, we turn any d-vector into other $(d+1)$-vector by adding a 0-component equal to 1.

Lemma 3.3. *Assume that* (W.1)–(W.3) *and* (D.1) *hold. For* $A, B \in \mathfrak{H}$,

$$\operatorname{Cov}(Z_n(A), Z_n(B)) = \frac{f^{2d+1}(\mathbf{x})}{\det A \det B} \sum_{i,\, j=1}^{d+1} \int (A^{-1}u)_{i-1}^2 \, (B^{-1}u)_{j-1}^2 \times K(A^{-1}u)\, K(B^{-1}u)\, du + \mathrm{o}(1).$$

Lemma 3.4. *Assume that* (W.1)–(W.3), (D.1), (R.1), *and* (R.2) *hold. For* $A, B \in \mathfrak{H}$,

$$\operatorname{Cov}(Y_n(A), Y_n(B)) = \frac{f^{2d+1}(\mathbf{x})}{\det A \det B} \sum_{p,\, q=1}^{d+1} c_{pq}(\mathbf{x}) \times \int (A^{-1}u)_{p-1}^2 \, (B^{-1}u)_{q-1}^2 \, K(A^{-1}u)\, K(B^{-1}u)\, du + \mathrm{o}(1),$$

with

$$c_{pq}(\mathbf{x}) = \begin{cases} s^2(\mathbf{x}), & \text{if } \ p = q = 1, \\ m(\mathbf{x})^2, & \text{otherwise.} \end{cases}$$

Lemma 3.5. *Assume that* (W.1)–(W.3), (D.1), *and* (R.1) *hold. For* $A, B \in \mathfrak{H}$,

$$\operatorname{Cov}(Y_n(A), Z_n(B)) = \frac{f^{2d+1}(\mathbf{x})\, m(\mathbf{x})}{\det A \det B} \sum_{i,\, j=1}^{d+1} \int (A^{-1}u)_{i-1}^2 \, (B^{-1}u)_{j-1}^2 \times K(A^{-1}u)\, K(B^{-1}u)\, du + \mathrm{o}(1).$$

Weak Convergence of the Auxiliary Processes

In order to prove the weak convergence of these processes, we will verify their tightness. For this purpose we need to establish bounds on the moments at the blocks (see Bickel and Wichura [3] for a definition of block).

Let us denote with $H_{i,\, j}(t) \in \mathfrak{H}$, a matrix having t in the (i, j) entry. Assume that s, t are real numbers such that $H_{i,\, j}(t + \lambda(s - t)) \in \mathfrak{H}\ \forall \lambda \in [0, 1]$. Different constants are denoted by $C_0, C_1, \ldots$. By using the projection of the statistics and applying the mean value theorem for vectorial functions, we have:

Lemma 3.6. *Assume that* (W.1)–(W.3), (D.1), (R.1), *and* (R.4) *hold. Then*:

(i) $\mathrm{E}(|Y_n(H_{i,\, j}(t)) - Y_n(H_{i,\, j}(s))|^p) \leqslant C_0\, |s - t|^p$,

(ii) $\mathrm{E}(|Z_n(H_{i,\, j}(t)) - Z_n(H_{i,\, j}(s))|^p) \leqslant C_1\, |s - t|^p$.

This lemma is crucial for proving the condition on the moments from which we will derive the tightness.

Let $\mathbf{B}=\prod_{i=1}^{d}\prod_{j=1}^{d}(s_{ij}, t_{ij}]$ and $\mathbf{D}=\prod_{i=1}^{d}\prod_{j=1}^{d}(u_{ij}, v_{ij}]$ *be two blocks in* $Gl(\mathbb{R}, d)$ *with* $\mu(\mathbf{B})>0$, $\mu(\mathbf{D})>0$. *The increment of a process* ϕ_n *(expressing* Y_n *or* Z_n*) around* **B** *is defined as in Bickel and Wichura* (1971):

$$\phi_n(\mathbf{B})=\sum_{i,j}\sum_{\delta_{i,j}=0,1}(-1)^{d^2-\Sigma_{i,j}\delta_{ij}} \times\phi_n(s_{11}+\delta_{11}(t_{11}-s_{11}), ..., s_{dd}+\delta_{dd}(t_{dd}-s_{dd})).$$

LEMMA 3.7. *Assume that* (W.1)–(W.3), (D.1), (R.1), *and* (R.2) *hold. Then*:

$$\mathrm{E}(|\phi_n(\mathbf{B})|^{d^2}\,|\phi_n(\mathbf{D})|^{d^2})\leqslant C_2\mu(\mathbf{B})\,\mu(\mathbf{D})$$

for a constant $C_2>0$ *which does not depend on* **B**, **D** *(*ϕ_n *denote either* Y_n *or* Z_n*).*

From the convergence of finite-dimensional distributions and the condition on the moments, we have

LEMMA 3.8. *Assume that* (W.1)–(W.3), (D.1), *and* (R.1)–(R.2) *hold. Then, the sequences* $\{Y_n(H), H\in\mathfrak{H}\}$ *and* $\{Z_n(H), H\in\mathfrak{H}\}$ *of random elements of* $\mathscr{C}(\mathfrak{H})$ *are tight.*

LEMMA 3.9. *Assume that* (W.1)–(W.3), (D.1)–(D.2) *hold. Then, we have*:

(i) $$Z_n(H)\Rightarrow Z(H),$$

where $Z(.)$ *is a multivariate Gaussian process with multivariate index, mean zero and covariance*:

$$\mathrm{Cov}(Z(A), Z(B)) =\frac{f^{2d+1}(\mathbf{x})}{\det A\det B}\sum_{i,j=1}^{d+1}\int(A^{-1}u)_{i-1}^2\,(B^{-1}u)_{j-1}^2\,K(A^{-1}u)\,K(B^{-1}u)\,du.$$

(ii) *Moreover, if* (R.1)–(R.4) *are satisfied, we have*:

$$Y_n(H)\Rightarrow Y(H),$$

where $Y(.)$ *is a multivariate Gaussian process with multivariate time, mean zero and covariance*:

$$\mathrm{Cov}(Y(A), Y(B)) =\frac{f^{2d+1}(\mathbf{x})}{\det A\det B}\sum_{p,q=1}^{d+1}c_{pq}(\mathbf{x})\int(A^{-1}u)_{p-1}^2\,(B^{-1}u)_{q-1}^2\,K(A^{-1}u)\,K(B^{-1}u)\,du,$$

with

$$c_{pq}(\mathbf{x}) = \begin{cases} s^2(\mathbf{x}), & \textit{if} \quad p=q=1, \\ m^2(\mathbf{x}), & \textit{otherwise.} \end{cases}$$

To finish complementary lemmas, we need to prove uniform convergence in probability of the factor $c_n(H) = \hat{m}_n(H)/f^{d+1}(\mathbf{x})$ arising in the decomposition of $R_n(H)$, towards the limit $c(\mathbf{x}) = m(\mathbf{x})/f^{d+1}(\mathbf{x})$.

