On Variance-Stabilizing Multivariate Nonparametric Regression Estimation

Kiheiji NISHIDA [a] and Yuichiro KANAZAWA [b]

(a) Corresponding Author. Specially Appointed Researcher, Graduate School of Medicine, Osaka University, 1-7 Yamadaoka, Suita, Osaka 565-0871, JAPAN. E-mail: kiheiji.nishida@gmail.com

(b) Professor, Graduate School of Systems and Information Engineering, University of Tsukuba, 1-1-1 Ten-noh-dai, Tsukuba, Ibaraki 305-8573, JAPAN. E-mail: kanazawa@sk.tsukuba.ac.jp

Key Words: Bandwidth selection, Local linear estimator, Local variable bandwidth, Variance-stabilization.

ABSTRACT

The MSE-minimizing local variable bandwidth for the univariate local linear estimator (the LL) is well-known. This bandwidth does not stabilize variance over the domain. Moreover, in regions where a regression function has zero curvature, the LL estimator is discontinuous. In this paper, we propose a variance-stabilizing (VS) local variable diagonal bandwidth matrix for the multivariate LL estimator. Theoretically, the VS bandwidth can outperform the multivariate extension of the MSE-minimizing local variable scalar bandwidth in terms of asymptotic MISE and can avoid discontinuity created by the MSE-minimizing bandwidth. We present an algorithm for estimating the VS bandwidth and simulation studies.

1 Introduction

Suppose that we are interested in exploring the association between a set of stochastic covariates $\mathbf{X} = (X_1, ..., X_p)$ and the response Y. Nonparametric approaches to explain the conditional expectation, such as $E[Y|\mathbf{X}] = m(\mathbf{x})$, are preferable in many cases. In this paper, we will concentrate on the nonparametric kernel-type local linear estimator (henceforth the LL estimator), a popular approach in curve estimation, presented, for example, by Ruppert

and Wand (1994).

Let us consider a $p+1$-row vector $(X_{i1},...,X_{ip},Y_i)$ of random variables. We assume that $\mathbf{x}_{i\cdot} = (x_{i1},...,x_{ip})$, $i = 1,...,n$, are the realizations of random explanatory vectors $\mathbf{X}_{i\cdot} = (X_{i1},...,X_{ip})$, i. i. d. with respect to i whose joint density function $f_{\mathbf{X}}(\mathbf{x})$ is bounded away from zero on compact support $I^p \in R^p$. The n sample realizations of $(X_{i1},...,X_{ip})$ can be written as the covariate matrix $(\mathbf{x}_{\cdot 1}, \mathbf{x}_{\cdot 2},..., \mathbf{x}_{\cdot p})$, where $\mathbf{x}_{\cdot j} = (x_{1j}, x_{2j},..., x_{nj})^T$, $j = 1,...,p$. We assume that the jth random explanatory variable $\mathbf{X}_{\cdot j}$ may be correlated with the kth variable $\mathbf{X}_{\cdot k}$, $j \neq k$, and that the response Y_i, $i = 1,...,n$, is influenced by the corresponding explanatory vector $\mathbf{X}_{i\cdot}$ in the form of $m(\mathbf{X}_{i\cdot})$ and the disturbance U_i as,

$$Y_i \;=\; m(\mathbf{X}_{i\cdot}) + U_i,$$

where $m(\cdot)$ is $m : R^p \to R$ function of the $\mathbf{X}_{i\cdot}$. The $U_i|\mathbf{X}_{i\cdot}$'s, $i = 1,...,n$, are random variables independent with respect to i and are assumed to be independent of $\mathbf{X}_{j\cdot}$, $i \neq j$. We assume the first two conditional moments of $U_i|\mathbf{X}_{i\cdot}$ are

$$E_{U_i|\mathbf{X}_{i\cdot}}\left[U_i|\mathbf{X}_{1\cdot} = \mathbf{x}_{1\cdot},...,\mathbf{X}_{n\cdot} = \mathbf{x}_{n\cdot}\right] = 0, \quad E_{U_i^2|\mathbf{X}_{i\cdot}}\left[U_i^2|\mathbf{X}_{1\cdot} = \mathbf{x}_{1\cdot},...,\mathbf{X}_{n\cdot} = \mathbf{x}_{n\cdot}\right] \;=\; \sigma^2(\mathbf{x}_{i\cdot}). \quad (1)$$

These standard assumptions are summarized as **S 1-4** below :

S 1 Pairs of random variables $(\mathbf{X}_{i\cdot}, Y_i)$ are i. i. d. with respect to i;

S 2 The density of $\mathbf{X}$ is continuous and $0 < C_f \leq f_{\mathbf{X}}(\mathbf{x}) \leq C^f$ on compact support I^p;

S 3 Column vectors $\mathbf{X}_{\cdot j}, j = 1,...,n$, of the covariate matrix $(\mathbf{X}_{\cdot 1}, \mathbf{X}_{\cdot 2},..., \mathbf{X}_{\cdot n})$ are linearly independent with respect to j;

S 4 The $U_i|\mathbf{X}_{i\cdot}$'s, $i = 1,...,n$, are random variables independent with respect to i, and assumed to be independent of $\mathbf{X}_{j\cdot}$, $i \neq j$. The first two conditional moments of $U_i|\mathbf{X}_{i\cdot}$ are given in (1).

Let $K_{\mathbf{X}}(\mathbf{t})$ be the non-negative real-valued p-dimensional kernel function, where $\mathbf{t} = (t_1,...,t_p)$, satisfying the standard set of assumptions **K 1** below :

K 1 Let $K_{\mathbf{X}}(\mathbf{t})$ be the non-negative real-valued p-dimensional kernel density function satisfying

(i) $K_{\mathbf{X}}(\mathbf{t})$ is a compactly supported, bounded kernel such that $\int \cdots \int K_{\mathbf{X}}(\mathbf{t})\mathbf{t}^T\mathbf{t}d\mathbf{t} = \mu_2(K_{\mathbf{X}})\mathbf{I}_p$, where $\mu_2(K_{\mathbf{X}}) \neq 0$ and $\mathbf{I}_p$ is the p-dimensional identity matrix;

(ii) $\int \cdots \int t_1^{l_1} \cdots t_p^{l_p} K_{\mathbf{X}}(\mathbf{t})d\mathbf{t} = 0$ for all nonnegative integers $l_1, ..., l_p$ such that their sum is odd.

We assume that $\mathbf{H}$ is a p-dimensional symmetric positive definite-bandwidth matrix. The LL estimator of $m(\cdot)$ is given by the solution for β_0 that minimizes, along with the other β_j, $j = 1, ..., p$, the sum of weighted least squares,

$$\begin{aligned} &\min_{\beta_0,\beta_1,...,\beta_p} \sum_{i=1}^{n} \left[Y_i - \beta_0 - \sum_{j=1}^{p} \beta_j (x_{ij} - x_j) \right]^2 K_{\mathbf{X}} \left((\mathbf{x}_{i\cdot} - \mathbf{x})\mathbf{H}^{-1} \right) \\ &= \min_{\beta_0,\beta_1,...,\beta_p} \left[\mathbf{Y} - \mathbf{D}(\mathbf{x})\boldsymbol{\beta} \right]^T \mathbf{W}(\mathbf{x}) \left[\mathbf{Y} - \mathbf{D}(\mathbf{x})\boldsymbol{\beta} \right], \end{aligned} \tag{2}$$

where

$$\mathbf{D}(\mathbf{x}) = \begin{pmatrix} 1 & x_{11} - x_1 & x_{12} - x_2 & \dots & x_{1p} - x_p \\ 1 & x_{21} - x_1 & x_{22} - x_2 & \dots & x_{2p} - x_p \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} - x_1 & x_{n2} - x_2 & \dots & x_{np} - x_p \end{pmatrix},$$

$\mathbf{W}(\mathbf{x}) = \text{diag}\left(K_{\mathbf{X}}((\mathbf{x}_{1\cdot} - \mathbf{x})\mathbf{H}^{-1}), ..., K_{\mathbf{X}}((\mathbf{x}_{n\cdot} - \mathbf{x})\mathbf{H}^{-1})\right)$ is the matrix controlling the weight reflecting the relevant data points in calculating the LL estimator at $\mathbf{x}$, $\boldsymbol{\beta} = (\beta_0, \beta_1, ..., \beta_p)^T$ is the local linear coefficient vector, $\boldsymbol{Y} = (Y_1, ..., Y_n)^T$ is the vector of responses with length n and the term $\beta_0 + \sum_{j=1}^{p} \beta_j (x_{ij} - x_j)$ is the linear approximation of $m(\mathbf{x})$ in the neighborhood of $\mathbf{x}$. Solving the minimization problem (2) with respect to $\boldsymbol{\beta}$ retaining its intercept term β_0, we obtain the LL estimator,

$$\widehat{m_{\mathbf{H}}^{LL}}(\mathbf{x}) = \mathbf{e}_1 \left[\mathbf{D}^T(\mathbf{x})\mathbf{W}(\mathbf{x})\mathbf{D}(\mathbf{x}) \right]^{-1} \left[\mathbf{D}^T(\mathbf{x})\mathbf{W}(\mathbf{x})\mathbf{Y} \right],$$

where $\mathbf{e}_1$ is a $1 \times (p+1)$ row vector with 1 as the first entry and 0 for all other entries.

The variance and the bias of the LL estimator are well known. See Ruppert and Wand (1994). With the standard set of assumptions on kernel **K 1** and the additional assumptions **S 5-7** concerning $\sigma^2(\mathbf{x})$, $m(\mathbf{x})$ and $\alpha_{ij}(\mathbf{x})$, where

$$\alpha_{ij}(\mathbf{x}) = \frac{\partial}{\partial x_j}\left[\frac{\partial m(\mathbf{x})}{\partial x_i}\right], \quad \text{for } i = 1, ..., p,\ j = 1, ..., p: \tag{3}$$

S 5 $m(\mathbf{x})$ is twice continuously differentiable with respect to x_i, $i = 1, ..., p$;

S 6

(i) $m(\mathbf{x})$ are not linear functions such as $b_0 + \sum_{i=1}^{p} b_i x_i$;

(ii) $\sum_{i=1}^{p} \alpha_{ii}(\mathbf{x}) \neq 0$ on some nonzero set within I^p;

S 7 The conditional variance is continuous and $0 < C_{\sigma^2} \leq \sigma^2(\mathbf{x}) \leq C^{\sigma^2}$ for compact support I^p,

and assumptions concerning $\mathbf{H}$ written as follows:

A 1 All the entries in $\mathbf{H}$ converge to 0 as $n \to \infty$;

A 2 $n|\mathbf{H}| \to \infty$ as $n \to \infty$,

the theoretical conditional variance of the LL estimator is written as follows:

$$V_{\mathbf{X}_{i\cdot},Y_i}\left[\widehat{m_{\mathbf{H}}^{LL}}(\mathbf{x})\middle|\, \mathbf{X}_{1\cdot} = \mathbf{x}_{1\cdot}, ..., \mathbf{X}_{n\cdot} = \mathbf{x}_{n\cdot}\right] = \frac{1}{n|\mathbf{H}|}\frac{\sigma^2(\mathbf{x})}{f_{\mathbf{X}}(\mathbf{x})} R(K_{\mathbf{X}})\left(1 + o_p\left(1\right)\right), \tag{4}$$

where $R(K_{\mathbf{X}}) = \int \cdots \int K_{\mathbf{X}}^2(\mathbf{t})d\mathbf{t}$. Similarly, the theoretical conditional bias for the LL estimator at $\mathbf{x}$ is known to be

$$\begin{aligned} &E_{\mathbf{X}_{i\cdot},Y_i}\left[\widehat{m_{\mathbf{H}}^{LL}}(\mathbf{x})\middle|\, \mathbf{X}_{1\cdot} = \mathbf{x}_{1\cdot}, ..., \mathbf{X}_{n\cdot} = \mathbf{x}_{n\cdot}\right] - m(\mathbf{x}) \\ &= \frac{\mu_2(K_{\mathbf{X}})}{2}\text{trace}\left[\mathbf{H}^T\nabla^2 m(\mathbf{x})\mathbf{H}\right] + o_p\left(\text{trace}\left(\mathbf{H}^T\mathbf{H}\right)\right), \end{aligned} \tag{5}$$

where the Hessian matrix is defined to be

$$\nabla^2 m(\mathbf{x}) = \begin{pmatrix} \alpha_{11}(\mathbf{x}) & \dots & \alpha_{1p}(\mathbf{x}) \\ \vdots & \ddots & \vdots \\ \alpha_{p1}(\mathbf{x}) & \dots & \alpha_{pp}(\mathbf{x}) \end{pmatrix}.$$

Because the leading terms in (4) and (5) do not depend on $\mathbf{X}_{1\cdot}, ..., \mathbf{X}_{n\cdot}$, they play the role of unconditional variance and bias, respectively, as stated in Ruppert and Wand (1994).

