On the Variance-stabilizing Multivariate Nonparametric Regression Estimation

Department of Social Systems and Management

Discussion Paper Series

No.1275

On the Variance-stabilizing Multivariate Nonparametric Regression Estimation

by

Kiheiji NISHIDA, and Yuichiro KANAZAWA

July 2011

UNIVERSITY OF TSUKUBA
Tsukuba, Ibaraki 305-8573
JAPAN

On the Variance-Stabilizing Multivariate Nonparametric Regression Estimation

Kiheiji NISHIDA [(a)] and Yuichiro KANAZAWA [(b)]

(a) Researcher, Graduate School of Systems and Information Engineering, University of Tsukuba, 1-1-1 Ten-noh-dai, Tsukuba, Ibaraki 305-8573, JAPAN.
E-mail: kiheiji.nishida@gmail.com

(b) Professor, Department of Social Systems and Management, 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

ABSTRACT

Fan and Gijbels (1992) propose a local variable bandwidth that produces MSE minimizing univariate locally linear estimator (henceforth the LL). Their estimator does not stabilize variance over the domain. Moreover in the regions where underlying regression function has curvature zero, their resulting LL estimator is discontinuous. In this paper, we propose a variance-stabilizing (VS) diagonal bandwidth matrix for the multivariate LL estimator that does not manufacture such discontinuity. Theoretically, VS bandwidth can outperform the natural multivariate extension of Fan and Gijbels (1992) estimator in terms of asymptotic MISE. We present an estimating procedure of VS bandwidth and a simulation study.

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 $\mathbf{Y}$. Nonparametric approaches to explain the conditional expectation such as $E[\mathbf{Y}|\mathbf{X}] = m(\mathbf{x})$ are preferable in many cases. In this paper, we will concentrate on nonparametric kernel-type locally linear estimator (henceforce the LL estimator) as in Ruppert and Wand (1994), a popular approach in curve estimation.

Let us consider a $p+1$ row vector $(X_{i1}, ..., X_{ip}, Y_i)$ of random variables. We assume $\mathbf{x}_{i\cdot} = (x_{i1}, ..., x_{ip})$, $i = 1, ..., n$, are the realizations of random explanatory vector $\mathbf{X}_{i\cdot} = (X_{i1}, ..., X_{ip})$, i.i.d. with respect to i and whose joint density function is denoted as $f_{\mathbf{X}}(\mathbf{x})$ on support $I^p \in R^p$. The n sample realizations of $(X_{i1}, ..., X_{ip})$ can be written as,

$$\begin{pmatrix} x_{11} & \dots & x_{1p} \\ \vdots & \ddots & \vdots \\ x_{n1} & \dots & x_{np} \end{pmatrix}, \tag{1}$$

and we define the i-th column of the matrix in (1) as $\mathbf{x}_{\cdot i}$. We assume that the j-th random explanatory variable $\mathbf{X}_{\cdot j}$ may be correlated with the k-th variable $\mathbf{X}_{\cdot k}$, $j \neq k$. We assume 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 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

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

In subsequent expositions, we use a set of standard assumptions **S 1-4** in Appendix 1 on the explanatory variables $\mathbf{X}_{i\cdot}$, on the disturbances U_i and on the responses Y_i.

Let $K_{\mathbf{X}}(\mathbf{x})$ be the p-dimensional kernel function and $\mathbf{H}_{\mathbf{X}}$ be p-dimensional symmetric positive definite bandwidth matrix satisfying **A 2** and **3** in Appendix 1. The LL estimator of $m(\cdot)$ is given by the solution for β_0 that minimizes along with other β_j, $j = 1, ..., p$, the weighted least squares,

$$\min_{\beta_0, \beta_1, ..., \beta_p} \sum_{i=1}^{n} \left[Y_i - \beta_0 - \sum_{j=1}^{p} \beta_j (x_j - x_{ij}) \right]^2 K_{\mathbf{X}}((\mathbf{x} - \mathbf{x}_{i\cdot})\mathbf{H}_{\mathbf{X}}^{-1})$$