LEMMA 3.10. *Assume* (W.1)–(W.2), (D.1)–(D.2), *and* (R.1)–(R.3). *Then*:

$$\hat{m}_n(H) \xrightarrow{P} m(x), \qquad \textit{as} \quad n \to \infty, \textit{uniformly in } \mathfrak{H}.$$

4. AUXILIARY RESULTS AND PROOFS

Results Involving Matrices

Let X be a $(n \times p)$-matrix, and s be a subset of $\{1, ..., n\}$. If X_s denotes the submatrix of X obtained from the rows of s, we have:

LEMMA 4.1 (Binet–Cauchy expansion). *Let X and Z be two $(n \times p)$-matrices $(n \geqslant p)$. Then*:

(i) $\det(X^T Z) = \sum_{s(p)} (\det X_s)(\det Z_s)$

(ii) $\det(X^T Z) = \binom{n-r}{r-p} \sum_{s(r)} \det(X_s^T Z_s)$; *for any $r \geqslant p$, where $\sum_{s(r)}$ means summation over all subsets of size r.*

Proof. (See Noble [25, p. 226] and Farebrother [14] for the use of this expansion in regression context). ∎

Proof of Theorem 2.1. It suffices to apply Lemma 4.1 to the numerator and to the denominator of $\hat{\alpha}$. Consequently,

$$\det \mathbf{X}^T\mathbf{W}\mathbf{X} = \sum_{\mathbf{j}} \det \mathbf{X}_{\mathbf{j}}^T \mathbf{W}_{\mathbf{j}} \mathbf{X}_{\mathbf{j}}$$

$$\det \mathbf{X}^T\mathbf{W}\mathbf{X}^{(1)}(\mathbf{Y}) = \sum_{\mathbf{j}} \det \mathbf{X}_{\mathbf{j}}^T \mathbf{W}_{\mathbf{j}} \mathbf{X}_{\mathbf{j}}^{(1)}(\mathbf{Y}_{\mathbf{j}}),$$

where $\sum_{\mathbf{j}}$ means summation over all subsets of size $(d+1)$ taken from $\{1, ..., n\}$. ∎

Results Involving the Optimal Choice of H_n

Proof of Lemma 2.1. To optimize the AMSE we must find a matrix A minimizing:

$$\text{trace}^2\,(A^T \mathscr{H}_m(\mathbf{x})\,A), \tag{4.1}$$

that is, minimizing the absolute value of the trace of $A^T \mathscr{H}_m(\mathbf{x})\,A$. Therefore, without loss of generality, we may suppose that $\mathscr{H}_m(\mathbf{x})$ is positive definite.

The rest of the proof is similar to that of Proposition 6 in Terrell and Scott [34], and it follows the optimal value of h_n,

$$h_n(\mathbf{x}) = n^{-1/(d+4)} \left\{ \frac{\sigma^2(\mathbf{x})\,R(K)}{d(\det \mathscr{H}_m(\mathbf{x}))^{2/d}\,f(\mathbf{x})} \right\}^{1/(d+4)},$$

and the optimal choice of A, which is any matrix satisfying

$$AA^T = a\mathscr{H}_m^{-1}(\mathbf{x}) \qquad \text{with} \quad a = \det \mathscr{H}_m(\mathbf{x})^{1/d}.$$

One possible choice of A is

$$A = a^{1/2} O(\mathbf{x})\,\Gamma(\mathbf{x})^{-1/2},$$

where $O(\mathbf{x})$ is an orthogonal matrix whose columns are eigenvectors and $\Gamma(\mathbf{x})$ is a diagonal matrix of the eigenvalues of $\mathscr{H}_m(\mathbf{x})$. ▮

Proof of Lemma 2.2. In this situation it holds that for any n a matrix A exists such that trace $(A^T \mathscr{H}_m(\mathbf{x})\,A) = 0$. Therefore, the leading term in the conditional bias is of a smaller order.

There are many possible choices of matrix A. One of the most simple is

$$A = a_1\,O(\mathbf{x})\,\Gamma_1^{1/2}(\mathbf{x}),$$

where $\Gamma_1(\mathbf{x})$ is a diagonal matrix such that trace$(\Gamma_1(\mathbf{x})\,\Gamma(\mathbf{x})) = 0$ and constant a_1 is such that det $A = 1$.

Let us suppose that there are l negative and $d - l$ positive eigenvalues. If λ is a negative one, then the corresponding element of Γ_1 will be $(-\lambda)^{-1}/l$, and if κ is a positive one, then the corresponding element of Γ_1 will be $(\kappa)^{-1}/(d-l)$. ▮

Moments of Random Determinants

To find moments associated with some determinants, we will use a representation allowing us to interchange the order between the expectation and the determinant operators. If X is a squared $(p \times p)$-matrix and X_π denotes this matrix when we permute the rows of X through the permutation π, it easily follows that

$$\det(X^T X) = \sum_{\pi} \det\{(X_\pi)^T \operatorname{Diag}(X_\pi)\}, \tag{4.2}$$

where $\sum_\pi$ means summation over all permutations of $(1, ..., p)$ and $\operatorname{Diag}(X)$ denotes the diagonal matrix having only diagonal elements of X. This last identity is obtained by writing

$$\det X = \sum_{\pi} (-1)^{\sigma(\pi)} \det(\operatorname{Diag}(X_\pi)),$$

where $\sigma(\pi)$ is the parity of permutation π.