If we assume

> **A 3** The data $\mathbf{X}_{i\cdot}$'s are distributed as approximately multivariate normal so that the p-dimensional bandwith matrix $\mathbf{H}$ is assumed to be diagonal, $\mathbf{H} = \text{diag}(h_{11}, h_{22}, ..., h_{pp})$,

the sphering approach as presented by Wand and Jones (1993) is appropriate, and we do not have to parametrize the off-diagonal elements of bandwidth matrix that reflect the correlation between explanatory variables. Therefore, the bandwidth matrix is diagonal.

As most nonparametric regression estimators choose their bandwidth by balancing the squared bias and the variance either globally or locally, they do not produce constant variance over all values of the combinations of the regressor variables, unless one is dealing with rare occasions where the variability of the response variable does not change with the density of the data points or where the covariate variables have a joint distribution whose density compensates for the aforementioned variability of the response. This heteroscedasticity is unsettling and should be avoided if possible. In a series of papers, Kanazawa (1992), Kanazawa (1993a,b), and Kanazawa, Kogure, and Lee (1999) showed that in density estimation the AIC, Kullback-Leibler loss (KLL), and Hellinger distance (HD) are equivalent to the variance-stabilizing integrated squared error (VSISE) and that the histogram cell width selection rules or window width selection rules based on AIC, KLL, or HD are asymptotically equivalent. This means that the resulting histogram estimates or kernel density estimates based on AIC, KLL, or HD all try to stabilize the variance of the density estimates over the domain. Considering these developments in nonparametric density estimation, we feel that it is essential to have in the toolbox for data analysis a nonparametric regression estimator that stabilizes the variance when the variances in error terms vary with observation.

One of the well-known bandwidths for the *univariate* LL estimator is the local variable bandwidth that minimizes the leading term of the mean squared error (asymptotic MSE or

AMSE)

$$AMSE\left(m(\mathbf{x}),\ \widehat{m_{\mathbf{H}}^{LL}}(\mathbf{x})\right)$$
$$= \frac{1}{n|\mathbf{H}|}\frac{\sigma^2(\mathbf{x})}{f_{\mathbf{X}}(\mathbf{x})}R(K_{\mathbf{X}}) + \frac{\mu_2^2(K_{\mathbf{X}})}{4}\left[\text{trace}\left[\mathbf{H}^T\nabla^2 m(\mathbf{x})\mathbf{H}\right]\right]^2,$$

as given by Ruppert and Wand (1994) for a multivariate setting. The LL estimator with such a bandwidth is heteroscedastic. On the other hand, Nishida and Kanazawa (2011) propose a local variable bandwidth that stabilizes the variance of the univariate Nadaraya-Watson estimator (Nadaraya, 1964, 1965, 1970; Watson, 1964; Watson and Leadbetter, 1963) (henceforth variance-stabilizing bandwidth or VS bandwidth). For the LL estimator, the VS bandwith is given by

$$h_{VS}(x) = \frac{\sigma^2(x)}{f_X(x)} \cdot \left[\frac{R(K_X)}{\mu_2^2(K_X)\left[\int_I \frac{w_X(x)\sigma^8(x)\alpha^2(x)}{f_X^4(x)}dx\right]}\right]^{\frac{1}{5}} n^{-\frac{1}{5}},$$

where $\alpha(x)$ is the second derivative function of $m(x)$ and $w_X(x)$ is a univariate weight function defined on I. This bandwidth can be criticized on the grounds that the MSE-minimizing local variable bandwidth in a *univariate setting* will always outperform the VS bandwidth in terms of the asymptotic mean integrated squared error (henceforth AMISE)

$$AMISE\left(m(\cdot),\ \widehat{m_{\mathbf{H}}^{LL}}(\cdot)\right) = \int\cdots\int_{I^p} AMSE\left(m(\mathbf{x}),\ \widehat{m_{\mathbf{H}}^{LL}}(\mathbf{x})\right) w_{\mathbf{X}}(\mathbf{x})d\mathbf{x},$$

where $w_{\mathbf{X}}(\mathbf{x})$ is a multivariate weight function defined on I^p. The result is brought about by the fact that the univariate VS bandwidth is calculated so as to minimize MISE among the class of bandwidths that stabilize variance over all local points $\mathbf{x}$. This constrained bandwidth choice cannot achieve the minimum MSE at every local point and thus cannot achieve minimum MISE over the support. We will show in Proposition 2-(i) that this bandwidth generates a larger AMISE than the MSE-minimizing local variable bandwidth in a univariate setting.

In a *multivariate regression setting*, however, this assertion is not necessarily true. In other words, we are able to find a variance-stabilizing estimator whose MISE can outperform

the MISE of a multivariate extension of the MSE-minimizing local variable bandwidth. This is possible because in a multivariate regression setting, we can reduce the sum of the MSE inflated due to the constraint by distributing the inflated MSE among different directions of coordinate axes. To do so, we employ a set of locally varying parameters that adjust the bias obtained after the variance is stabilized, or we introduce the local variable bandwidth matrix that negates the variable part of the "unconditional" variance in (4) given by

$$\mathbf{H_{VS}}(\mathbf{x}) = h_0 \cdot \begin{pmatrix} \left[\frac{\sigma^2(\mathbf{x})}{f_{\mathbf{X}}(\mathbf{x})}\right]^{\eta_{11}(\mathbf{x})} & 0 & 0 & \dots & 0 \\ 0 & \left[\frac{\sigma^2(\mathbf{x})}{f_{\mathbf{X}}(\mathbf{x})}\right]^{\eta_{22}(\mathbf{x})} & 0 & \dots & 0 \\ \vdots & \vdots & \ddots & & \\ 0 & \dots & \dots & & \left[\frac{\sigma^2(\mathbf{x})}{f_{\mathbf{X}}(\mathbf{x})}\right]^{\eta_{pp}(\mathbf{x})} \end{pmatrix}, \tag{6}$$

where h_0 is a global parameter and $\eta_{ii}(\mathbf{x})$, $i = 1, ..., p$, are the local parameters, both to be estimated, satisfying

$$\sum_{i=1}^{p} \eta_{ii}(\mathbf{x}) = 1, \tag{7}$$

$$-\infty < \eta_{ii}(\mathbf{x}) < \infty. \tag{8}$$

Both the global parameter h_0 and the local parameters, $\eta_{ii}(\mathbf{x})$, $i = 1, ..., p$, can be determined so as to optimize AMISE. This optimized bandwidth matrix can outperform a *multivariate* extension of the MSE-minimizing local variable bandwidth,

$$\mathbf{H}_{var}(\mathbf{x}) \;=\; \left[\frac{R(K_{\mathbf{X}})\sigma^2(\mathbf{x})}{\mu_2^2(K_{\mathbf{X}}) f_{\mathbf{X}}(\mathbf{x}) \left[\sum_{i=1}^{p} \alpha_{ii}(\mathbf{x})\right]^2}\right]^{\frac{1}{p+4}} p^{\frac{1}{p+4}} \cdot n^{-\frac{1}{p+4}} \cdot \mathbf{I}_p, \tag{9}$$

which minimizes AMSE at every $\mathbf{x}$ among the class of local variable scalar bandwidth matrices,

$$\mathbf{H}_{var}(\mathbf{x}) = h_{00}(\mathbf{x}) \cdot \mathbf{I}_p$$

(henceforth the MSE-minimizing local variable scalar bandwidth matrix or the MSE-minimizing scalar bandwidth matrix). The proposed VS bandwidth matrix is given in Proposition 1 along with remarks.

One type of VS bandwidth matrix has a practical advantage over the MSE-minimizing scalar bandwidth matrix in (9) in that it avoids a discontinuity often encountered by (9): the denominator of the scalar in the MSE-minimizing scalar bandwidth matrix in (9) is zero at the points satisfying $\sum_{i=1}^{p} \alpha_{ii}(\mathbf{x}_*) = 0$. As a result, the diagonal elements in the bandwidth matrix takes infinitely large value and $\widehat{m^{LL}_{\mathbf{H_X}}}(\mathbf{x}_*)$ takes $\mathbf{e_1}\widehat{\boldsymbol{\beta}^{OLS}}$, the intercept term of the OLS estimator. Because $m(\mathbf{x}_*)$ does not coincide with $\mathbf{e_1}\widehat{\boldsymbol{\beta}^{OLS}}$ in general, this invites considerble bias at the corresponding points. However, the VS bandwidth matrix is continuous at these points as long as the local parameters $\eta_{ii}(\mathbf{x})$, $i = 1, ..., p$, are all constant and a standard assumption such as **S 2** is made. We explain this issue in Section 4.

In Section 2, we introduce the local variable VS bandwidth matrix that minimizes AMISE and show a sufficient condition that enables the proposed VS bandwidth matrix to outperform (9). In Section 3, we present the outline of an algorithm for estimating the VS bandwidth matrix and two simulation studies to evaluate the performance of our proposed estimator. In Section 4, we give Discussion. The detailed estimation algorithm is shown in Appendix 1. To assist the reader, we give a brief overview of all bandwidth matrices used throughout the paper in Appendix 2.

2 Introduction of the variance-stabilizing local variable bandwidth matrix

Proposition 1. *The theoretically variance-stabilizing local variable diagonal bandwidth matrix for the multivariate LL estimator,*

$$\mathbf{H}^*_{\mathbf{VS}}(\mathbf{x}) \;=\; h_0^* \cdot \mathrm{diag}\left(\left[\frac{\sigma^2(\mathbf{x})}{f_{\mathbf{X}}(\mathbf{x})}\right]^{\eta^*_{11}(\mathbf{x})}, ..., \left[\frac{\sigma^2(\mathbf{x})}{f_{\mathbf{X}}(\mathbf{x})}\right]^{\eta^*_{pp}(\mathbf{x})} \right), \tag{10}$$

which minimizes asymptotic MISE is given by the following optimized parameters h_0^* and $\eta^*_{ii}(\mathbf{x})$, $i = 1, ..., p$.

(i) *The optimal global parameter* h_0^* *is given by*

$$h_0^* = \left[\frac{R(K_{\mathbf{X}})}{\mu_2^2(K_{\mathbf{X}})T_{VS}(\eta_{11}^*(\mathbf{x}),...,\eta_{pp}^*(\mathbf{x}))}\right]^{\frac{1}{p+4}} \cdot p^{\frac{1}{p+4}} \cdot n^{-\frac{1}{p+4}}, \tag{11}$$

where

$$T_{VS}(\eta_{11}(\mathbf{x}),...,\eta_{pp}(\mathbf{x})) = \int \cdots \int_{I^p} w_{\mathbf{X}}(\mathbf{x}) \left[\sum_{i=1}^{p} \alpha_{ii}(\mathbf{x}) \left[\frac{\sigma^2(\mathbf{x})}{f_{\mathbf{X}}(\mathbf{x})}\right]^{2\eta_{ii}(\mathbf{x})}\right]^2 d\mathbf{x}. \tag{12}$$

(ii) *The optimal local parameters* $\eta_{ii}^*(\mathbf{x})$, $i = 1,...,p$, *are given by*

$$\eta_{ii}^*(\mathbf{x}) = \frac{\ln\left[\frac{\prod_{j=1}^{p}\alpha_{jj}(\mathbf{x})}{[\alpha_{ii}(\mathbf{x})]^p}\left[\frac{\sigma^2(\mathbf{x})}{f_{\mathbf{X}}(\mathbf{x})}\right]^2\right]}{\ln\left[\frac{\sigma^2(\mathbf{x})}{f_{\mathbf{X}}(\mathbf{x})}\right]^{2p}} \tag{13}$$

if $\alpha_{ii}(\mathbf{x}) > 0$, $i = 1,...,p$, *or* $\alpha_{ii}(\mathbf{x}) < 0$, $i = 1,...,p$.

Remark 1. If $\alpha_{ii}(\mathbf{x}) = 0, i = 1,...,p$, at some points in the domain, the criterion function presented later in (17) takes a zero minimum value for every $\eta_{ii}(\mathbf{x})$, $i = 1,...,p$. At the points, any set of values $\eta_{ii}^*(\mathbf{x})$, $i = 1,...,p$, satisfying (7) is available.

Remark 2. In general, if $\alpha_{ii}(\mathbf{x})$'s, $i = 1,...,p$, are not of the same sign, $\eta_{ii}^*(\mathbf{x})$'s, $i = 1,...,p$, are not uniquely determined when $p \geq 3$. In a special case where the function (18), presented later, does not have local maximum or minimum values at $\mathbf{x}$, the optimal set of parameters $\eta_{ii}^*(\mathbf{x})$, $i = 1,...,p$, is given by any set of values satisfying

$$\sum_{i=1}^{p} \alpha_{ii}(\mathbf{x}) \left[\frac{\sigma^2(\mathbf{x})}{f_{\mathbf{X}}(\mathbf{x})}\right]^{2\eta_{ii}(\mathbf{x})} = 0, \quad \text{subject to } \sum_{i=1}^{p} \eta_{ii}(\mathbf{x}) = 1.$$

Remark 3. If $\alpha_{qq}(\mathbf{x})$ is zero and the rest of the $\alpha_{ii}(\mathbf{x})$'s, $i = 1,...,p, i \neq q$, are non-zero, we consider the $p-1$-dimensional minimization problem of (17) with the q-th variable left out of the minimization problem.