$$= \min_{\beta_0,\beta_1,...,\beta_p} [\mathbf{Y} - \mathbf{D}(\mathrm{x})\boldsymbol{\beta}]^T \mathbf{W}(\mathrm{x}) [\mathbf{Y} - \mathbf{D}(\mathrm{x})\boldsymbol{\beta}], \tag{2}$$

where

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

$\mathbf{W}(\mathrm{x}) = \mathrm{diag}\left(K_{\mathbf{X}}((\mathrm{x} - \mathrm{x}_{1\cdot})\mathbf{H}_{\mathbf{X}}^{-1}), ..., K_{\mathbf{X}}((\mathrm{x} - \mathrm{x}_{n\cdot})\mathbf{H}_{\mathbf{X}}^{-1})\right)$ is the matrix controlling the weight reflecting the relevant data points in calculating the LL estimator at 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 at length n and the term $\beta_0 + \sum_{j=1}^{p} \beta_j x_j$ is the linear approximation of $m(\mathrm{x})$ at the neighbourhood of $(x_1, ..., x_p)$ of x. Solving the minimization problem (2) with respect to $\boldsymbol{\beta}$ and retaining its intercept term β_0, we obtain the LL estimator,

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

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

Variance and bias of the LL estimator have been well known as in Ruppert and Wand (1994). With the standard set of assumptions on kernel **K 1** and the additional assumptions **A 2-7** on $f_{\mathbf{X}}(\mathrm{x})$, $\sigma^2(\mathrm{x})$, $m(\mathrm{x})$ and $\mathbf{H_X}$ in Appendix 1, the theoretical unconditional variance of the LL estimators is written as

$$V_{\mathbf{X}_{i\cdot},Y_i}\left[\widehat{m_{\mathbf{H_X}}^{LL}}(\mathrm{x})\right] = \frac{1}{n|\mathbf{H_X}|}\frac{\sigma^2(\mathrm{x})}{f_{\mathbf{X}}(\mathrm{x})}\left[\int \cdots \int K_{\mathbf{X}}^2(\mathbf{t})d\mathbf{t}\right] + o(1), \tag{3}$$

where $\mathbf{t} = (t_1, ..., t_p)$. Similarly the theoretical unconditional bias for the LL estimator at $\mathbf{x}$ is known to be

$$E_{\mathbf{X}_{i\cdot},Y_i}\left[\widehat{m_{\mathbf{H_X}}^{LL}}(\mathrm{x})\right] - m(\mathrm{x}) = \frac{\mu_2(K_{\mathbf{X}})}{2}\mathrm{trace}\left[\mathbf{H}_{\mathbf{X}}^T \nabla^2 m(\mathrm{x})\mathbf{H_X}\right] + o(1), \tag{4}$$

where the variance of kernel and the Hesse matrix are respectively defined to be

$$\mu_2(K_{\mathbf{X}}) = \int \cdots \int \mathbf{t}\mathbf{t}^T K_{\mathbf{X}}(\mathbf{t})d\mathbf{t}, \qquad \nabla^2 m(\mathrm{x}) = \begin{pmatrix} \frac{\partial^2 m(\mathbf{x})}{\partial x_1 \partial x_1} & \cdots & \frac{\partial^2 m(\mathbf{x})}{\partial x_1 \partial x_p} \\ \vdots & \ddots & \vdots \\ \frac{\partial^2 m(\mathbf{x})}{\partial x_p \partial x_1} & \cdots & \frac{\partial^2 m(\mathbf{x})}{\partial x_p \partial x_p} \end{pmatrix}.$$

If we can assume **A 1** in Appendix 1 whereby the sphering approach as in Wand and Jones (1993) is appropriate, we do not have to parametrize the off-diagonal elements of bandwidth matrix that reflect the correlation between explanatory variables, so that bandwidth matrix is diagonal.