Now if X is a $(n \times p)$ data matrix in a random design, whose rows are independent, and s is a subset of indices of size p, it follows that $(X_{\pi(s)})^T$ $\operatorname{Diag}(X_{\pi(s)})$ is a $(p \times p)$-matrix whose columns are independent, and we can interchange the expectation and the determinant operators; that is,

$$\mathrm{E}(\det\{(X_{\pi(s)})^T \operatorname{Diag}(X_{\pi(s)})\}) = \det\{\mathrm{E}[(X_{\pi(s)})^T \operatorname{Diag}(X_{\pi(s)})]\}.$$

$U_n(\mathbf{x}; H_n)$ and $V_n(\mathbf{x}; H_n)$ as U-Statistics

Let us define the following matrices, associated with the exact expectation of the statistics $U_n(\mathbf{x}; H_n)$ and $V_n(\mathbf{x}; H_n)$,

$$\Psi_1(\mathbf{x}; H_n) = \int_{\mathbb{R}^d} \begin{bmatrix} 1 \\ u \end{bmatrix} [m(\mathbf{x} + H_n u), u^T]\, K(u)\, f(\mathbf{x} + H_n u)\, du,$$

$$\Psi_2(\mathbf{x}; H_n) = \int_{\mathbb{R}^d} \begin{bmatrix} 1 \\ u \end{bmatrix} [1, u^T]\, K(u)\, f(\mathbf{x} + H_n u)\, du.$$

Suppose (W.1)–(W.2), (D.1)–(D.2), and (R.1)–(R.3) hold. A Taylor series expansions of the functions $m(\,.\,)$ and $f(\,.\,)$, stand the diagonal expansion for the determinant of the sum of two matrices (see Searle [28]) gives us

$$\begin{aligned} \det \Psi_2(\mathbf{x}; H_n) &\\ = f(\mathbf{x})^{d+1} &+ f(\mathbf{x})^d \left\{\tfrac{1}{2}(\mu_4 + (d-1)\,\mu_{22} + 1)\operatorname{trace}(H_n^T \mathscr{H}_f(\mathbf{x})\, H_n)\right\} \\ &- f(\mathbf{x})^{d-1} \operatorname{trace}(H_n D_f^T(\mathbf{x})\, D_f(\mathbf{x})\, H_n^T) + o(\operatorname{trace}(H_n \mathbf{J} H_n^T)) \end{aligned} \tag{4.3}$$

$$\begin{aligned} \det \Psi_1(\mathbf{x}; H_n) &\\ = m(\mathbf{x})\, f(\mathbf{x})^{d+1} &+ f(\mathbf{x})^{d+1}\, \tfrac{1}{2} \operatorname{trace}(H_n^T \mathscr{H}_m(\mathbf{x})\, H_n) \\ &+ m(\mathbf{x})\, f(\mathbf{x})^d \left\{\tfrac{1}{2}(\mu_4 + (d-1)\,\mu_{22} + 1)\operatorname{trace}(H_n^T \mathscr{H}_f(\mathbf{x})\, H_n)\right\} \\ &- m(\mathbf{x})\, f(\mathbf{x})^{d-1} \operatorname{trace}(H_n^T D_f(\mathbf{x})\, D_f^T(\mathbf{x})\, H_n) + o(\operatorname{trace}(H_n^T \mathbf{J} H_n)). \end{aligned} \tag{4.4}$$

Let $\mathbf{j}$ be a subset of $(d+1)$ different indices. When $i \in \mathbf{j}$, it is easily shown that

$$\mathrm{E}(\det \mathbf{X}_{\mathbf{j}}^T \mathbf{W}_{\mathbf{j}} \mathbf{X}_{\mathbf{j}} \mid X_i) = d!\,(\det H_n)^2\, K_H(X_i - \mathbf{x}) \sum_{p=1}^{d+1} \det \Psi_2^{(p)}((\mathbf{z}_i)_p\, \mathbf{z}_i)$$

with

$$\mathbf{z}_i = \begin{bmatrix} 1 \\ H_n^{-1}\,(X_i - \mathbf{x}) \end{bmatrix} \tag{4.5}$$

and, providing there is no confusion, we denote

$$\det \Psi_2^{(p)}(X_i) \equiv \det \Psi_2^{(p)}((\mathbf{z}_i)_p\, \mathbf{z}_i).$$

In a similar way, we have

$$\begin{aligned} &\mathrm{E}(\det \mathbf{X}_{\mathbf{j}}^T \mathbf{W}_{\mathbf{j}} \mathbf{X}_{\mathbf{j}}^{(1)}(\mathbf{Y}_{\mathbf{j}}) \mid Y_i, X_i) \\ &\quad = d!\,(\det H_n)^2\, K_H(X_i - \mathbf{x}) \det \Psi_1^{(1)}(Y_i \mathbf{z}_i) \\ &\qquad + d!\,(\det H_n)^2\, K_H(X_i - \mathbf{x}) \sum_{p=2}^{d+1} \det \Psi_1^{(p)}((\mathbf{z}_i)_p\, \mathbf{z}_i), \end{aligned} \tag{4.6}$$

where $\mathbf{z}_i$ is defined as (4.5). Providing there is no confusion, we denote,

$$\det \Psi_1^{(p)}(X_i, Y_i) \equiv \begin{cases} \det \Psi_1^{(1)}(Y_i \mathbf{z}_i), & \text{if } \; p = 1, \\ \det \Psi_1^{(p)}((\mathbf{z}_i)_p\, \mathbf{z}_i), & \text{if } \; p \neq 1. \end{cases}$$

LEMMA 4.2. *If the corresponding expectations are assumed to exist, the Hájek projections of the statistics $V_n(\mathbf{x}; H_n)$ and $U_n(\mathbf{x}; H_n)$ are given by*

$$\begin{aligned} \text{(i)}\quad \hat{V}_n(\mathbf{x}) - \det \Psi_2(\mathbf{x}; H_n) = \frac{1}{n} \sum_{i=1}^{n} \sum_{p=1}^{d+1} \{ &K_H(X_i - \mathbf{x}) \det \Psi_2^{(p)}(X_i) \\ &- \det \Psi_2(\mathbf{x}; H_n) \}. \end{aligned} \tag{4.7}$$

$$\begin{aligned} \text{(ii)}\quad \hat{U}_n(\mathbf{x}) - \det \Psi_1(\mathbf{x}; H_n) = \frac{1}{n} \sum_{i=1}^{n} \sum_{p=1}^{d+1} \{ &K_H(X_i - \mathbf{x}) \det \Psi_1^{(p)}(Y_i, X_i) \\ &- \det \Psi_1(\mathbf{x}; H_n) \}. \end{aligned} \tag{4.8}$$

Proof. Let i be a fixed index from $\{1, ..., n\}$. There are $\binom{n-1}{d}$ subsets of size $(d+1)$ that include the index i. Moreover, from $(d+1)!$ ordered tuples associated to one of these subsets, there are $d!$ of these, so that i is fixed in a position p. Then, using (4.6), we obtain the following expression of the conditional expectation of U_n, given X_i:

$$\mathrm{E}(U_n(\mathbf{x}; H_n) \mid X_i) = \frac{\binom{n-1}{d} d!}{\binom{n}{d+1} (d+1)!} \sum_{p=1}^{d+1} \{K_H(X_i - \mathbf{x}) \det \Psi_2^{(p)}(X_i)\}$$

$$+ \frac{\binom{n-1}{d+1} (d+1)!}{\binom{n}{d+1} (d+1)!} \det \Psi_2(\mathbf{x}; H_n).$$

By adding up when i varies, we have the first required result. The second is derived with the same arguments, by replacing $\Psi_2^{(p)}(X_i)$ with $\Psi_1^{(q)}(X_i, Y_i)$ and $\Psi_2(\mathbf{x}; H_n)$ with $\Psi_1(\mathbf{x}; H_n)$. ∎

As we have mentioned earlier, U_n and V_n are symmetric statistics whose kernels depend on n through H_n. The following lemma is a consequence of this, and its proof is very similar to that of Lemmas A and B in Serfling [29, pp. 185–186].