Remark 4. If $f_{\mathbf{X}}(\mathbf{x})$ and $\sigma^2(\mathbf{x})$ are of the same functional form, we do not have to employ our proposed method because the variance (4) is already constant. If $\sigma^2(\mathbf{x})/f_{\mathbf{X}}(\mathbf{x}) = 1$ at some points in the domain, any set of values $\eta_{ii}(\mathbf{x})$, $i = 1,...,p$, satisfying (7) is available at the corresponding points.

Remark 5. The assumption **S 6**-(i)(ii) is necessary to guarantee the existence of h_0^* in (11), which requires that $T_{VS}(\eta_{11}(\mathbf{x}), ..., \eta_{pp}(\mathbf{x}))$ in (12) is not equal to zero.

Proof of (i). We first choose h_0, given $\eta_{ii}(\mathbf{x})$, $i = 1, ..., p$, to minimize AMISE. Integrating the square of the leading term in (5) and (4) over the support I^p, the leading term of the MISE between $\widehat{m^{LL}_{\mathbf{H}(\mathbf{x})}}(\mathbf{x})$ and $m(\mathbf{x})$ is

$$\int \cdots \int_{I^p} \left[\frac{1}{n|\mathbf{H}(\mathbf{x})|} \frac{\sigma^2(\mathbf{x})}{f_{\mathbf{X}}(\mathbf{x})} R(K_{\mathbf{X}}) + \frac{\mu_2^2(K_{\mathbf{X}})}{4} \left[\text{trace} \left[\mathbf{H}(\mathbf{x})^T \nabla^2 m(\mathbf{x}) \mathbf{H}(\mathbf{x}) \right] \right]^2 \right] w_{\mathbf{X}}(\mathbf{x}) d\mathbf{x}. \quad (14)$$

Substituting $\mathbf{H_{VS}}(\mathbf{x})$ in (6) for $\mathbf{H}(\mathbf{x})$ in (14), we obtain

$$\frac{1}{nh_0^p} R(K_{\mathbf{X}}) + \frac{h_0^4}{4} \mu_2^2(K_{\mathbf{X}}) T_{VS}(\eta_{11}(\mathbf{x}), ..., \eta_{pp}(\mathbf{x})). \quad (15)$$

The optimal global parameter (11) minimizes (15) with respect to h_0. □

Proof of (ii). We then optimize $\eta_{ii}(\mathbf{x})$, $i = 1, ..., p$, in terms of the AMISE. Substituting h_0^* in (11) for h_0 in (15), we obtain the AMISE, optimized in terms of h_0, written as

$$\left[\frac{[R(K_{\mathbf{X}})]^{\frac{4}{p+4}}}{[\mu_2^2(K_{\mathbf{X}})]^{-\frac{p}{p+4}} [T_{VS}(\eta_{11}(\mathbf{x}), ..., \eta_{pp}(\mathbf{x}))]^{-\frac{p}{p+4}}} \right] \left(p^{\frac{-p}{p+4}} + \frac{p^{\frac{4}{p+4}}}{4} \right) \cdot n^{-\frac{4}{p+4}}. \quad (16)$$

To minimize (16), the term $T_{VS}(\eta_{11}(\mathbf{x}), ..., \eta_{pp}(\mathbf{x}))$ defined in (12) must be minimized in terms of $\eta_{ii}(\mathbf{x})$ $i = 1, ..., p$. For such $\eta_{ii}(\mathbf{x})$, $i = 1, ..., p$, we solve the following constrained minimization problem in terms of $\eta_{ii}(\mathbf{x})$, $i = 1, ..., p$, at every point of $\mathbf{x}$,

$$\min_{\eta_{11}(\mathbf{x}), ..., \eta_{pp}(\mathbf{x})} \left[\sum_{i=1}^{p} \alpha_{ii}(\mathbf{x}) \left[\frac{\sigma^2(\mathbf{x})}{f_{\mathbf{X}}(\mathbf{x})} \right]^{2\eta_{ii}(\mathbf{x})} \right]^2, \text{ subject to } \sum_{i=1}^{p} \eta_{ii}(\mathbf{x}) = 1. \quad (17)$$

Let $G(\cdot)$ denote the $p-1$-variate function with respect to $\eta_{11}(\mathbf{x})$, ..., $\eta_{p-1p-1}(\mathbf{x})$,

$$G(\eta_{11}(\mathbf{x}), ..., \eta_{p-1p-1}(\mathbf{x})) = \left[\sum_{i=1}^{p-1} \alpha_{ii}(\mathbf{x}) \left[\frac{\sigma^2(\mathbf{x})}{f_{\mathbf{X}}(\mathbf{x})} \right]^{2\eta_{ii}(\mathbf{x})} \right] + \alpha_{pp}(\mathbf{x}) \left[\frac{\sigma^2(\mathbf{x})}{f_{\mathbf{X}}(\mathbf{x})} \right]^{2\left[1 - \sum_{i=1}^{p-1} \eta_{ii}(\mathbf{x})\right]}. \quad (18)$$

Differentiating (18) with respect to $\eta_{ii}(\mathbf{x})$, $i = 1, ..., p-1$, and equating the outcome to zero, we obtain the following simultaneous equations

$$\frac{\partial G(\eta_{11}(\mathbf{x}), ..., \eta_{p-1p-1}(\mathbf{x}))}{\partial \eta_{ii}(\mathbf{x})} = 2 \left[\alpha_{ii}(\mathbf{x}) \left[\frac{\sigma^2(\mathbf{x})}{f_{\mathbf{X}}(\mathbf{x})} \right]^{2\eta_{ii}(\mathbf{x})} - \alpha_{pp}(\mathbf{x}) \left[\frac{\sigma^2(\mathbf{x})}{f_{\mathbf{X}}(\mathbf{x})} \right]^{2\left[1 - \sum_{i=1}^{p-1} \eta_{ii}(\mathbf{x})\right]} \right] \ln \left[\frac{\sigma^2(\mathbf{x})}{f_{\mathbf{X}}(\mathbf{x})} \right]$$

$$= 0, \quad i = 1, ..., p-1. \tag{19}$$

Solving simultaneous equations (19) and (7) with respect to $\eta_{ii}(\mathbf{x})$, $i = 1, ..., p$, we obtain the following first-order condition,

$$\eta^*_{ii}(\mathbf{x}) = \frac{\ln \left[\frac{\prod_{j=1}^p \alpha_{jj}(\mathbf{x})}{[\alpha_{ii}(\mathbf{x})]^p} \left[\frac{\sigma^2(\mathbf{x})}{f_{\mathbf{X}}(\mathbf{x})} \right]^2 \right]}{\ln \left[\frac{\sigma^2(\mathbf{x})}{f_{\mathbf{X}}(\mathbf{x})} \right]^{2p}}, \quad i = 1, ..., p. \tag{20}$$

To check the second-order condition, we examine the principal minors of order $k = 1, .., p-1$,

$$\begin{aligned} \Delta_k(\mathbf{x}) &= \begin{vmatrix} A_{11}(\mathbf{x}) & A_{12}(\mathbf{x}) & \dots & A_{1k}(\mathbf{x}) \\ A_{21}(\mathbf{x}) & A_{22}(\mathbf{x}) & \dots & A_{2k}(\mathbf{x}) \\ \vdots & \vdots & \ddots & \vdots \\ A_{k1}(\mathbf{x}) & A_{k2}(\mathbf{x}) & \dots & A_{kk}(\mathbf{x}) \end{vmatrix} \\ &= \left[2 \ln \left[\frac{\sigma^2(\mathbf{x})}{f_{\mathbf{X}}(\mathbf{x})} \right] \right]^{2k} \sum_{j=1}^{k+1} \prod_{i=1, i \neq j}^{k+1} \left[\alpha_{ii}(\mathbf{x}) \left[\frac{\sigma^2(\mathbf{x})}{f_{\mathbf{X}}(\mathbf{x})} \right]^{2\eta^*_{ii}(\mathbf{x})} \right], \end{aligned} \tag{21}$$

where

$$A_{ij}(\mathbf{x}) = \left. \frac{\partial^2 G(\eta_{11}(\mathbf{x}), ..., \eta_{p-1p-1}(\mathbf{x}))}{\partial \eta_{ii}(\mathbf{x}) \partial \eta_{jj}(\mathbf{x})} \right|_{\eta_{11}(\mathbf{x}) = \eta^*_{11}(\mathbf{x}), ..., \eta_{p-1p-1}(\mathbf{x}) = \eta^*_{p-1p-1}(\mathbf{x})}, \; i = 1, ..., k, \; j = 1, ..., k.$$

If $\alpha_{ii}(\mathbf{x}) > 0$, $i = 1, ..., p$, the sequence of the principal minors (21) is $\Delta_1(\mathbf{x}) > 0$, $\Delta_2(\mathbf{x}) > 0$, ..., $\Delta_{p-1}(\mathbf{x}) > 0$. This means that the function in (18) takes a positive minimum value under the first-order condition (20). On the other hand, if $\alpha_{ii}(\mathbf{x}) < 0$, $i = 1, ..., p$, the sequence of the principal minors (21) is $\Delta_1(\mathbf{x}) < 0$, $\Delta_2(\mathbf{x}) > 0$, $\Delta_3(\mathbf{x}) < 0$,... This means that the criterion function in (18) takes a negative maximum value under the first-order condition (20). Because the criterion function in (17) is the square of the function in (18), the first-order condition (20) optimizes the minimization problem (17). □

Remark 6. Interpretations of the two parameters are as follows. The parameter h_0 plays a role in controlling the AMISE globally. See (15). As for $\eta_{ii}(\mathbf{x})$, $i = 1, ..., p$, this set of functions is intended to cancel out the variable part $\sigma^2(\mathbf{x})/f_{\mathbf{X}}(\mathbf{x})$ of the variance (4) and to reduce the AMSE locally and therefore the AMISE globally. Furthermore, the local parameters

$\eta_{ii}(\mathbf{x})$, $i = 1, ..., p$, serve to stabilize the variance at the expense of bias. Particularly when $0 \leq \eta_{ii}(\mathbf{x}) \leq 1$, $i = 1, ..., p$, the parameters $\eta_{ii}(\mathbf{x})$, $i = 1, ..., p$, can be interpreted as the fractional rate of the power of the squared bias that should be distributed to every coordinate axis, $x_1, ..., x_p$ from (12).

Remark 7. Suppose that $\eta_{ii}(\mathbf{x})$, $i = 1, ..., p$, do not depend on $\mathbf{x}$ such as $\eta_{11}, ..., \eta_{pp}$, and $\sum_{i=1}^{p} \eta_{ii} = 1$ at all points $\mathbf{x}$. These globally determined parameters can also achieve the purpose of canceling out the term $\sigma^2(\mathbf{x})/f_{\mathbf{X}}(\mathbf{x})$ in the leading term of (4). However, these globally determined η_{ii}'s cannot reduce the AMISE as much as the locally determined $\eta_{ii}(\mathbf{x})$'s. Furthermore, optimizing $\eta_{11}, ..., \eta_{pp}$ requires numerical calculation. One easy way to obtain $\eta_{11}, ..., \eta_{pp}$ is to set $\eta_{11} = ... = \eta_{pp} = 1/p$. This choice of local parameters avoids the discontinuity created by (9) as explained in Section 4.

We illustrate the theoretical strength of our proposed VS bandwidth matrix over the MSE-minimizing scalar bandwidth matrix through a proposition and an example. Let $\gamma(\mathbf{x})$ denote the ratio of the following two "density" functions,

$$\gamma(\mathbf{x}) = \frac{\sigma^2(\mathbf{x})}{\int \cdots \int_{I^p} \sigma^2(\mathbf{x}) d\mathbf{x}} \Big/ \frac{f_{\mathbf{X}}(\mathbf{x})}{\int \cdots \int_{I^p} f_{\mathbf{X}}(\mathbf{x}) d\mathbf{x}} \, . \tag{22}$$

When the VS bandwidth matrix and the MSE-minimizing scalar bandwidth matrix are employed, the respective AMISE's are written as

$$AMISE\left(m(\cdot),\ \widehat{m^{LL}_{\mathbf{H_{VS}}}}(\cdot)\right) = C_1 \cdot n^{-\frac{4}{p+4}} \cdot \left[\int \cdots \int_{I^p} w_{\mathbf{X}}(\mathbf{x}) \gamma^{\frac{4}{p}}(\mathbf{x}) \left[p \prod_{i=1}^{p} |\alpha_{ii}(\mathbf{x})|^{\frac{1}{p}} \right]^2 d\mathbf{x} \right]^{\frac{p}{p+4}}, \tag{23}$$

$$AMISE\left(m(\cdot),\ \widehat{m^{LL}_{\mathbf{H}_{var}}}(\cdot)\right) = C_1 \cdot n^{-\frac{4}{p+4}} \cdot \int \cdots \int_{I^p} w_{\mathbf{X}}(\mathbf{x}) \left[\gamma^{\frac{4}{p}}(\mathbf{x}) \left[\sum_{i=1}^{p} \alpha_{ii}(\mathbf{x}) \right]^2 \right]^{\frac{p}{p+4}} d\mathbf{x}, \tag{24}$$

where $C_1 = \left(p^{-p/(p+4)} + p^{4/(p+4)}/4\right) \left[R(K_{\mathbf{X}})\right]^{4/(p+4)} \left[\mu_2^2(K_{\mathbf{X}})\right]^{p/(p+4)} \left[\int \cdots \int_{I^p} \sigma^2(\mathbf{x}) d\mathbf{x}\right]^{4/(p+4)} > 0$. We obtain the following proposition as to the magnitude relationship in terms of the AMISE between (23) and (24).