As most nonparametric regression estimators choose their bandwidth by balancing the bias squared and variance either globally or locally, they do not produce constant variance over all values of combinations of regressor variable, unless one is dealing with rare occasions where the variability of response variable does not vary with the density of data points or where the covariate variables have joint distribution whose density compensates for the aforementioned variability of the response. This heteroscedastic nature is unsettling and, if possible, ought to be avoided.

Fan and Gijbels (1992) introduced the local variable bandwidth for the *univariate* LL estimator that minimizes the leading term of mean squared error (we call asymptotic MSE or AMSE),

$$\begin{aligned} &AMSE(m(\mathbf{x}),\ \widehat{m^{LL}_{\mathbf{H_X}}}(\mathbf{x})) \\ &= \frac{1}{n|\mathbf{H_X}|}\frac{\sigma^2(\mathbf{x})}{f_{\mathbf{X}}(\mathbf{x})}\left[\int\cdots\int K^2_{\mathbf{X}}(\mathbf{t})d\mathbf{t}\right] + \frac{\mu_2^2(K_{\mathbf{X}})}{4}\left[\text{trace}\left[\mathbf{H}^T_{\mathbf{X}}\nabla^2 m(\mathbf{x})\mathbf{H_X}\right]\right]^2. \end{aligned}$$

In the paper, they mentioned the possibility of the local variable bandwidth that stabilizes the variance of the nonparametric regression estimator (henceforth variance-stabilizing bandwidth or VS bandwidth), which according to our calculation come out to be

$$h_{VS}(x) \;=\; \frac{\sigma^2(x)}{f_X(x)}\cdot\left[\frac{\left[\int K^2_X(t)dt\right]}{\mu_2^2(K_X)\left[\int_I \frac{\sigma^8(x)[m^{(2)}(x)]^2}{f^3_X(x)}dx\right]}\right]^{\frac{1}{5}} n^{-\frac{1}{5}}.$$

They criticized such choice of bandwidth on the ground that the MSE minimizing local variable bandwidth in *univariate setting* will always outperform the VS bandwidth in terms of asymptotic mean integrated squared error (henceforth AMISE)

$$AMISE(m(\mathbf{x}),\ \widehat{m^{LL}_{\mathbf{H_X}}}(\mathbf{x})) = \int\cdots\int_{I^2} AMSE(m(\mathbf{x}),\ \widehat{m^{LL}_{\mathbf{H_X}}}(\mathbf{x}))f_{\mathbf{X}}(\mathbf{x})d\mathbf{x}.$$

The result is brought about by the fact that the univariate VS bandwidth is calculated so as to minimize MISE among the class of the 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.

In *multivariate regression setting*, however, their claim is not neccesary true. In other words, we are able to find a variance-stabilizing estimator whose MISE can outperform the MISE of a multivariate extension of Fan and Gijbels estimator. This is possible because in multivariate regression setting we can reduce sum of the MSE inflated by 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 (3) given by

$$\mathbf{H}_{\mathbf{VS}}(\mathbf{x}) = h_0 \cdot \begin{pmatrix} \left[\frac{\sigma^2(\mathbf{x})}{f_{\mathbf{X}}(\mathbf{x})}\right]^{\eta_1(\mathbf{x})} & 0 & 0 \; \dots & 0 \\ 0 & \left[\frac{\sigma^2(\mathbf{x})}{f_{\mathbf{X}}(\mathbf{x})}\right]^{\eta_2(\mathbf{x})} & 0 \; \dots & 0 \\ \vdots & \vdots & \ddots & \\ 0 & \dots & \dots & \left[\frac{\sigma^2(\mathbf{x})}{f_{\mathbf{X}}(\mathbf{x})}\right]^{\eta_p(\mathbf{x})} \end{pmatrix}, \tag{5}$$