Lemma 4.3. *If the corresponding moments of order r $(r \geqslant 2)$ are assumed to exist, we have*

$$\mathrm{E}(U_n(\mathbf{x}) - \mathrm{E}(U_n(\mathbf{x})))^r = O((n \det H_n)^{-[(1/2)(r+1)]}), \tag{4.9}$$

where $[\,.\,]$ denotes the integer part. If the symmetric statistic is degenerate up to the order c (see Serfling [29] for the definition), then

$$\mathrm{E}(U_n(\mathbf{x}) - \mathrm{E}(U_n(\mathbf{x})))^r = O((n \det H_n)^{-[(1/2)(rc+1)]}).$$

Corollary 4.1. *If the corresponding moments are assumed to exist we have*

$$\mathrm{E}(U_n(\mathbf{x}) - \hat{U}_n(\mathbf{x})))^r = O((n \det H_n)^{-r}).$$

Proof. Showing that $U_n(\mathbf{x}) - \hat{U}_n(\mathbf{x})$ is a symmetric statistic degenerate up to the order 2, is straightforward. From the previous Lemma 4.3, it follows that

$$\mathrm{E}(U_n(\mathbf{x}) - \hat{U}_n(\mathbf{x})))^r = O((n \det H_n)^{-[(1/2)(2r+1)]}) = O((n \det H_n)^{-r}). \quad ∎$$

Asymptotic Normality

Let $\Psi_{k,(p,i)}(\mathbf{x}; H_n)$ $(k = 1, 2)$ be the cofactor associated to the (p, i) element of matrix $\Psi_k(\mathbf{x}; H_n)$ $(k = 1, 2)$.

We will denote, for now:

$$B_n = (n \det H_n)^{1/2}\, \mathrm{E}(\hat{V}_n - f(\mathbf{x})^{d+1}) \\ = (n \det H_n)^{1/2}\, (\det \Psi_2(\mathbf{x}; H_n) - f(\mathbf{x})^{d+1}) \tag{4.10}$$

and

$$S_n = \mathrm{Var}\{(nh_n^d)^{1/2}\, (\hat{V}_n - f(\mathbf{x})^{d+1})\} \\ = h_n^d\, \mathrm{Var}\left\{K_H(X - \mathbf{x}) \sum_{p=1}^{d+1} \det \Psi_2^{(p)}(X)\right\}. \tag{4.11}$$

Lemma 4.4. *Assume* (W.1)–(W.2), (D.1), *and* (H.2) *hold. Then*

$$\sup_{z \in \mathbb{R}} \left| \Pr\{(n \det H_n)^{1/2}\, (\hat{V}_n - f(\mathbf{x})^{d+1}) \leqslant z\} - \Phi\left(\frac{z - B_n}{S_n^{1/2}}\right)\right| = o(1).$$

Proof. We will apply the Berry–Essen inequality to the following random variables,

$$Z_i = \sum_{p=1}^{d+1} \{K_H(X_i - \mathbf{x}) - \det \Psi_2^{(p)}(X_i) \det \Psi_2(\mathbf{x}; H_n)\}, \qquad i = 1, ..., n,$$

which are independent, with zero mean and a common distribution function. For a detailed proof, see Cristóbal and Alcalá [7]. ▮

If we consider approximations to B_n and S_n, by using Lemma 4.2 of Cao [5], we can obtain the following corollary. Thus, we consider:

$$B(\mathbf{x}) = (\det H)^{1/2} \left\{\frac{f(\mathbf{x})^d}{2}\, (\mu_4 + (d-1)\, \mu_{22} + 1)\, \mathrm{trace}(H^T \mathscr{H}_f(\mathbf{x})\, H)\right\} \\ - (\det H)^{1/2}\, f(\mathbf{x})^{d-1}\, \mathrm{trace}(H^T D_f(\mathbf{x})\, D_f^T(\mathbf{x})\, H)$$

and

$$S(\mathbf{x}) = d_K f^{2d+1}(\mathbf{x}).$$

Corollary 4.2. *Assume* (W.1)–(W.2), (D.1)–(D.2), *and* (H.2) *hold. Then*

$$\sup_{z \in \mathbb{R}} \left| \Pr\{(n \det H_n)^{1/2}\, (\hat{V}_n - f(\mathbf{x})^{d+1}) \leqslant z\} - \Phi\left(\frac{z - B(\mathbf{x})}{S(\mathbf{x})^{1/2}}\right)\right| = o(1).$$

Proof. We must bound the order between B_n and $B(\mathbf{x})$ as well as between S_n and $S(\mathbf{x})$. From the assumption (D.2) and (H.2), some more detailed expansions of det $\Psi_2(\mathbf{x}; H_n)$ are followed, and we can obtain the order of the different traces of the minors of the matrix. Thus, we have

$$B_n = B(\mathbf{x}) + o(1). \tag{4.12}$$

For the variance, we must reason on the cofactors of $\Psi_2(\mathbf{x})$, and a little basic algebra gives us:

$$\det \Psi_{2(p,q)}(\mathbf{x}) = \begin{cases} f(\mathbf{x})^d + O(h_n^2) & \text{if it is a principal cofactor,} \\ O(h_n) & \text{if it is a nonprincipal cofactor.} \end{cases}$$

From these approximations and the symmetry of $K^2(u)$, we obtain the order in the difference

$$\begin{aligned} S_n - S(\mathbf{x}) &= f(\mathbf{x})^{2d+1}\, d_K + o(1) - h_n^d(f(\mathbf{x})^{2(d+1)} + o(1)) - f(\mathbf{x})^{2d+1}\, d_K \\ &= o(1). \end{aligned} \tag{4.13}$$

Finally, from (4.12) and (4.13), Lemma 2.4 of Cao [5], and using the triangular inequality, we obtain

$$\sup_{z \in \mathbb{R}} \left| \Pr\{(n \det H_n)^{1/2}\, (\hat{U}_n - f(\mathbf{x})^{d+1}) \leqslant z\} - \Phi\left(\frac{z - B(\mathbf{x})}{S(\mathbf{x})^{1/2}}\right) \right| = o(1). \quad \blacksquare$$