Proposition 2. *Suppose that* $\alpha_{ii}(\mathbf{x}) > 0$, $i = 1, ..., p$, *or* $\alpha_{ii}(\mathbf{x}) < 0$, $i = 1, ..., p$, *holds at every* $\mathbf{x}$. *The magnitude relationship of the AMISE between the VS bandwidth matrix in* (10) *and the MSE-minimizing scalar bandwidth matrix in* (9) *is then determined as follows.*
(i) *When* $p = 1$, *the* $AMISE\left(m(\cdot),\ \widehat{m^{LL}_{h_{VS}}}(\cdot)\right)$ *is always larger than the* $AMISE\left(m(\cdot),\ \widehat{m^{LL}_{h_{var}}}(\cdot)\right)$.
(ii) *When* $p > 1$, *a sufficient condition for which the* $AMISE\left(m(\cdot),\ \widehat{m^{LL}_{\mathbf{H_{VS}}}}(\cdot)\right)$ *is smaller than the* $AMISE\left(m(\cdot),\ \widehat{m^{LL}_{\mathbf{H}_{var}}}(\cdot)\right)$ *is*

$$\gamma^{\frac{4}{p}}(\mathbf{x})\left[\sum_{i=1}^{p}\alpha_{ii}(\mathbf{x})\right]^2 = C, \quad \text{at every } \mathbf{x}, \text{ where } C > 0 \text{ is any positive constant.} \tag{25}$$

Proof of (i). When $p = 1$, by Hölder's inequality, we obtain

$$\begin{aligned} &AMISE\left(m(\cdot),\ \widehat{m^{LL}_{h_{var}}}(\cdot)\right) - AMISE\left(m(\cdot),\ \widehat{m^{LL}_{h_{VS}}}(\cdot)\right) \\ &= \frac{5}{4}\cdot n^{-\frac{4}{5}}\left[R(K_X)\right]^{\frac{4}{5}}\left[\mu_2^2(K_X)\right]^{\frac{1}{5}}\left[\int_I \sigma^2(x)dx\right]^{\frac{4}{5}} \\ &\quad\times\left[\int_I w_X^{\frac{4}{5}}(x)\left[w_X(x)\alpha^2(x)\gamma^4(x)\right]^{\frac{1}{5}}dx - \left[\int_I w_X(x)dx\right]^{\frac{4}{5}}\left[\int_I w_X(x)\alpha^2(x)\gamma^4(x)dx\right]^{\frac{1}{5}}\right] \le 0. \end{aligned}$$

□

Proof of (ii). When $p > 1$, if we employ (25), we obtain the following relation

$$\begin{aligned} &AMISE^{\frac{p+4}{p}}\left(m(\cdot),\ \widehat{m^{LL}_{\mathbf{H}_{var}}}(\cdot)\right) - AMISE^{\frac{p+4}{p}}\left(m(\cdot),\ \widehat{m^{LL}_{\mathbf{H_{VS}}}}(\cdot)\right) \\ &= \left[C_1\right]^{\frac{p+4}{p}}\cdot C\cdot n^{-\frac{4}{p}}\cdot\int\cdots\int_{I^p} w_{\mathbf{X}}(\mathbf{x})\left[\frac{\left[\sum_{i=1}^{p}\alpha_{ii}(\mathbf{x})\right]^2 - \left[p\prod_{i=1}^{p}\left|\alpha_{ii}(\mathbf{x})\right|^{\frac{1}{p}}\right]^2}{\left[\sum_{i=1}^{p}\alpha_{ii}(\mathbf{x})\right]^2}\right]d\mathbf{x}. \end{aligned} \tag{26}$$

Because $\left[\sum_{i=1}^{p}\alpha_{ii}(\mathbf{x})\right]^2 - \left[p\prod_{i=1}^{p}\left|\alpha_{ii}(\mathbf{x})\right|^{1/p}\right]^2 \ge 0$ always holds at every $\mathbf{x}$, equation (26) is always greater than or equal to zero under the sufficient condition (25). □

Example 1. Let a twice differentiable bivariate regression curve be $m(x_1, x_2) = \cos(3x_1) + \cos(3x_2)$, with covariates distributed as $\tau(x_1, x_2)$, the bivariate normal density having its mean $(0, 0)$ and the variance-covariance matrix $\mathrm{diag}(0.35^2, 0.35^2)$ truncated on the domain $[-0.5, 0.5]\times[-0.5, 0.5]$. In this setting, 71.71% of the data points distributed as

$N((0,0)^T, \text{diag}(0.35^2, 0.35^2))$ are included in the domain $[-0.5, 0.5] \times [-0.5, 0.5]$. The conditional variance $\sigma^2(x_1, x_2)$ is determined by (25) and is $\tau(x_1, x_2)/(9\cos(3x_1) + 9\cos(3x_2))$. See Figure 1. For this example, the $AMISE\left(m(\cdot),\ \widehat{m^{LL}_{\mathbf{H_{VS}}}}(\cdot)\right)$ is 3.2982 % smaller than the $AMISE\left(m(\cdot),\ \widehat{m^{LL}_{\mathbf{H}_{var}}}(\cdot)\right)$.

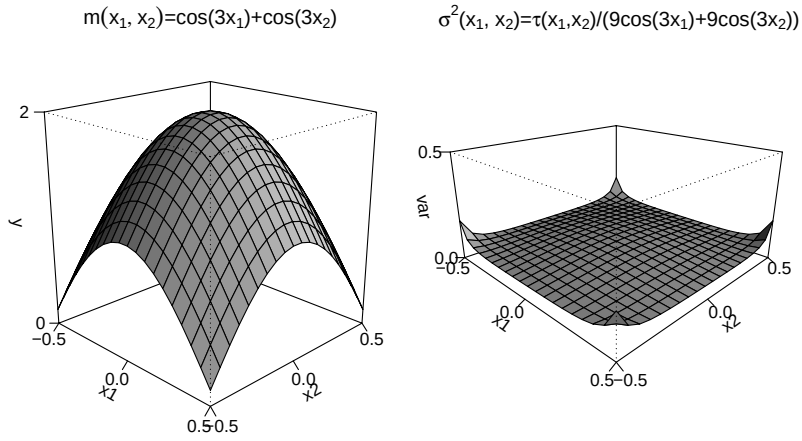

Figure 1: The true regression function $m(x_1, x_1) = \cos(3x_1) + \cos(3x_2)$ on the left and the conditional variance function $\sigma^2(x_1, x_2) = \tau(x_1, x_2)/(9\cos(3x_1) + 9\cos(3x_2))$ on the right in Example 1, where $\tau(x_1, x_2)$ is the bivariate normal density $N((0,0), \text{diag}(0.35^2, 0.35^2))$ truncated on $[-0.5, 0.5] \times [-0.5, 0.5]$.

Remark 8. Proposition 2-(ii) results from the fact that the VS bandwidth matrix has more flexibility in its matrix form than the MSE-minimizing scalar bandwidth matrix in (9). The p-variate VS bandwidth matrix has $p-1$ local parameters at a given point $\mathbf{x}$ and one global parameter, while the MSE-minimizing scalar bandwidth matrix has one local parameter at the same given point $\mathbf{x}$. However, in a univariate setting, the VS bandwidth matrix is reduced to having one global parameter, while the MSE-minimizing scalar bandwidth matrix remains to have one local parameter at a given point $\mathbf{x}$. As a result, the VS bandwidth matrix will not be able to outperform the MSE-minimizing scalar bandwidth matrix by definition as in Proposition 2-(i).

3 Estimating the variance-stabilizing bandwidth matrix

To estimate the VS bandwidth matrix, the global parameter h_0^* in (11), the local parameters $\eta_{ii}^*(\mathbf{x})$, $i = 1, ..., p$ in (13), $\sigma^2(\mathbf{x})$ and $f_{\mathbf{X}}(\mathbf{x})$ must be estimated. The basic idea is to individually estimate components in (11) and (13), $\widehat{f_{\mathbf{X}}}(\mathbf{x})$, $\widehat{\alpha_{ii}}(\mathbf{x})$, $i = 1, ..., p$, $\widehat{\sigma^2}(\mathbf{x})$ and plug these estimators into (11) and (13). In the course of sequentially estimating necessary components, we also employ well-established ideas such as "pilot estimator," "cross-validation." This idea guarantees weak consistency of the LL estimator with the VS bandwidth matrix, while simultaneously achieving homoscedasticity of $\widehat{m^{LL}_{\widehat{\mathbf{H}_{\mathbf{VS}}}}}(\mathbf{x})$, as long as the components $\widehat{f_{\mathbf{X}}}(\mathbf{x})$ and $\widehat{\sigma^2}(\mathbf{x})$ in (10) are respectively weakly consistent estimators.

We present an illustrative example of the plug-in algorithm in Appendix 1 for bivariate $f_{X_1,X_2}(x_1, x_2)$. First, we estimate $f_{X_1,X_2}(x_1, x_2)$. Second, we estimate the second derivative of $m(x_1, x_2)$, $\alpha_{11}(x_1, x_2)$, $\alpha_{22}(x_1, x_2)$. Third, we estimate $\sigma^2(x_1, x_2)$. Fourth, we estimate h_0. Fifth, we estimate $\mathbf{H_{VS}}(\mathbf{x})$.

In estimating $\alpha_{11}(x_1, x_2)$ and $\alpha_{22}(x_1, x_2)$, we use the quartic polynomial "pilot" estimator of $m(x_1, x_2)$, proposed by Fan and Gijbels (1995) to allow for flexibility in estimating the second derivative of $m(\mathbf{x})$ in (3). To estimate $\sigma^2(x_1, x_2)$, we employ the "residual-based" estimator in Fan and Yao (1998). This estimator smoothes squared residuals $(Y_i - \widehat{m}(\mathbf{x}))^2$ by the Nadaraya-Watson regression estimator. To calculate the squared residuals, we estimate $m(x_1, x_2)$ by the LL estimator with its bandwidth estimated so as to minimize cross-validation statistics among the class of global scalar bandwidth matrices that appear later in (33). The bivariate extension of the residual-based estimator appears in (34), and we compute the bandwidth of the residual-based estimator so as to minimize cross-validation statistics among the class of global scalar bandwidth matrices in (35).

Simulation studies

We present two simulation studies. The first simulation does not satisfy condition (25), whereas the second does. In both simulation cases, we repeat the process $M = 200$ times

at points from -0.495 to 0.495 at 0.01 increments in both directions for $n = 500,\ 1{,}000,\ 5{,}000,\ 10{,}000$ and $15{,}000$. In the simulations, we would first like to see if the proposed algorithm obtains $\widehat{h}_0^*$ close to h_0^* in (11) and if the estimator converges to the true regression function. We would also like to see if the proposed estimator of the VS bandwidth matrix in (36) actually stabilizes the variances of the LL estimator. We also evaluate our proposed estimator of the VS bandwidth matrix relative to the theoretically MSE-minimizing scalar bandwidth matrix in (9). We choose two simulation settings that guarantee that no points of $\mathbf{x}$ satisfy $\sum_{i=1}^{p} \alpha_{ii}(\mathbf{x}) = 0$ in their respective domains and that the MSE-minimizing scalar bandwidth matrix in (9) does not produce discontinuous points as mentioned in Section 1. As a kernel, we employ a bivariate Gaussian. As a weighting function $w_{\mathbf{X}}(\mathbf{x})$, we employ $\widehat{f_{\mathbf{X}}}(\mathbf{x})$.