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

$$\sum_{i=1}^{p} \eta_i(\mathbf{x}) = 1, \tag{6}$$

$$-\infty < \eta_i(\mathbf{x}) < \infty. \tag{7}$$

Both the global parameter h_0 and the local parameters, $\eta_i(\mathbf{x})$, $i = 1, ..., p$, can be determined so as to optimize AMISE. This optimized bandwidth can outperform *multivariate* extension of Fan and Gijbels (1992) local variable bandwidth,

$$\mathbf{H}_{var}(\mathbf{x}) = \left[\frac{\left[\int \cdots \int K_{\mathbf{X}}^2(\mathbf{t}) d\mathbf{t}\right] \sigma^2(\mathbf{x})}{\mu_2^2(K_{\mathbf{X}}) f_{\mathbf{X}}(\mathbf{x}) \left[\sum_{i=1}^{p} \alpha_i(\mathbf{x})\right]^2}\right]^{\frac{1}{p+4}} p^{\frac{1}{p+4}} \cdot n^{-\frac{1}{p+4}} \cdot \mathbf{I}_p, \tag{8}$$

$$\alpha_i(\mathbf{x}) = \frac{\partial^2 m(\mathbf{x})}{\partial x_i^2}, \quad \text{for } \; i = 1, ..., p, \tag{9}$$

which minimizes AMSE at every $\mathbf{x}$ among the class of diagonal variable bandwidth matrix,

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

The proposed VS bandwidth matrix is given in Proposition 1 along with the Remarks.

Our proposed VS bandwidth has practical strength over the MSE minimizing variable bandwidth in (8) in that it avoids discontinuity often encountered by (8): The denominator of the MSE minimizing local variable bandwidths in (8) are zero at the points satisfying $\left[\sum_{i=1}^{p} \alpha_i(\mathbf{x}_*)\right]^2 = 0$. Then, the bandwidth take infinitely large value and $\widehat{m_{\mathbf{H_X}}^{LL}}(\mathbf{x}_*)$ takes $\widehat{\mathbf{e_1}\boldsymbol{\beta}^{OLS}}$, the intercept term of OLS estimator. Since, $m(\mathbf{x}_*)$ does not coincide with $\widehat{\mathbf{e_1}\boldsymbol{\beta}^{OLS}}$ in general, this invites large bias at the corresponding points. However, our proposed VS bandwidth is continuous on these points as long as a standard assumption such as **A 4** in Appendix 1, is placed. We explain this issue on discussion.

In section 2, we introduce the VS bandwidth that minimizes AMISE and show a sufficient condition that enables the proposed VS bandwidth matrix to outperform (8). In section 3, we present a basic idea of estimating VS bandwidth matrix and simulation to check the performance of our proposed estimator in bivariate setting. The estimation algorithm is shown in Appendix 2. In section 4, we give discussions.

2 Introduction of the variance-stabilizing bandwidth

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

$$\mathbf{H_{VS}}(\mathbf{x}) \;=\; h_0^* \cdot \mathrm{diag}\left(\left[\frac{\sigma^2(\mathbf{x})}{f_{\mathbf{X}}(\mathbf{x})}\right]^{\eta_1^*(\mathbf{x})}, ..., \left[\frac{\sigma^2(\mathbf{x})}{f_{\mathbf{X}}(\mathbf{x})}\right]^{\eta_p^*(\mathbf{x})} \right), \tag{10}$$

minimizing asymptotic MISE is given by the following optimized parameters h_0^* and $\eta_i^*(\mathbf{x})$, $i = 1, ..., p$.