From the projection of the statistic $U_n(\mathbf{x}; H_n)$, we have similar results. We will suppose that all functions arising are bounded. Let us define

$$\begin{aligned} B_n &= (n \det H_n)^{1/2}\, \mathrm{E}(\hat{U}_n - m(\mathbf{x})\, f(\mathbf{x})^{d+1}) \\ &= (nh_n^d)^{1/2}\, (\det \Psi_1(\mathbf{x}; H_n) - m(\mathbf{x})\, f(\mathbf{x})^{d+1}). \end{aligned}$$

Observe that the different det $\Psi_1^{(p)}(X_i, Y_i)$ only contain those values of Y_i associated to the first column (i.e., $p = 1$). The term associated to variance is

$$\begin{aligned} S_n &= \mathrm{Var}\{(nh_n^d)^{1/2}\, (\hat{U}_n - m(\mathbf{x})\, f(\mathbf{x})^{d+1})\} \\ &= h_n^d\, \mathrm{Var}\left\{K_H(X - \mathbf{x}) \sum_{p=1}^{d+1} \det \Psi_1^{(p)}(Y, X)\right\}. \end{aligned}$$

LEMMA 4.5. *Assume* (W.1)–(W.2), (D.1)–(D.2), (R.1), (R.2), (R.4), *and* (H.2) *hold. Then*

$$\sup_{z \in \mathbb{R}} \left| \Pr\{(n \det H_n)^{1/2}\, (\hat{U}_n - m(\mathbf{x})\, f(\mathbf{x})^{d+1}) \leqslant z\} - \Phi\left(\frac{z - B_n}{S_n^{1/2}}\right) \right| = o(1).$$

Proof. The proof is similar to that of Lemma 4.6; by applying the Berry–Essen inequality to variables,

$$Z_i = K_H(X_i - \mathbf{x}) \sum_{p=1}^{d+1} \det \Psi_1^{(p)}(Y_i, X_i) - (d+1) \det \Psi_1(\mathbf{x}; H_n), \qquad i = 1, ..., n,$$

which are independent with zero mean and common distribution function. ▮

If we consider only the leading terms of B_n and S_n, we obtain a similar result. Now, these terms are:

$$\begin{aligned} B(\mathbf{x}) = {} & f(\mathbf{x})^{d+1} (\det H)^{1/2} \tfrac{1}{2} \operatorname{trace}(H^T \mathscr{H}_m(\mathbf{x}) H) \\ & + m(\mathbf{x}) f(\mathbf{x})^d (\det H)^{1/2} \{ \tfrac{1}{2}(\mu_4 + (d-1) \mu_{22} + 1) \operatorname{trace}(H^T \mathscr{H}_f(\mathbf{x}) H) \} \\ & - m(\mathbf{x}) f(\mathbf{x})^{d-1} (\det H)^{1/2} \operatorname{trace}(H^T D_f(\mathbf{x}) D_f(\mathbf{x})^T H) \end{aligned}$$

and

$$S(\mathbf{x}) = f^{2d+1}(\mathbf{x}) \{ s^2(\mathbf{x}) d_{1K} + m^2(\mathbf{x}) d_{2K} \}.$$

Corollary 4.3. *Assume* (W.1)–(W.2), (D.1)–(D.2), *and* (R.1)–(R.4) *hold. Then*

$$\sup_{z \in \mathbb{R}} \left| P\{ (n \det H_n)^{1/2} (\hat{U}_n - m(\mathbf{x}) f(\mathbf{x})^{d+1}) \leqslant z \} - \Phi\left(\frac{z - B(\mathbf{x})}{S(\mathbf{x})^{1/2}} \right) \right| = o(1).$$

Proof. Similar to the proof of Corollary 4.4. ▮

Other Proofs of Lemmas and Theorems

Proof of Lemma 3.1. Remember that $H_n = n^{-1/(d+4)} H$, with $H \in \mathfrak{H}$. Let B_n^* be a block matrix, with block elements

$$\begin{aligned} B_{n,11}^* &= \tfrac{1}{2} \operatorname{trace}(H_n^T \mathscr{H}_f(\mathbf{x}) H_n), \\ B_{n,12}^* &= H_n^T D_f(\mathbf{x}), \\ B_{n,21}^* &= B_{n,12}^{*T}, \\ B_{n,22}^* &= \int u u^T (u^T H_n^T \mathscr{H}_f(\mathbf{x}) H_n u) K(u) \, du, \end{aligned}$$

and we define $\Psi_2^*(\mathbf{x}; H_n) = f(\mathbf{x}) I_{d+1} + B_n^*$.

The diagonal expansion (see Searle [28]) leads us to write $\det(f(\mathbf{x}) I_{d+1} + A)$ as

$$\sum_{k=0}^{d+1} f(\mathbf{x})^{d+1-k} \operatorname{trace}_k (A),$$

where A is $(d+1)\times(d+1)$ matrix and $\mathrm{trace}_k(A)$ denotes the sum of the principal minors of order k ($\mathrm{trace}_0(A)=1$ and $\mathrm{trace}_{d+1}(A)=\det(A)$).

Using this diagonal expansion, it is easy to see that

$$n^{2/(d+4)}[f(x)^d \,\mathrm{trace}_1\,(B_n^*) \\ + f(\mathbf{x})^{d-1}\,\mathrm{trace}_2\,(B_n^*)] - \zeta(H) \xrightarrow[n\to\infty]{} 0, \qquad \text{uniformly in } \mathfrak{H},$$

and that $n^{2/(d+4)}\,\mathrm{trace}_k\,(B_n^*) \xrightarrow[n\to\infty]{} 0$, $(k\geqslant 3)$ uniformly in $\mathfrak{H}$, and then

$$n^{2/(d+4)} \det \Psi_2^*(\mathbf{x}; H_n) - n^{2/(d+4)} f(\mathbf{x})^{d+1} - \zeta(H) \xrightarrow[n\to\infty]{} 0, \qquad \text{uniformly in } \mathfrak{H}.$$

Let $B_n = \Psi_2(\mathbf{x}\colon H_n) - f(\mathbf{x})\, I_{d+1}$ be the matrix with the remaining terms of the Taylor expansion up to order two of $\Psi_2(\mathbf{x}; H_n)$. The second derivatives continuity implies that

$$n^{2/(d+4)}[\mathrm{trace}_1\,(B_n) - \mathrm{trace}_1\,(B_n^*)] \xrightarrow[n\to\infty]{} 0, \qquad \text{uniformly in } \mathfrak{H},$$

and

$$n^{2/(d+4)}[\mathrm{trace}_2\,(B_n) - \mathrm{trace}_2\,(B_n^*)] \xrightarrow[n\to\infty]{} 0, \qquad \text{uniformly in } \mathfrak{H},$$

and, as before, $n^{2/(d+4)}\,\mathrm{trace}_k\,(B_n) \xrightarrow[n\to\infty]{} 0$, $k\geqslant 3$, uniformly in $\mathfrak{H}$ and, therefore,

$$n^{2/(d+4)}[\det \Psi_2(\mathbf{x}; H_n) - \det \Psi_2^*(\mathbf{x}; H_n)] \xrightarrow[n\to\infty]{} 0, \qquad \text{uniformly in } \mathfrak{H},$$

from part (i) of the lemma follows.