Simulation 1. The first simulation setting is as follows. Let $I \times I$ denote $[-0.5, 0.5] \times [-0.5, 0.5]$. The density function $f_{X_1,X_2}(x_1, x_2)$ is a bivariate normal $N((0,0)^T, \mathrm{diag}(0.25^2, 0.25^2))$ truncated on the compact domain $[-0.5, 0.5] \times [-0.5, 0.5]$. In this setting, 91.1% of the data points distributed as $N((0,0)^T, \mathrm{diag}(0.25^2, 0.25^2))$ are included in the domain $[-0.5, 0.5] \times [-0.5, 0.5]$. The true regression function and the conditional variance function, respectively, are set to $m(x_1, x_2) = 1 - x_1^2 - x_2^2$, as in the left panel of Figure 2, and $\sigma^2(x_1, x_2) = 0.25 + 0.5x_1^2 + 0.5x_2^2$, as in the right panel. In this setup, the variance measured in terms of (4) becomes large near the boundaries. We intend that the curvatures of the true regression function are constant and can be easily estimated in this setup.

Table 1 shows the results of Simulation 1. In the table, we present means and standard deviations and the ratios $\widehat{h}_0^*/h_0^*$ of $M = 200$ simulated $\widehat{h}_0^*$ for $n = 500,\ 1{,}000,\ 5{,}000,\ 10{,}000$ and $15{,}000$. These numbers show that the estimator $\widehat{h}_0^*$ converges to h_0^* and is stable.

For the two bandwidth matrices, we also present the estimates of the MISE defined by

$$\begin{aligned} \widehat{MISE_{\widehat{\mathbf{H}}}} &= \widehat{MISE}\left(m(x_1, x_2),\ \widehat{m_{\widehat{\mathbf{H}}}^{LL}}(x_1, x_2)\right) \\ &= \frac{1}{M} \sum_{\mathbf{T}=1}^{M} \left[\int \int_{I^2} f_{X_1,X_2}(x_1, x_2) \left[m(x_1, x_2) - \widehat{m_{\widehat{\mathbf{H}}}^{LL(\mathbf{T})}}(x_1, x_2) \right]^2 dx_1 dx_2 \right], \quad (27) \end{aligned}$$

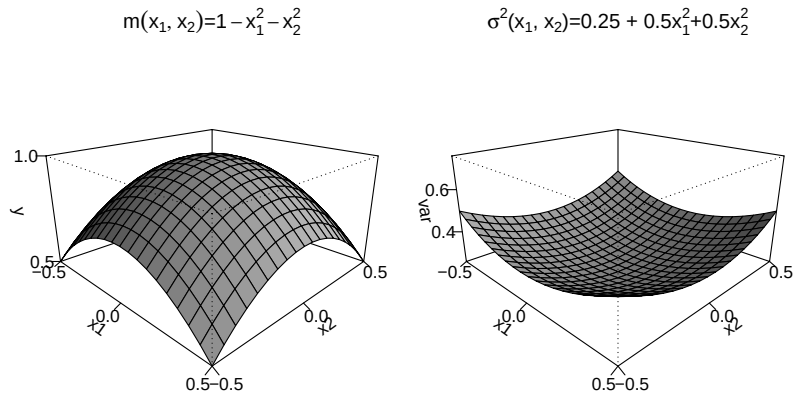

Figure 2: The true regression function on the left and the true conditional variance function on the right in Simulation 1.

where $\widehat{m^{LL}_{\widehat{\mathbf{H}}}}^{(\mathbf{T})}(x_1, x_2)$ is the LL estimator calculated ($\mathbf{T}$)-th generated sample of size n, and the two ratios $\widehat{MISE_{\widehat{\mathbf{H_{VS}}}}}/\widehat{MISE_{\mathbf{H}_{var}}}$ and $\widehat{MISE_{\mathbf{H_{VS}}}}/\widehat{MISE_{\mathbf{H}_{var}}}$, respectively, are denoted by Ratio 1 and Ratio 2. From these numbers for the $\widehat{MISE_{\widehat{\mathbf{H_{VS}}}}}$, we see that the $\widehat{MISE_{\widehat{\mathbf{H_{VS}}}}}$'s approach to zero as n increases and thus, $\widehat{MSE_{\widehat{\mathbf{H_{VS}}}}}$'s approach zero as n increases, except for the countable number of points (x_1, x_2). Therefore, pointwise convergence of $\widehat{m^{LL}_{\widehat{\mathbf{H_{VS}}}}}(x_1, x_2)$ to $m(x_1, x_2)$ in the sense of the mean square and thus weak consistency of $\widehat{m^{LL}_{\widehat{\mathbf{H_{VS}}}}}(x_1, x_2)$ to $m(x_1, x_2)$ are supported by the simulation results. In addition, the Ratio 1's in Table 1 show that the price of homoscedasticity of the estimate decreases considerably as the sample size n increases.

To see if the variance is stabilized by the proposed VS bandwidth matrix, we present in Figure 3 boxplots of sample variances of $\widehat{m^{LL}_{\widehat{\mathbf{H_{VS}}}}}^{(\mathbf{T})}(x_1, x_2)$ and $\widehat{m^{LL}_{\mathbf{H}_{var}}}^{(\mathbf{T})}(x_1, x_2)$, $\mathbf{T} = 1, ..., M$, at 0.05 intervals on the x_1 axis for sample sizes $n = 500$, $1{,}000$, $5{,}000$, $10{,}000$ and $15{,}000$. The two horizontally aligned panels for the same sample size in Figure 3 share the same scale in terms of the y axis, but the scale of the y axis is shrunk from top to bottom. Because $m(x_1, x_2) = 1 - x_1^2 - x_2^2$ is exchangeable with x_1 and x_2 and so are $\sigma^2(x_1, x_2) = 0.25 + 0.5x_1^2 + 0.5x_2^2$, we only plot how the variance is stabilized along the x_1 axis. From Figure 3, we see that comparatively smaller variances are achieved by the estimator of the VS

bandwidth matrix near the boundaries with sample sizes greater than $1,000$, which suggests that the estimator of the VS bandwidth matrix stabilizes the variance of the LL estimator when the sample size is large. Table 3 summarizes Figure 3.

We see from Table 3 that both the sample means and the standard deviations of the sample variances under both the VS and MSE bandwidth matrices diminish as the sample size increases. We also see that the estimator of the VS bandwidth matrix achieves smaller standard deviations of the sample variance relative to the theoretically MSE-minimizing scalar bandwidth matrix when the sample size is $1,000$. When the sample size is greater than $5,000$, smaller sample means and standard deviations of the sample variance are achieved by the estimator of the VS bandwidth matrix relative to the theoretically MSE-minimizing scalar bandwidth matrix.

Simulation 2. For the second simulation study, we employ the setting in Example 1. In this setup, the curvature of the true regression function varies across the domain. We expect that estimation for Simulation 2 is more "difficult" than for Simulation 1 for this reason. We also expect that the estimator of the VS bandwidth matrix can outperform the theoretically MSE-minimizing scalar bandwidth matrix in terms of (27).

In running simulation 2, we observe that the estimator for $\sigma^2(\mathbf{x})$ is severely affected by boundary effects. To reduce the effects in $\widehat{\sigma^2}(\mathbf{x})$, we employ a weight function $w_{\sigma^2}(\mathbf{x})$ for $\sigma^2(\mathbf{x})$ and replace $\widehat{\sigma^2}(\mathbf{x})$ with $\widehat{\sigma^2}(\mathbf{x})w_{\sigma^2}(\mathbf{x})$ in the estimation of h_0^*. As a weight function $w_{\sigma^2}(\mathbf{x})$, we employ the bivariate normal density $N((0,0), \mathrm{diag}(0.15, 0.15))$.

Table 2 shows the results of Simulation 2. In the table, means and standard deviations of $\widehat{h}_0^*$, the ratios $\widehat{h}_0^*/h_0^*$, the estimates of the MISE for the two bandwidth matrices, and the Ratio 1's and the Ratio 2's are presented as in Simulation 1. The diminishing size of standard deviation of $\widehat{h}_0^*$ shows that $\widehat{h}_0^*$ converges to h_0^* and is stable as well. From the diminishing size of $\widehat{MISE}_{\widehat{\mathbf{H}}_{\mathbf{VS}}}$, pointwise convergence in the sense of mean square and thus weak consistency of $\widehat{m^{LL}_{\widehat{\mathbf{H}}_{\mathbf{VS}}}}(x_1, x_2)$ to $m(x_1, x_2)$ are confirmed as well. As for the Ratio 1's as well as the Ratio 2's in Table 2, they reach values less than unity, as we expected in Proposition 2-(ii).

To check if the variance is stabilized, we present boxplots of sample variances of $\widehat{m^{LL}_{\widehat{\mathbf{H}}_{\mathbf{VS}}}}^{(\mathbf{T})}(x_1, x_2)$ and $\widehat{m^{LL}_{\mathbf{H}_{var}}}^{(\mathbf{T})}(x_1, x_2)$, $\mathbf{T} = 1, ..., M$, in Figure 4 and their summary in Table 4. From the two horizontally aligned panels in Figure 4, we see that comparatively smaller variances are achieved by the estimator of the VS bandwidth matrix near the center of the domain when the sample size is greater than $1,000$. The numbers in Table 4 show that comparatively smaller sample means and standard deviations of the sample variances are achieved by the estimator of the VS bandwidth matrix when the sample size is greater than $1,000$.

n	$\widehat{h_0^*}$		$\widehat{h_0^*}/h_0^*$	$\widehat{MISE_{\widehat{\mathbf{H}}_{\mathbf{VS}}}}$	$\widehat{MISE_{\mathbf{H}_{var}}}$	Ratio 1	Ratio 2
	mean	std.dev.	mean				
500	0.2233	0.0441	1.1100	$2.3752 \cdot 10^{-2}$	$6.5010 \cdot 10^{-3}$	3.6535	1.1312
1,000	0.1980	0.0254	1.1045	$5.8148 \cdot 10^{-3}$	$3.9022 \cdot 10^{-3}$	1.4901	1.1482
5,000	0.1494	0.0059	1.0901	$1.4316 \cdot 10^{-3}$	$1.2244 \cdot 10^{-3}$	1.1692	1.1321
10,000	0.1327	0.0036	1.0873	$8.4876 \cdot 10^{-4}$	$7.5256 \cdot 10^{-4}$	1.1278	1.1672
15,000	0.1239	0.0024	1.0863	$6.1867 \cdot 10^{-4}$	$5.5025 \cdot 10^{-4}$	1.1243	1.1668

Table 1: Results of Simulation 1: Estimation of h_0^*, $\widehat{MISE_{\widehat{\mathbf{H}}_{\mathbf{VS}}}}$ and $\widehat{MISE_{\mathbf{H}_{var}}}$. We also calculate the Ratio 1 $= \widehat{MISE_{\widehat{\mathbf{H}}_{\mathbf{VS}}}}/\widehat{MISE_{\mathbf{H}_{var}}}$ and the Ratio 2 $= \widehat{MISE_{\mathbf{H}_{\mathbf{VS}}}}/\widehat{MISE_{\mathbf{H}_{var}}}$.

n	$\widehat{h_0^*}$		$\widehat{h_0^*}/h_0^*$	$\widehat{MISE_{\widehat{\mathbf{H}}_{\mathbf{VS}}}}$	$\widehat{MISE_{\mathbf{H}_{var}}}$	Ratio 1	Ratio 2
	mean	std.dev.	mean				
500	0.2797	0.0092	1.0670	$5.0603 \cdot 10^{-3}$	$4.8577 \cdot 10^{-3}$	1.0417	0.9975
1,000	0.2477	0.0051	1.0607	$2.9405 \cdot 10^{-3}$	$2.9934 \cdot 10^{-3}$	0.9823	0.9979
5,000	0.1888	0.0015	1.0573	$8.8754 \cdot 10^{-4}$	$9.2046 \cdot 10^{-4}$	0.9642	0.9985
10,000	0.1682	0.0009	1.0570	$5.3203 \cdot 10^{-4}$	$5.5444 \cdot 10^{-4}$	0.9595	0.9986
15,000	0.1571	0.0006	1.0566	$4.0719 \cdot 10^{-4}$	$4.2249 \cdot 10^{-4}$	0.9637	0.9987

Table 2: Results of Simulation 2 : Estimation of h_0^*, $\widehat{MISE_{\widehat{\mathbf{H}}_{\mathbf{VS}}}}$ and $\widehat{MISE_{\mathbf{H}_{var}}}$. We also calculate the Ratio 1 $= \widehat{MISE_{\widehat{\mathbf{H}}_{\mathbf{VS}}}}/\widehat{MISE_{\mathbf{H}_{var}}}$ and the Ratio 2 $= \widehat{MISE_{\mathbf{H}_{\mathbf{VS}}}}/\widehat{MISE_{\mathbf{H}_{var}}}$.