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

$$h_0^* \;=\; \left[\frac{\int \cdots \int K_{\mathbf{X}}^2(\mathbf{t})d\mathbf{t}}{\mu_2^2(K_{\mathbf{X}}) T_{VS}(\eta_1^*(\mathbf{x}), ..., \eta_p^*(\mathbf{x}))}\right]^{\frac{1}{p+4}} \cdot p^{\frac{1}{p+4}} \cdot n^{-\frac{1}{p+4}}, \tag{11}$$

where

$$T_{VS}(\eta_1^*(\mathbf{x}),...,\eta_p^*(\mathbf{x})) = \int \cdots \int_{I^p} f_{\mathbf{X}}(\mathbf{x}) \left[\sum_{i=1}^{p} \alpha_i(\mathbf{x}) \left[\frac{\sigma^2(\mathbf{x})}{f_{\mathbf{X}}(\mathbf{x})} \right]^{2\eta_i^*(\mathbf{x})} \right]^2 d\mathbf{x}. \tag{12}$$

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

$$\eta_i^*(\mathbf{x}) = \frac{\ln \left[\frac{\prod_{j=1}^{p} \alpha_j(\mathbf{x})}{[\alpha_i(\mathbf{x})]^p} \left[\frac{\sigma^2(\mathbf{x})}{f_{\mathbf{X}}(\mathbf{x})} \right]^2 \right]}{\ln \left[\frac{\sigma^2(\mathbf{x})}{f_{\mathbf{X}}(\mathbf{x})} \right]^{2p}} \tag{13}$$

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

Remark 1. If $\alpha_i(\mathbf{x}) = 0, i = 1,...,p$, the criterion function to appear in (16) takes zero minimum value for every $\eta_i(\mathbf{x})$, $i = 1,...,p$. At the points, any set of values $\eta_i^*(\mathbf{x})$, $i = 1,...,p$, satisfying (6) is available.

Remark 2. If $\alpha_i(\mathbf{x})$'s, $i = 1,...,p$, are not of the same sign, $\eta_i^*(\mathbf{x})$'s, $i = 1,...,p$, are not uniquely determined in general when $p \geq 3$. In a special case where the function (17) to appear does not have local maximal nor minimal values at $\mathbf{x}$, the optimal set of parameters $\eta_i^*(\mathbf{x})$, $i = 1,...,p$, is given by any set of values satisfying

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

Remark 3. If $\alpha_q(\mathbf{x})$ is zero and the rest of $\alpha_i(\mathbf{x})$'s, $i = 1,...,p, i \neq q$, are non-zero, we consider the $p-1$ dimensional minimization problem of (16) with the q-th variable left out of the minimization problem.

Proof of (i). We first choose h_0 given $\eta_i(\mathbf{x})$, $i = 1,...,p$, to minimize AMISE. Integrating square of the unconditional bias in (4) and the unconditional variance in (3) over the support I^p, the leading term of MISE between $\widehat{m_{\mathbf{H_X}}^{LL}}(\mathbf{x})$ and $m(\mathbf{x})$ is

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

Substituting $\mathbf{H_X}$ in (14) for $\mathbf{H_{VS}}(\mathbf{x})$ in (5), we obtain

$$\frac{1}{nh_0^p}\left[\int\cdots\int K_{\mathbf{X}}^2(\mathbf{t})d\mathbf{t}\right]+\frac{h_0^4}{4}\mu_2^2(K_{\mathbf{X}})T_{VS}(\eta_1(\mathbf{x}),...,\eta_p(\mathbf{x})). \tag{14}$$

The optimal global parameter (11) minimizes (14) with respect to h_0. □

Proof of (ii). We then optimize $\eta_i(\mathbf{x})$, $i=1,...,p$, in terms of AMISE. Plugging h_0^* in (11) into h_0 in (14), we obtain AMISE optimized in terms of h_0 written as

$$\left[\frac{\left[\int\cdots\int K_{\mathbf{X}}^2(\mathbf{t})d\mathbf{t}\right]^{\frac{4}{p+4}}}{\left[\mu_2^2(K_{\mathbf{X}})\right]^{-\frac{p}{p+4}}\left[T_{VS}(\eta_1(\mathbf{x}),...,\eta_p(\mathbf{x}))\right]^{-\frac{p}{p+4}}}\right]\left(p^{\frac{-p}{p+4}}+\frac{p^{\frac{4}{p+4}}}{4}\right)\cdot n^{-\frac{4}{p+4}}. \tag{15}$$