Proof of (ii) is similar to part (i). ∎

Proof of Lemma 3.2. It follows from Lemma 4.3 and Corollaries 4.1, 4.2, and 4.3. ∎

Proof of Lemma 3.3. We can simplify by analyzing the Hájek projection, i.e., $\hat{Z}(A) = n^{2/(d+4)}[\mathrm{E}(\hat{V}_n(A) - \mathrm{E}\hat{V}_n(A))]$, since

$$\mathrm{E}(V_n(A)\, V_n(B)) = \mathrm{E}(\hat{V}_n(A)\, \hat{V}_n(B)) + o(n^{-4/(d+4)}).$$

As in this process we have sums of random variables i.i.d., the covariance in it is easily found:

$$\mathrm{Cov}(\hat{Z}_n(A), \hat{Z}_n(B)) \\ = \frac{1}{\det A \det B} \int \sum_{p,\,q=1}^{d+1} \det \Psi_2^{(p)}(A^{-1}u) \det \Psi_2^{(q)}(B^{-1}u) \\ \times K(A^{-1}u)\, K(B^{-1}u)\, f(\mathbf{x} + n^{-1/(d+4)}u)\, du \\ - n^{-d/(d+4)} \det \Psi_2(\mathbf{x}; A) \det \Psi_2(\mathbf{x}; B).$$

The second term on the right-hand side is of order $o(1)$ and the principal cofactors of diagonal elements are dominant over the determinants in the first term. We can write them as $f^d(\mathbf{x}) + o(1)$, and they do not depend on A and B. So, the result follows. ▮

Proof of Lemma 3.4. According to the above discussion, it is sufficient to find the asymptotic covariance between the elements of the Hájek projection. As the functions arising are continuous and bounded, we have

$$\begin{aligned}
&\mathrm{Cov}(\hat{Y}_n(A), \hat{Y}_n(B)) \\
&\quad = \frac{1}{\det A \det B} \int\!\!\int \sum_{p,q=1}^{d+1} \det \Psi_1^{(p)}(A^{-1}u, v) \det \Psi_1^{(q)}(B^{-1}u, v) \\
&\qquad \times K(A^{-1}u)\, K(B^{-1}u)\, f_{XY}(\mathbf{x} + n^{-1/(d+4)}, u, v)\, du\, dv \\
&\qquad - n^{-d/(d+4)} \det \Psi_1(\mathbf{x}; A) \det \Psi_1(\mathbf{x}; B).
\end{aligned}$$

Remember that the variable Y only appears in the case $p = 1$. So, if $p = q = 1$, after we apply the Fubini theorem, function $s^2(\mathbf{x} + n^{-1/(d+4)}u)$ is in the first term on the right-hand side. If either $p = 1$ or $q = 1$ (but not both), the corresponding function is $m(\mathbf{x} + n^{-1/(d+4)}u)$. As these functions are continuous, the expansion of $\det \Psi_1(\mathbf{x}; H_n)$ by the principal cofactors complete the proof. ▮

Proof of Lemma 3.5. Again, it is sufficient to find the covariance between the Hájek projections of $Y_n(A)$ and $Z_n(B)$. The same as in Lemmas 3.3 and 3.4, we have

$$\begin{aligned}
&\mathrm{Cov}(\hat{Y}_n(A), \hat{Z}_n(B)) \\
&\quad = \frac{1}{\det A \det B} \int\!\!\int \sum_{p,q=1}^{d+1} \det \Psi_1^{(p)}(A^{-1}u, v) \det \Psi_2^{(q)}(B^{-1}u) \\
&\qquad \times K(A^{-1}u)\, K(B^{-1}u)\, f_{XY}(\mathbf{x} + n^{-1/(d+4)}u, v)\, du\, dv \\
&\qquad - n^{-d/(d+4)} \det \Psi_1(\mathbf{x}; A) \det \Psi_2(\mathbf{x}; B).
\end{aligned}$$

Applying the Fubinni theorem and analyzing case $p = 1$, the result follows. ▮

Proof of Lemma 3.6. First, we will prove (ii). From the connection between matrices and vectors of $\mathbb{R}^{d^2}$ through the vec(.) operator ($H_{ij}(u) \sim \mathrm{vec}(H_{ij}(u))$), and the application of mean value theorem for vector functions, there exists a $\theta \in (0, 1)$ such that

$$V_n(H_{nij}(t)) - V_n(H_{nij}(s)) \\ = \mathrm{d}V_n(H_{nij}(t+\theta(s-t); n^{-1/(d+4)}(t-s)\, E_{ij}), \tag{4.14}$$

where E_{ij} denotes the elemental matrix with a 1 in the (i, j) entry and zeros elsewhere, and the differential operator is denoted by "d." We study the differential in (4.14) for any matrix H:

$$\mathrm{d}V_n(H; \mathrm{d}H) = \frac{1}{n(n-1)\cdots(n-d)} \sum_{\mathbf{j}} (\det \mathbf{X}_{\mathbf{j}}^T \mathbf{X}_{\mathbf{j}})\, \mathrm{d}\left(\frac{\det \mathbf{W}_{\mathbf{j}}(H)}{(\det H)^2}\right). \tag{4.15}$$

Applying basic differential matricial calculus (Magnus and Neudecker [19]), we can see that

$$\mathrm{d}\left(\frac{\det \mathbf{W}_{\mathbf{j}}(H)}{(\det H)^2}\right) = -\left\{(d+3)\frac{\det \mathbf{W}_{\mathbf{j}}(H)}{(\det H)^2}\,\mathrm{trace}(H^{-1}\mathrm{d}H) \right. \\ \left. + \sum_{i=1}^{d+1} \frac{\det \mathbf{W}_{\mathbf{j}}^{(i)}(H)}{(\det H)^2}\right\}, \tag{4.16}$$

where the matrix $\mathbf{W}_{\mathbf{j}}^{(i)}(H)$ is a diagonal matrix agreeing with $\mathbf{W}_{\mathbf{j}}(H)$ except in the ith diagonal element, which is $\zeta_K(H; H^{-1}(X_{\mathbf{j}_i} - \mathbf{x}))$, instead of $K_H(X_{\mathbf{j}_i} - \mathbf{x})$, where $\zeta_K(H; u)$ is

$$\zeta_K(H; u) = \mathrm{trace}(uD_K^T(u)\, H^{-1}\mathrm{d}H)(\det H)^{-1}$$

with $D_K(u)$ denoting the gradient vector of $K(u)$. Note that the statistic in (4.16) is still symmetric in the observations, which allows us to use the results involving such statistics.