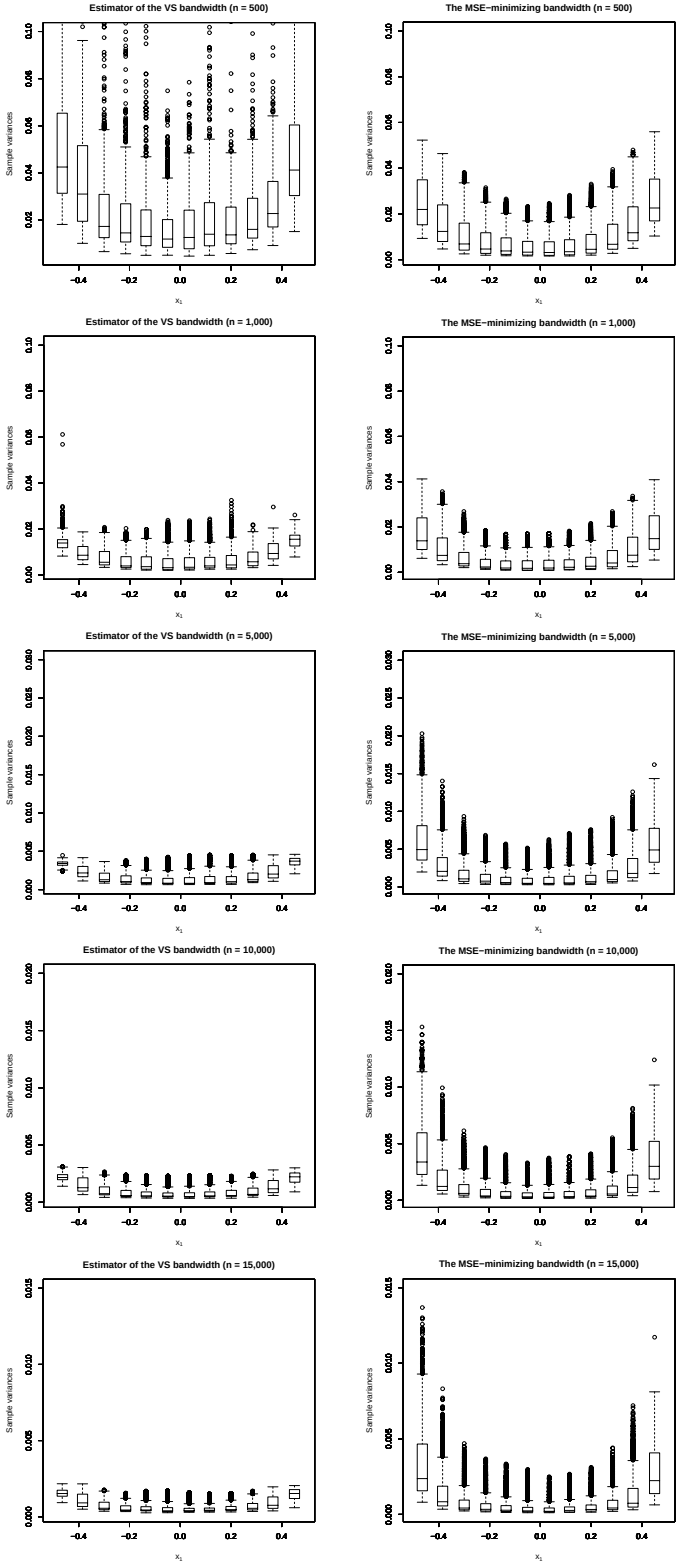

Figure 3: Results of Simulation 1: Distributions of sample variance at different points.

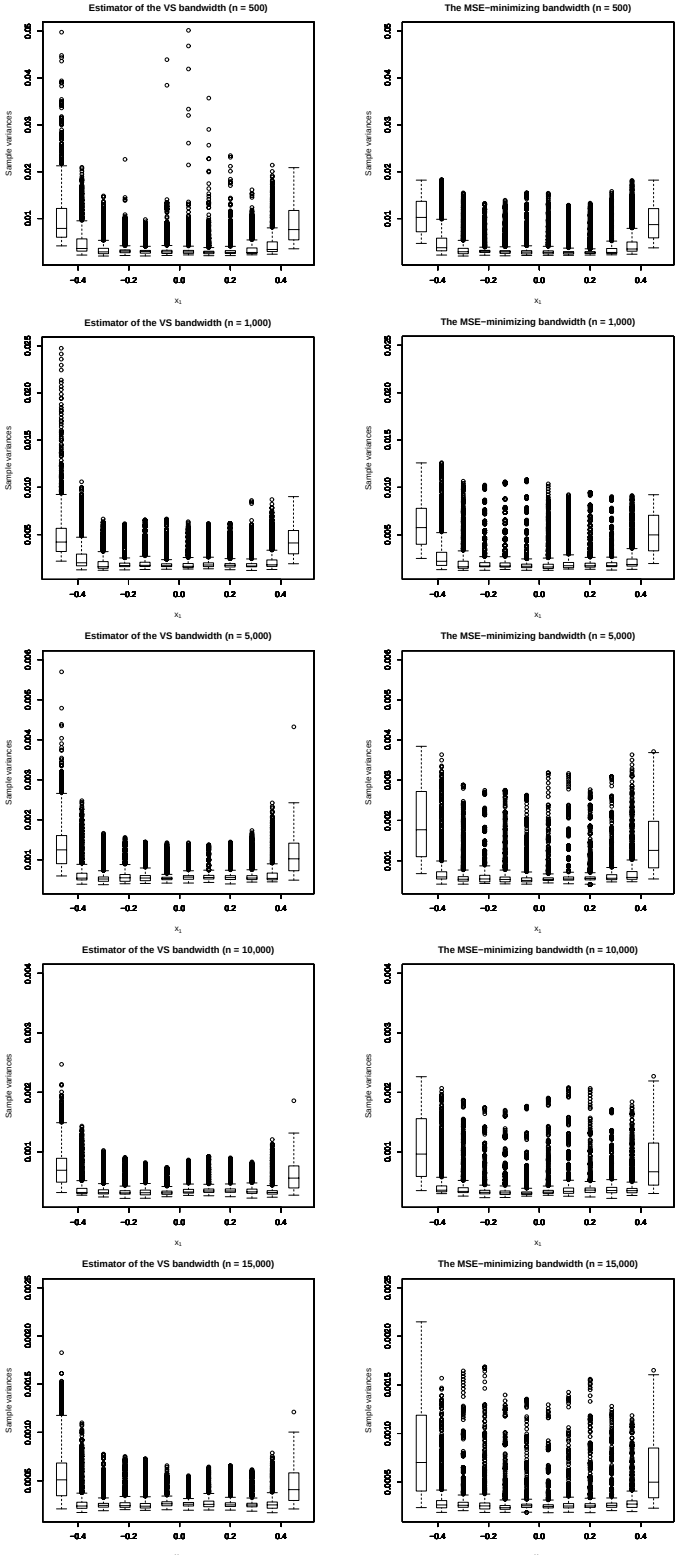

Figure 4: Results of Simulation 2: Distributions of sample variance at different points.

Mean and std.dev. of sample variances of the LL at 100×100 points ($M = 200$)						
n	Estimator of the VS bandwidth (36)			The MSE-minimizing bandwidth (9)		
	mean	std.dev.	coef.var.	mean	std.dev.	coef.var.
500	$4.7300 \cdot 10^{-2}$	$1.4231 \cdot 10^{-1}$	3.0087	$1.2400 \cdot 10^{-2}$	$1.1266 \cdot 10^{-2}$	0.9085
1,000	$8.6363 \cdot 10^{-3}$	$5.9634 \cdot 10^{-3}$	0.6905	$7.3632 \cdot 10^{-3}$	$6.7093 \cdot 10^{-3}$	0.9111
5,000	$1.9783 \cdot 10^{-3}$	$1.1709 \cdot 10^{-3}$	0.5918	$2.4565 \cdot 10^{-3}$	$2.8193 \cdot 10^{-3}$	1.1476
10,000	$1.1826 \cdot 10^{-3}$	$7.4524 \cdot 10^{-4}$	0.6301	$1.5716 \cdot 10^{-3}$	$2.0150 \cdot 10^{-3}$	1.2820
15,000	$8.3596 \cdot 10^{-4}$	$5.0301 \cdot 10^{-4}$	0.6017	$1.2015 \cdot 10^{-3}$	$1.6795 \cdot 10^{-3}$	1.3978

Table 3: Results of Simulation 1: Summary of Figure 3 results. The variance of the proposed estimator is stabilized as n increases. In addition, it performs better than the MSE-minimizing scalar bandwidth matrix when $n \geq 1,000$.

Mean and std.dev. of sample variances of the LL at 100×100 points ($M = 200$)						
n	Estimator of the VS bandwidth (36)			The MSE-minimizing bandwidth (9)		
	mean	std.dev.	coef.var.	mean	std.dev.	coef.var.
500	$5.0099 \cdot 10^{-3}$	$4.7212 \cdot 10^{-3}$	0.9423	$4.9717 \cdot 10^{-3}$	$3.6874 \cdot 10^{-3}$	0.7416
1,000	$2.7006 \cdot 10^{-3}$	$2.1761 \cdot 10^{-3}$	0.8057	$2.9425 \cdot 10^{-3}$	$2.1800 \cdot 10^{-3}$	0.7408
5,000	$7.3644 \cdot 10^{-4}$	$4.7057 \cdot 10^{-4}$	0.6389	$8.6432 \cdot 10^{-4}$	$6.6812 \cdot 10^{-4}$	0.7730
10,000	$4.2390 \cdot 10^{-4}$	$2.3318 \cdot 10^{-4}$	0.5500	$5.0773 \cdot 10^{-4}$	$3.8363 \cdot 10^{-4}$	0.7555
15,000	$3.1847 \cdot 10^{-4}$	$1.8108 \cdot 10^{-4}$	0.5686	$3.8035 \cdot 10^{-4}$	$2.9317 \cdot 10^{-4}$	0.7707

Table 4: Results of Simulation 2: Summary of Figure 4 results. The variance of the proposed estimator is stabilized as n increases. In addition, it performs better than the MSE-minimizing scalar bandwidth matrix when $n \geq 1,000$.

4 Discussion

In this paper, we propose an optimal multivariate variance-stabilizing (VS) bandwidth in Proposition 1 by combining two components, the globally determined h_0^* in (11) and a variable component that can be determined in principle for every $\mathbf{x}$ in the domain from (13), in a manner described in (6). Fan and Gijbels call this class of variable bandwidth a *local* variable bandwidth (See Fan and Gijbels 1992, p.2024). In Proposition 2, we give a sufficient condition in (25) under which our proposed VS bandwidth matrix theoretically outperforms the MSE-minimizing scalar bandwidth matrix. This proposition reveals that our VS bandwidth matrix can outperform the MSE-minimizing scalar bandwidth matrix (9) in terms of the AMISE.

In Section 3 and Appendix 1, respectively, we present the concept and the corresponding algorithm for estimating the VS bandwidth matrix. In Tables 1 and 2, we present simulation studies to show that the global parameter h_0^* is successfully estimated. The results in Figures 3 and 4 and in Tables 3 and 4 also show that, under the proposed VS bandwidth matrix selection algorithm, the variance of the LL estimator is stabilized as the sample size increases in comparison with the theoretically MSE-minimizing scalar bandwidth matrix.

Another way of introducing the idea of *variable* bandwidth is to combine two components, the globally determined $h_{n,\text{opt}}$ in (2.10) and a variable component defined only at the data points X_j, as in (2.9) in Fan and Gijbels (1992, p. 2013). They call this bandwidth a *global* variable bandwidth. Although their final recommended choice for the variable part is different from ours, they nonetheless briefly entertain the possibility of employing the variance-stabilized expression in the form of $\sigma^2(X_j)/f_X(X_j)$ (Fan and Gijbels, 1992, p.2014) in a *univariate* setting. These two ways of defining variable bandwidth, one as a *local* variable bandwidth and the other as a *global* variable bandwidth, seem different in concept but are not different in the context of asymptotic variance stabilization. This is because, if the variance of the estimated regression function is asymptotically stabilized at every $\mathbf{x}$, it is also asymptotically stabilized at the data points X_j. Conversely, if the variance is stabilized at the data points X_j where these data points are not restricted, it is also asymptotically

stabilized everywhere in the domain. These similar but subtly different ways of defining variable bandwidth give rise to two regression function estimates that are essentially the same, although the *local* variable bandwidth offers somewhat more accuracy but increased computational burden, while the global variable bandwidth offers somewhat less accuracy but reduced computational burden.

It can be argued that another type of MSE-minimizing local variable bandwidth matrix that minimizes the AMSE among the class of diagonal bandwidth matrices,

$$\mathbf{H}_{var+}(\mathbf{x}) = \operatorname{diag}\left(h_{11}(\mathbf{x}), ..., h_{pp}(\mathbf{x})\right), \tag{28}$$

should be compared with (10). This bandwidth matrix is a local counterpart to the class of bandwidth matrices proposed by Yang and Tschernig (1999) for the multivariate LL estimator. However, we feel that the class of VS bandwidth matrices that ought to be compared with (28) is the one that minimizes the AMISE among the class of

$$\mathbf{H}_{\mathbf{VS}^{+}}(\mathbf{x}) = \operatorname{diag}\left(h_{11}\left[\frac{\sigma^2(\mathbf{x})}{f_{\mathbf{X}}(\mathbf{x})}\right]^{\eta_{11}(\mathbf{x})}, ..., h_{pp}\left[\frac{\sigma^2(\mathbf{x})}{f_{\mathbf{X}}(\mathbf{x})}\right]^{\eta_{pp}(\mathbf{x})}\right) \tag{29}$$

because the number of parameters $h_{11}(\mathbf{x}), ..., h_{pp}(\mathbf{x})$ employed in (28) and the number of parameters $h_{11}, ..., h_{pp}$ in (29) are the same. Although the class of VS bandwidth matrices as defined in (29) is more suitable for estimation using a more complex data-generating process, optimizing (29) in terms of the AMISE is computationally complex in general.