To minimize (15), the term $T_{VS}(\eta_1(\mathbf{x}),...,\eta_p(\mathbf{x}))$ defined in (12) must be minimized in terms of $\eta_i(\mathbf{x})$ $i=1,...,p$. For such $\eta_i(\mathbf{x})$, $i=1,...,p$, we solve the following constrained minimization problem in terms of $\eta_i(\mathbf{x})$, $i=1,...,p$, at every point of $\mathbf{x}$,

$$\min_{\eta_1(\mathbf{x}),...,\eta_p(\mathbf{x})}\left[\sum_{i=1}^{p}\alpha_i(\mathbf{x})\left[\frac{\sigma^2(\mathbf{x})}{f_{\mathbf{X}}(\mathbf{x})}\right]^{2\eta_i(\mathbf{x})}\right]^2,\ \text{subject to}\ \sum_{i=1}^{p}\eta_i(\mathbf{x})=1. \tag{16}$$

Let $G(\cdot)$ donote the $p-1$ variate function with respect to $\eta_1(\mathbf{x})$, ..., $\eta_{p-1}(\mathbf{x})$,

$$G(\eta_1(\mathbf{x}),...,\eta_{p-1}(\mathbf{x}))=\left[\sum_{i=1}^{p-1}\alpha_i(\mathbf{x})\left[\frac{\sigma^2(\mathbf{x})}{f_{\mathbf{X}}(\mathbf{x})}\right]^{2\eta_i(\mathbf{x})}\right]+\alpha_p(\mathbf{x})\left[\frac{\sigma^2(\mathbf{x})}{f_{\mathbf{X}}(\mathbf{x})}\right]^{2\left[1-\sum_{i=1}^{p-1}\eta_i(\mathbf{x})\right]}. \tag{17}$$

Differentiating (17) with respect to $\eta_i(\mathbf{x})$, $i=1,...,p-1$, and equating the outcome to be zero, we obtain the following simultaneous equations

$$\begin{aligned}\frac{\partial G(\eta_1(\mathbf{x}),...,\eta_{p-1}(\mathbf{x}))}{\partial\eta_i(\mathbf{x})} &= 2\left[\alpha_i(\mathbf{x})\left[\frac{\sigma^2(\mathbf{x})}{f_{\mathbf{X}}(\mathbf{x})}\right]^{2\eta_i(\mathbf{x})}-\alpha_p(\mathbf{x})\left[\frac{\sigma^2(\mathbf{x})}{f_{\mathbf{X}}(\mathbf{x})}\right]^{2\left[1-\sum_{i=1}^{p-1}\eta_i(\mathbf{x})\right]}\right]\ln\left[\frac{\sigma^2(\mathbf{x})}{f_{\mathbf{X}}(\mathbf{x})}\right]\\ &= 0,\quad i=1,...,p-1.\end{aligned} \tag{18}$$

Solving the simultaneous equations (18) and (6) with respect to $\eta_i(\mathbf{x})$, $i=1,...,p$, we obtain the first order condition,

$$\eta_i^*(\mathbf{x})=\frac{\ln\left[\frac{\prod_{j=1}^{p}\alpha_j(\mathbf{x})}{[\alpha_i(\mathbf{x})]^p}\left[\frac{\sigma^2(\mathbf{x})}{f_{\mathbf{X}}(\mathbf{x})}\right]^2\right]}{\ln\left[\frac{\sigma^2(\mathbf{x})}{f_{\mathbf{X}}(\mathbf{x})}\right]^{2p}},\quad i=1,...,p. \tag{19}$$