In our case (if t_θ denotes point $t+\theta(s-t)$, and $H_{nij}(t_\theta) = n^{-1/(d+4)} H_{ij}(t_\theta)$, with $H_{ij}(t_\theta) \in \mathscr{H}$ and $\mathrm{d}H = n^{-1/(d+4)}(t-s)\, E_{ij}$), it follows that

$$\mathrm{trace}(H_n^{-1}\mathrm{d}H_n) = h_n^{ij} n^{-1/(d+4)}(s-t) = h^{ij}(s-t) \tag{4.17}$$

with h_n^{ij} and h^{ij} denoting the (i, j) entry of the matrices $H_{nij}^{-1}(t_\theta)$ and $H_{ij}^{-1}(t_\theta)$. Moreover, we have

$$\zeta_K(H_{nij}(t_\theta); u) = (\det H_{nij}(t_\theta))^{-1}\, u_j D_K^T(u)\, h^{(i)}(s-t), \tag{4.18}$$

where $h^{(i)}$ denotes the ith column of $H_{ij}(t_\theta)$.

By using (4.17) and (4.18), the differential (4.16) can be expressed as

$$(s-t)\,\frac{1}{(\det H_{nij}(t_\theta))^2}\,\tilde{w}_{\mathbf{j}}(H_{nij}(t_\theta))$$

and recall that $\tilde{w}_{\mathbf{j}}$ is a symmetric statistic.

Under the assumptions concerning the kernel and its first-order derivatives (K.3), we can assure that the pth moments of the obtained symmetric statistic exist.

Lemma 4.3 implies that

$$\mathrm{E}(|\mathrm{d}V_n(H_{ij}(t_\theta)) - \mathrm{E}(\mathrm{d}V_n(H_{ij}(t_\theta)))|^p) = O(n^{\{-2/(d+4)\}\,p}).$$

From the above expressions, it follows that

$$\begin{aligned}\mathrm{E}(&|Z_n(H_{i,\,j}(t)) - Z_n(H_{i,\,j}(s))|^p)\\ &= |s-t|^p\, n^{2p/(d+4)}\mathrm{E}(|\mathrm{d}V_n(H_{ij}(t_\theta)) - \mathrm{E}(\mathrm{d}V_n(H_{ij}(t_\theta)))|^p)\\ &= |s-t|^p\, O(1).\end{aligned}$$

The same reasoning is applied to the first part of lemma, since the differential associated to process U_n does not depend on the observations $Y_1, ..., Y_n$ and agrees with the differential given. ▮

Proof of Lemma 3.7. In view of the Cauchy–Schwarz inequality, it is sufficient to prove that

$$\mathrm{E}(|\phi_n(\mathbf{B})|^{2p}) \leqslant C_2\mu^2(\mathbf{B}) \qquad \text{for} \quad p = d^2.$$

Let (i, j) be any pair of fixed indices associated to an entry of the matrices of parameters. Then, using the definition of block and the c_r-inequality, we can bound the above moment by

$$\mathrm{E}(|\phi_n(\mathbf{B})|^{2p}) \leqslant C_1 \sum_{l=0}^{2^{p-1}-2} \mathrm{E}\,|\phi_n(B_l(t_{ij})) - \phi_n(B_l(s_{ij}))|^{2p},$$

where B_l denotes any 2^{p-1} matrices in the definition of block, whose elements are $s_{pq} + \delta_{pq}(t_{pq} - s_{pq})$, except the (i, j) element, which is either t_{ij} or s_{ij}.

Applying Lemma 3.6 to each term in the above summation, it follows that a constant exists such as

$$\mathrm{E}(|\phi_n(\mathbf{B})|^{2p}) \leqslant C_2\,|t_{ij} - s_{ij}|^{2p},$$

which is obtained for any pair (i, j) of indices given, Then, we have

$$\mathrm{E}(|\phi_n(\mathbf{B})|^{2p}) \leqslant C_2 \min_{(i,j)} |t_{ij} - s_{ij}|^{2p} \leqslant C_2 \mu^2(\mathbf{B}). \quad \blacksquare$$

Proof of Lemma 3.8. As function $K(.)$ is continuous, these processes belong to $\mathscr{C}(\mathfrak{H})$.

For a fixed value $H_0 \in \mathfrak{H}$, from the asymptotic normality and Theorem 6.2 of Billingsley [4], the tightness of processes $Y_n(H_0)$ and $Z_n(H_0)$ holds. The condition in the moments of these processes, given in Lemma 3.7, imply the tightness of $Y_n(.)$ and $Z_n(.)$ (see Theorem 3 of Bickel and Wichura [3]). ∎

Proof of Lemma 3.9. Lemma 3.8 shows the tightness of these processes. The convergence of the finite-dimensional distributions to the corresponding limits are followed from Lemmas 3.2, 3.3, and 3.4 and the Cramer–Wold device. It is straightforward that the limit distributions agree with the corresponding finite-dimensional distributions of $Y(.)$ and $Z(.)$. ∎

Proof of Lemma 3.10. By the continuous mapping theorem (see Theorem 5.1 of Billingsley [4] and Corollary 1) and the weak convergence of the processes $Z_n(H)$ and $Y_n(H)$, we have

$$\sup_{H \in \mathfrak{H}} |Z_n(H)| \xrightarrow{\mathscr{L}} \sup_{H \in \mathfrak{H}} |Z(H)|$$

and

$$\sup_{H \in \mathfrak{H}} |Y_n(H)| \xrightarrow{\mathscr{L}} \sup_{H \in \mathfrak{H}} |Y(H)|.$$

Hence, applying Slutsky's theorem:

$$\sup_{H \in \mathfrak{H}} |V_n(H) - \mathrm{E}(V_n(H))| = n^{-2/(d+4)} \sup_{H \in \mathfrak{H}} |Z_n(H)| \xrightarrow{P} 0, \qquad \text{as} \quad n \to \infty.$$

A similar reasoning can be applied to uniform convergence of $U_n(H)$. Finally, combining these results with Lemma 3.1, we complete the proof. ∎

Proof of Theorem 3.1. First, observe that

$$c_n(H)\, Z_n(H) \Rightarrow c(\mathbf{x})\, Z(H),$$

with $c_n(H) = \hat{m}_n(H)/f^{d+1}(\mathbf{x})$. The last result follows from the equicontinuity in probability of $\hat{m}_n(H)$ and the same arguments as Müller and Prewitt [22, p. 181] adapted to the metric space of regular matrices with the Frobenius norm.