It is also possible to propose the variance-stabilizing local variable full-bandwidth matrix $\mathbf{H}_{\mathbf{VS}^{++}}(\mathbf{x})$ that minimizes the AMISE. Because $\mathbf{H}_{\mathbf{VS}^{++}}(\mathbf{x})$ is more flexible than $\mathbf{H}_{\mathbf{VS}}(\mathbf{x})$ and $\mathbf{H}_{\mathbf{VS}^{+}}(\mathbf{x})$ in its matrix form, it is advantageous in the situations where sphering is inadvisable, such as in multimodal density settings of $\mathbf{X}$ or asymmetric data-generating processes. In a bivariate setting, one simple form of the VS full-bandwidth matrix is written as

$$\mathbf{H}_{\mathbf{VS}^{++}}(\mathbf{x}) = \begin{pmatrix} h_{11}\left[\frac{\sigma^2(\mathbf{x})}{f_{\mathbf{X}}(\mathbf{x})}\right]^{\eta_{11}(\mathbf{x})} & h_{12}\left[\frac{\sigma^2(\mathbf{x})}{f_{\mathbf{X}}(\mathbf{x})}\right]^{\frac{1}{2}} \\ h_{12}\left[\frac{\sigma^2(\mathbf{x})}{f_{\mathbf{X}}(\mathbf{x})}\right]^{\frac{1}{2}} & h_{22}\left[\frac{\sigma^2(\mathbf{x})}{f_{\mathbf{X}}(\mathbf{x})}\right]^{\eta_{22}(\mathbf{x})} \end{pmatrix}, \tag{30}$$

where

$$h_{11}h_{22} - h_{12}^2 > 0,\ h_{11},\ h_{22} > 0,$$

$$-\infty < \eta_{11}(\mathbf{x}) < \infty, \quad \eta_{11}(\mathbf{x}) + \eta_{22}(\mathbf{x}) = 1.$$

It is far more involved in terms of computation even in bivariate setting.

When we estimate the second derivative of $m(\mathbf{x})$ that appears in (12) and (13) in the estimation procedure, we employ a quartic polynomial estimator, as proposed by Fan and Gijbels (1995). The rule of thumb helps us estimate $\alpha_{ii}(\mathbf{x})$, $i = 1, ..., p$, with a comparatively smaller computational burden. However, it fails if the true regression curve shows a large degree of fluctuation over the domain. In this case, a more refined approach, such as employing a local cubic estimator, as proposed by Yang and Tschernig (1999), is needed. In this paper, we focus mainly on the performance of the estimator of the VS bandwidth matrix, so we employ a true regression function receptive to the quartic polynomial estimator in simulation studies.

To illustrate the issue of discontinuous MSE-minimizing LL estimators that we alluded to in the Introduction and Remark 7, we plot bivariate LL estimators based on the VS bandwidth matrix with its local parameters all set to be constant, and based on the theoretically MSE-minimizing scalar bandwidth matrix. We employ a true regression function, $m(x_1, x_2) = x_1^4 + x_2^4$. For the regression function, the denominator of the MSE-minimizing scalar bandwidth matrix takes a value of zero at $(x_{1*}, x_{2*}) = (0, 0)$. In other words,

$$\alpha_{ii}(x_{1*}, x_{2*}) = 0, \ i = 1, 2, \quad \text{or} \quad \left.\frac{\partial^2 m(x_1, x_2)}{\partial x_i^2}\right|_{x_1 = x_{1*}, x_2 = x_{2*}} = 0, \ i = 1, 2,$$

occurs at $(x_{1*}, x_{2*}) = (0, 0)$, where the curvature $|\partial^2 m(x_1, x_2)/\partial x_i^2| / \left[1 + [\partial m(x_1, x_2)/\partial x_i]^2\right]^{3/2}$, $i = 1, 2$, of $m(x_1, x_2)$ is zero at this (x_{1*}, x_{2*}).

To illustrate the problem numerically, we calculate the VS bandwidth matrix $\mathbf{H_{VS}}(\mathbf{x})$ with its local parameters set to $1/2$ and the theoretically MSE-minimizing scalar bandwidth matrix $\mathbf{H}_{var}(\mathbf{x})$, respectively, with $\sigma^2(x_1, x_2) = 0.2 + 0.1x_1^2 + 0.1x_2^2$, $f_{X_1, X_2}(x_1, x_2)$, a normal distribution with a mean of $(0, 0)$, a variance-covariance matrix $\mathrm{diag}(1, 1)$ truncated on $[-1.0, 1.0] \times [-1.0, 1.0]$, and a bivariate Gaussian kernel. The result is shown in the two bottom panels in Figure 5. To illustrate what effects these choices of bandwidth matrices have on the actual estimation of $m(x_1, x_2)$, we generate a data set $((X_{i1}, X_{i2}), Y_i)$, $i = 1, ..., 10,000$,

from the true functions and calculate the bivariate LL estimators using these bandwidth matrices. The result is shown in the two upper panels in Figure 5. To highlight the area around $(0,0)$, we show the results limited to $[-0.5,0.5] \times [-0.5,0.5]$ for all of the panels.

The upper left panel is a plot of the LL estimator with the VS bandwidth matrix, while the upper right panel is a plot of the LL estimator with the theoretically MSE-minimizing scalar bandwidth matrix. The bottom two panels of the figure are plots of the size of the first diagonal element in the corresponding bandwidth matrices at every point (x_1, x_2). As expected, we find a discontinuous point at $(x_{1*}, x_{2*}) = (0,0)$ in the upper right panel, which we do not see in the upper left panel. Although the MSE-minimizing scalar bandwidth matrix generates small vertical fluctuations in the LL estimator over most of the range, one discontinuous point at $(x_{1*}, x_{2*}) = (0,0)$ in the upper right panel shows that the LL estimator is considerbly in error in the vicinity of this point. On the other hand, while the VS bandwidth matrix with constant local parameters generates larger fluctuations in the LL estimator over most of the range, it does not have a discontinuous point.

Appendix 1

We present an algorithm to compute the multivariate LL estimator with the VS bandwidth matrix. For illustrative purposes, we consider a bivariate situation.

Stage 1. Estimation of $f_{\mathbf{X}}(x_1, ..., x_p)$.

When $p = 2$, to estimate $f_{X_1,X_2}(x_1, x_2)$, we employ the bivariate kernel density estimator,

$$\widehat{f_{\widehat{\mathbf{H_F}}}}(x_1, x_2) = \frac{1}{n|\widehat{\mathbf{H_F}}|} \sum_{i=1}^{n} K_{X_1,X_2}\left((x_1 - X_{i1}, x_2 - X_{i2})\widehat{\mathbf{H_F}}^{-1}\right),$$

where $K_{X_1,X_2}(\cdot, \cdot)$ is a bivariate Gaussian and the global diagonal bandwidth matrix $\widehat{\mathbf{H_F}}$ is denoted as

$$\widehat{\mathbf{H_F}} = \mathrm{diag}(\widehat{h_{f11}}, \widehat{h_{f22}}). \tag{31}$$

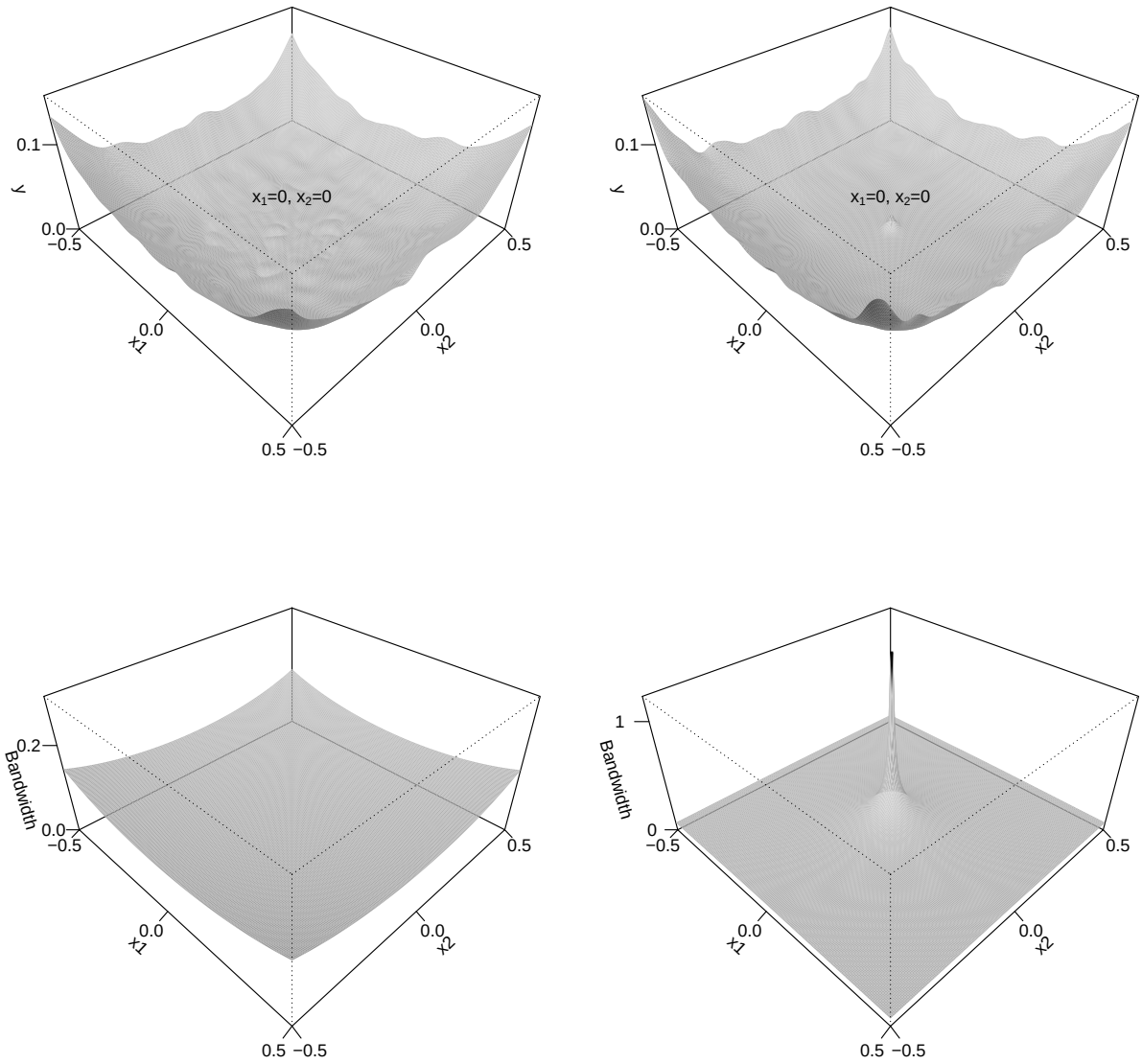

Figure 5: The upper left panel is a plot of the LL estimator with the VS bandwidth matrix with its local parameters set to be $1/2$, while the upper right panel is a plot of the LL estimator with the MSE-minimizing scalar bandwidth matrix for $m(x_1, x_2) = x_1^4 + x_2^4$. The lower left and right panels are plots of the corresponding sizes of the first diagonal element in the VS bandwidth matrix and the MSE-minimizing scalar bandwidth matrix, respectively, at every point (x_1, x_2). To highlight the area around $(0, 0)$, we show the results limited to $[-0.5, 0.5] \times [-0.5, 0.5]$ for all of the panels.

Assuming **A 3**, we employ Scott's rule (Scott 1992, p.152) written as

$$\widehat{h_{f11}} = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(X_{i1}-\bar{X}_{\cdot 1})^2}\cdot n^{-\frac{1}{6}}, \quad \widehat{h_{f22}} = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(X_{i2}-\bar{X}_{\cdot 2})^2}\cdot n^{-\frac{1}{6}}.$$

Stage 2. Estimation of $\alpha_{ii}(x_1,...,x_p)$, $i=1,...,p$.

This stage consists of three steps.