Combining this result with $Y_n(H) \Rightarrow Y(H)$, the weak convergence of $\{R_n(H)\}$ on $\mathscr{C}(\mathfrak{H})$ is derived. The expectation and the covariance structure of the limiting process are derived from Lemmas 3.1, 3.3, 3.4, and 3.5 and Slutsky's theorem. ▌

ACKNOWLEDGMENTS

The authors are grateful to the referees and the editor for their many valuable comments and suggestions which have greatly improved the paper.

REFERENCES

1. I. Abramson, Arbitrariness of the pilot estimator in adaptive kernel methods, *J. Multivariate Anal.* **12** (1982), 562–567.
2. P. K. Bhattacharya and H.-G. Müller, Asymptotics for nonparametric regression, *Sankhyā Ser. A* **55** (1993), 420–441.
3. P. J. Bickel and M. J. Wichura, Convergence criteria for multiparameter stochastic processes and some applications, *Ann. Math. Statist.* **42** (1971), 1656–1670.
4. P. Billingsley, "Convergence of Probability Measures," Wiley, New York, 1968.
5. R. Cao, "Aplicaciones y Nuevos Resultados del Métodos Bootstrap en la Estimación no Paramétrica de Curvas," Ph.D. dissertation, University of Santiago, Spain, 1990.
6. W. S. Cleveland, Robust locally weighted regression and smoothing scatterplots, *J. Amer. Statist. Assoc.* **74** (1979), 829–836.
7. J. A. Cristóbal and J. T. Alcalá, "Bandwidth Matrix Processes in Local Linear Smoothing of a Multivariate Regression Function," Technical Report, Sec. I, No. 5 University of Zaragoza, Spain, 1995.
8. R. L. Eubank, "Spline Smoothing and Nonparametric Regression," Marcel Dekker, New York, 1988.
9. J. Fan, Design-adaptive nonparametric regression, *J. Amer. Statist. Assoc.* **87** (1992), 998–1004.
10. J. Fan, T. Gasser, I. Gijbels, M. Brockmann, and J. Engel, "On Nonparametric Estimation via Local Polynomial Regression," Discussion Paper, Vol. 9511, Institute of Statistics, Catholic University of Louvain, Louvain-la-Neuve, Belgium, 1995.
11. J. Fan and I. Gijbels, Variable bandwidth and local linear regression smoothers, *Ann. Statist.* **20** (1992), 2008–2036.
12. J. Fan and I. Gijbels, "Local Polynomial Modelling and Its Applications," Chapman & Hall, London, 1996.
13. J. Fan and Q. Yao, Efficient estimation of conditional variance functions in stochastic regression, unpublished manuscript (1997).
14. R. W. Farebrother, Relations among subset estimators: A bibliographical note, *Technometrics* **22** (1985), 29–33.
15. T. Gasser and H.-G. Müller, Estimating regression functions and their derivatives by the kernel method, *Scand. J. Statist.* **11** (1984), 171–185.
16. W. Härdle, "Applied Nonparametric Regression," Cambridge Univ. Press, Boston, 1990.
17. T. J. Hastie and R. Tibshirani, "Generalized Additive Models," Chapman & Hall, London, 1990.
18. A. M. Krieger and J. Pickands III, Weak convergence and efficient density estimation at a point, *Ann. Statist.* **9** (1981), 1066–1078.

19. J. R. Magnus and H. Neudecker, "Matrix Differential Calculus," Wiley, Chichester, 1988.
20. Y. P. Mack and H.-G. Müller, Adaptive nonparametric estimation of a multivariate regression function, *J. Multivariate Anal.* **23** (1987), 169–182.
21. H.-G. Müller, "Nonparametric Regression Analysis of Longitudinal Data," Springer-Verlag, Berlin/New York, 1988.
22. H.-G. Müller and K. A. Prewitt, Multiparameter bandwidth processes and adaptive surface smoothing, *J. Multivariate Anal.* **47** (1993), 1–21.
23. H.-G. Müller and U. Stadtmüller, Variable bandwidth kernel estimators of regression curves, *Ann. Statist.* **15** (1987), 182–201.
24. E. A. Nadaraya, On estimating regression, *Theoret. Probab. Appl.* **9** (1964), 141–142.
25. B. Noble, "Applied Linear Algebra," Prentice-Hall, New York, 1969.
26. D. Ruppert and M. P. Wand, Multivariate locally weighted least squares regression, *Ann. Statist.* **22** (1994), 1346–1370.
27. D. W. Scott, "Multivariate Density Estimation: Theory, Practice and Visualization," Wiley, New York, 1992.
28. S. Searle, "Matrix Algebra Useful for Statistics," Wiley, New York, 1982.
29. R. J. Serfling, "Approximation Theorems of Mathematical Statistics," Wiley, New York, 1980.
30. J. S. Simonoff, "Smoothing in Statistics," Springer-Verlag, New York, 1996.
31. C. J. Stone, Consistent nonparametric regression, *Ann. Statist.* **5** (1977), 595–645.
32. C. J. Stone, Optimal rates of convergence for nonparametric estimators, *Ann. Statist.* **8** (1980), 1348–1360.
33. C. J. Stone, Optimal global rates of convergence for nonparametric regression, *Ann. Statist.* **10** (1982), 1040–1053.
34. G. R. Terrell and D. W. Scott, Variable kernel density estimation, *Ann. Statist.* **20** (1992), 1236–1265.
35. G. Wahba, "Spline Models for Observational Data," SIAM, Philadelphia, 1990.
36. M. P. Wand and M. C. Jones, Comparison of smoothing parametrizations in bivariate kernel density estimation, *J. Amer. Statist. Assoc.* **88** (1993), 520–528.
37. M. P. Wand and M. C. Jones, "Kernel Smoothing," Chapman & Hall, London, 1995.
38. G. S. Watson, Smooth regression analysis, *Sankhyā Ser. A* **26** (1964), 359–372.