Step 1. Following Fan and Gijbels (1995), we estimate the quartic polynomial pilot estimator $\check{m}(x_1,x_2)$ of the form,

$$\begin{aligned}\check{m}(x_1,x_2) = {} & \widehat{t_0}+\widehat{t_1}x_1+\widehat{t_2}x_1^2+\widehat{t_3}x_1^3+\widehat{t_4}x_1^4+\widehat{t_5}x_2+\widehat{t_6}x_2^2+\widehat{t_7}x_2^3+\widehat{t_8}x_2^4 \\ & +\widehat{t_9}x_1x_2+\widehat{t_{10}}x_1x_2^2+\widehat{t_{11}}x_1x_2^3+\widehat{t_{12}}x_1^2x_2+\widehat{t_{13}}x_1^2x_2^2+\widehat{t_{14}}x_1^3x_2,\end{aligned}$$

by OLS.

Step 2. We select the best model that minimizes AIC by removing insignificant terms. We denote the predicted value at (x_1,x_2) as $\check{m}^{OLS}(x_1,x_2)$.

Step 3. We calculate point estimates of $\partial^2\check{m}^{OLS}(x_1,x_2)/\partial x_i^2$, $i=1,2$.

Stage 3. Estimation of $\sigma^2(x_1,...,x_p)$.

We employ $\widehat{m^{LL}_{\widehat{\mathbf{H}_\mathbf{M}}}}(x_1,x_2)$ to calculate the squared residuals,

$$\widehat{r}^2(X_{i1},X_{i2}) = \left(Y_i-\widehat{m^{LL}_{\widehat{\mathbf{H}_\mathbf{M}}}}(X_{i1},X_{i2})\right)^2, \quad i=1,...,n, \tag{32}$$

where $\widehat{\mathbf{H}_\mathbf{M}}$ is the global scalar bandwidth matrix defined to be

$$\widehat{\mathbf{H}_\mathbf{M}} = \mathrm{diag}(\widehat{h_m},\widehat{h_m}). \tag{33}$$

The estimator $\widehat{\mathbf{H}_\mathbf{M}}$ is selected so as to minimize the cross-validation statistics,

$$CV(\widehat{h_m}) = \frac{1}{n}\sum_{i=1}^{n}\left[Y_i-\widehat{m^{LL}_{-i,\widehat{\mathbf{H}_\mathbf{M}}}}(X_{i1},X_{i2})\right]^2,$$

where $\widehat{m^{LL}_{-i,\widehat{\mathbf{H_M}}}}(X_{i1}, X_{i2})$ is the leave-one-out LL estimator with its i-th element left out. Then, we construct a residual-based variance estimator,

$$\widehat{\sigma^2_{\widehat{\mathbf{H_V}}}}(X_{i1}, X_{i2}) = \frac{\sum_{i=1}^{n} K_{X_1,X_2}\left(\frac{X_{j1}-X_{i1}}{\widehat{h_v}}, \frac{X_{j2}-X_{i2}}{\widehat{h_v}}\right) \widehat{r}^2(X_{i1}, X_{i2})}{\sum_{i=1}^{n} K_{X_1,X_2}\left(\frac{X_{j1}-X_{i1}}{\widehat{h_v}}, \frac{X_{j2}-X_{i2}}{\widehat{h_v}}\right)}, \tag{34}$$

where $\widehat{\mathbf{H_V}}$ is the global scalar bandwidth matrix, defined to be

$$\widehat{\mathbf{H_V}} = \mathrm{diag}(\widehat{h_v}, \widehat{h_v}). \tag{35}$$

As an estimator of the bandwidth in (35), we employ the following bandwidth that minimizes the cross-validation statistics with respect to $\widehat{h_v}$,

$$CV(\widehat{h_v}) = \frac{1}{n}\sum_{i=1}^{n}\left[\widehat{r}^2(X_{i1}, X_{i2}) - \widehat{\sigma^2_{-i,\widehat{\mathbf{H_V}}}}(X_{i1}, X_{i2})\right]^2,$$

where $\widehat{\sigma^2_{-i,\widehat{\mathbf{H_V}}}}(X_{i1}, X_{i2})$ is the leave-one-out residual-based estimator with the i-th residual element left out. The bandwidth minimized in terms of cross-validation statistics is equivalent to the one that minimizes average squared error on average. The average squared error in mean is asymptotically equivalent to the MISE as presented in Marron and Härdle (1986). This is the reason for employing cross-validation statistics.

Remark 9. At the end of Stage 3, we are able to estimate $\widehat{\eta^*_{ii}}(x_1, x_2)$, $i = 1, 2$. If $\widehat{\alpha_{11}}(x_1, x_2) = \widehat{\alpha_{22}}(x_1, x_2) = 0$ occurs at the point (x_1, x_2) as indicated in Remark 1, we set $\widehat{\eta^*_{ii}}(x_1, x_2) = 1/2, i = 1, 2$. Similarly, if $\widehat{\alpha_{11}}(x_1, x_2) \neq 0$, $\widehat{\alpha_{22}}(x_1, x_2) = 0$, or $\widehat{\alpha_{11}}(x_1, x_2) = 0$, $\widehat{\alpha_{22}}(x_1, x_2) \neq 0$, occurs at the point (x_1, x_2) as pointed out in Remark 3, we set $\widehat{\eta^*_{11}}(x_1, x_2) = 1, \widehat{\eta^*_{22}}(x_1, x_2) = 0$, or $\widehat{\eta^*_{11}}(x_1, x_2) = 0$, $\widehat{\eta^*_{22}}(x_1, x_2) = 1$, respectively.

Stage 4. Compute $\widehat{h^*_0}$.

Once we have obtained $\widehat{f_{\widehat{\mathbf{H_F}}}}(x_1, x_2)$, $\widehat{\sigma^2_{\widehat{\mathbf{H_V}}}}(x_1, x_2)$, $\widehat{\alpha_{ii}}(x_1, x_2)$, $\widehat{\eta^*_{ii}}(x_1, x_2)$, $i = 1, 2$, in **Stages 1-3**, we can obtain the global

$$\widehat{h^*_0} = \left[\frac{R(K_{X_1,X_2})}{\mu_2^2(K_{X_1,X_2})\widehat{T_{VS}}(\widehat{\eta^*_{11}}(x_1, x_2), \widehat{\eta^*_{22}}(x_1, x_2))}\right]^{\frac{1}{6}} 2^{\frac{1}{6}} \cdot n^{-\frac{1}{6}},$$

by numerically integrating the function of the form

$$\widehat{T_{VS}}\left(\widehat{\eta^*_{11}}(x_1,x_2),\widehat{\eta^*_{22}}(x_1,x_2)\right)=\int\int_{I^2} w_{X_1,X_2}(x_1,x_2)\left[\sum_{i=1}^{2}\widehat{\alpha_{ii}}(x_1,x_2)\left[\frac{\widehat{\sigma^2_{\mathbf{H_V}}}(x_1,x_2)}{\widehat{f_{\mathbf{H_F}}}(x_1,x_2)}\right]^{2\widehat{\eta^*_{ii}}(x_1,x_2)}\right]^2 dx_1dx_2.$$

The weight function $w_{X_1,X_2}(x_1,x_2)$ is generally set to be $\widehat{f_{\mathbf{H_F}}}(x_1,x_2)$.

Stage 5. So far, one global $\widehat{h^*_0}$ and, at every point (x_1,x_2), $\left[\widehat{\sigma^2_{\mathbf{H_V}}}(x_1,x_2)/\widehat{f_{\mathbf{H_F}}}(x_1,x_2)\right]$ and $\widehat{\eta^*_{ii}}(x_1,x_2)$, $i=1,2$, are obtained. With the estimated VS bandwidth matrix,

$$\widehat{\mathbf{H_{VS}}}(x_1,x_2)=\widehat{h^*_0}\cdot\mathrm{diag}\left(\left[\frac{\widehat{\sigma^2_{\mathbf{H_V}}}(x_1,x_2)}{\widehat{f_{\mathbf{H_F}}}(x_1,x_2)}\right]^{\widehat{\eta^*_{11}}(x_1,x_2)},\left[\frac{\widehat{\sigma^2_{\mathbf{H_V}}}(x_1,x_2)}{\widehat{f_{\mathbf{H_F}}}(x_1,x_2)}\right]^{\widehat{\eta^*_{22}}(x_1,x_2)}\right)\cdot 2^{\frac{1}{6}}\cdot n^{-\frac{1}{6}}, \quad (36)$$

we calculate the bivariate LL estimator at every point (x_1,x_2) in the domain.

Appendix 2

A brief overview of all diagonal bandwidth matrices used in this paper.

(a) global scalar bandwidth matrix $\mathbf{H}=h\mathbf{I}_p$: $\widehat{\mathbf{H_M}}$ in (33) and $\widehat{\mathbf{H_V}}$ in (35).

(b) local variable scalar bandwidth matrix $\mathbf{H}(\mathbf{x})=h(\mathbf{x})\mathbf{I}_p$: $\mathbf{H}_{var}(\mathbf{x})$ in (9).

(c) global diagonal bandwidth matrix $\mathbf{H}=\mathrm{diag}(h_{11},...,h_{pp})$: $\widehat{\mathbf{H_F}}$ in (31).

(d) local variable diagonal bandwidth matrix $\mathbf{H}(\mathbf{x})=\mathrm{diag}(h_{11}(\mathbf{x}),...,h_{pp}(\mathbf{x}))$: $\mathbf{H}_{var+}(\mathbf{x})$ in (28).

(e) local variable diagonal bandwidth matrices with restrictions :

  (i) local variable diagonal VS bandwidth with scalar global parameter h_0 : $\mathbf{H_{VS}}(\mathbf{x})$ in (6).

  (ii) local variable diagonal VS bandwidth with diagonal global parameters $h_{11},...,h_{pp}$: $\mathbf{H_{VS^+}}(\mathbf{x})$ in (29).

(f) local variable full-bandwidth matrix with restrictions: $\mathbf{H_{VS^{++}}}(\mathbf{x})$ in (30).

Acknowledgements

Kiheiji NISHIDA acknowledges financial support from the Japan Society for the Promotion of Science under Grant-in-Aid for Research Activity Start-up 24830048. Yuichiro KANAZAWA is also grateful for financial support from the Japan Society for the Promotion of Science under Grant-in-Aide for Scientific Research (C)(2)12680310, (C)(2)16510103, and (B)20310081. Lastly, we thank two anonymous referees for many constructive comments on earlier version of the manuscript.

References

Fan, J. and Gijbels, I. (1992). Variable Bandwidth and Local Linear Regression Smoothers. The Annals of Statistics 20:2008-2036.

Fan, J. and Gijbels, I. (1995). Adaptive Order Polynomial Fitting: Bandwidth Robustification and Bias Reduction. Journal of Computational and Graphical Statistics 4:213-227.

Fan, J. and Yao, Q. (1998). Efficient Estimation of Conditional Variance Functions in Stochastic Regression. Biometrika 85:645-660.

Kanazawa, Y. (1992). An Optimal Variable Cell Histogram Based on the Sample Spacings. The Annals of Statistics 20: 291-304.

Kanazawa, Y. (1993a). Hellinger Distance and Akaike's Information Criterion for the Histogram. Statistics and Probability Letters 17: 293-298.

Kanazawa, Y. (1993b). Hellinger Distance and Kullback-Leibler Loss for the Kernel Density Estimator. Statistics and Probability Letters 18: 315-321.

Kanazawa, Y., Kogure, A., and Lee, S.G. (1999). On the Asymptotic Equivalence of Hellinger Distance and Kullback-Leibler Loss. Journal of the Japan Statistical Society 29: 1-21.

Marron, J.S. and Härdle, W. (1986). Random Approximations to Some Measures of Accuracy in Nonparametric Curve Estimation. Journal of Multivariate Analysis 20:91-113.

Nadaraya, E.A. (1964) On Estimating Regression. Theory of Probability and Its Applications. 9:141-142.

Nadaraya, E.A. (1965). On Nonparametric Estimation of Density Functions and Regression Curves. Theory of Probability and Its Applications 10:186-190.

Nadaraya, E.A. (1970). Remarks on Nonparametric Estimates for Density Functions and Regression Curves. Theory of Probability and Its Applications 15:134-137.

Nishida, K. and Kanazawa, Y. (2011). Introduction to the Variance-Stabilizing Bandwidth for the Nadaraya-Watson Regression Estimator. Bulletin of Informatics and Cybernetics 43:53-66.

Ruppert, D. and Wand, M.P. (1994). Multivariate Locally Weighted Least Squares Regression. The Annals of Statistics 22:1346-1370.

Scott, D.W. (1992). Multivariate Density Estimation: Theory, Practice, and Visualization. New York, Chichester: John Wiley & Sons.

Wand, M.P. and Jones, M.C. (1993). Comparison of Smoothing Parametrizations in Bivariate Kernel Density Estimation. Journal of the American Statistical Association 88:520-528.

Watson, G.S. (1964). Smooth Regression Analysis. Sankhyā Series A 26:359-372.

Watson, G.S. and Leadbetter, M.R. (1963). On the Estimation of Probability Density, I. Annals of Mathematical Statistics 34:480-491.

Yang, L. and Tschernig, R. (1999). Multivariate Bandwidth Selection for Local Linear Regression. Journal of Royal Statistical Society, Series B 61:793-815.