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

$$\Delta_k(\mathbf{x}) = \begin{vmatrix} A_{11}(\mathbf{x}) & A_{12}(\mathbf{x}) & \ldots & A_{1k}(\mathbf{x}) \\ A_{21}(\mathbf{x}) & A_{22}(\mathbf{x}) & \ldots & A_{2k}(\mathbf{x}) \\ \vdots & \vdots & \ddots & \vdots \\ A_{k1}(\mathbf{x}) & A_{k2}(\mathbf{x}) & \ldots & A_{kk}(\mathbf{x}) \end{vmatrix}$$

$$= \left[2\ln\left[\frac{\sigma^2(\mathbf{x})}{f_{\mathbf{X}}(\mathbf{x})}\right]\right]^{2k} \sum_{j=1}^{k+1} \prod_{i=1, i\neq j}^{k+1} \left[\alpha_i(\mathbf{x}) \left[\frac{\sigma^2(\mathbf{x})}{f_{\mathbf{X}}(\mathbf{x})}\right]^{2\eta_i^*(\mathbf{x})}\right], \tag{20}$$

where

$$A_{ij}(\mathbf{x}) = \left.\frac{\partial^2 G(\eta_1(\mathbf{x}), ..., \eta_{p-1}(\mathbf{x}))}{\partial\eta_i(\mathbf{x})\partial\eta_j(\mathbf{x})}\right|_{\eta_1(\mathbf{x})=\eta_1^*(\mathbf{x}),...,\eta_{p-1}(\mathbf{x})=\eta_{p-1}^*(\mathbf{x})}, \; i = 1, ..., k, \; j = 1, ..., k.$$

If $\alpha_i(\mathbf{x}) > 0$, $i = 1, ..., p$, the sequence of the principal minors (20) are $\Delta_1(\mathbf{x}) > 0$, $\Delta_2(\mathbf{x}) > 0$, ..., $\Delta_{p-1}(\mathbf{x}) > 0$. This means the function (17) takes positive minimal value under the first order condition (19). On the other hand, if $\alpha_i(\mathbf{x}) < 0$, $i = 1, ..., p$, the sequence of the principal minors (20) are $\Delta_1(\mathbf{x}) < 0$, $\Delta_2(\mathbf{x}) > 0$, $\Delta_3(\mathbf{x}) < 0$,... This means the criterion function (17) takes negative maximal value under the first order condition (19). Since the criterion function in (16) is the squared of the function (17), the first order condition (19) optimizes the minimization problem (16). □

Remark 4. Interpretation of the two parameters are as folows. As for h_0, see (14). The parameter h_0 plays a role to control AMISE globally. As for the interpretation of $\eta_i(\mathbf{x})$, $i = 1, ..., p$, this set of parameters are intended to cancel out the variable part $\sigma^2(\mathbf{x})/f_{\mathbf{X}}(\mathbf{x})$ of the variance (3) and to reduce AMSE locally. Furthermore, the local parameters $\eta_i(\mathbf{x})$, $i = 1, ..., p$, serve as an adjustment to stabilize the variance at the expense of bias. Especially when $0 \leq \eta_i(\mathbf{x}) \leq 1$, $i = 1, ..., p$, the parameters $\eta_i(\mathbf{x})$, $i = 1, ..., p$, can be interpreted as the fractional rate of the power of the squared bias that should be distributed to every coordinate axis, $x_1, ..., x_p$.

Remark 5. Suppose that $\eta_i(\mathbf{x})$, $i = 1, ..., p$, do not depend on $\mathbf{x}$ like as $\eta_1, ..., \eta_p$, and $\sum_{i=1}^p \eta_i = 1$ at all points $\mathbf{x}$. These globally determined parameters can also achieve the purpose to cancel out the term $\sigma^2(\mathbf{x})/f_{\mathbf{X}}(\mathbf{x})$ in (3). However, this globally determined η_i's cannot reduce AMISE so much as the locally determined $\eta_i(\mathbf{x})$'s.

We present and explain theoretical strength of our proposed VS bandwidth over the MSE minimizing variable bandwidth. Let $\gamma(\mathbf{x})$ denote the ratio of two "density" functions,

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

Then, when employed VS bandwidth and MSE minimizing variable bandwidth, AMISE's are respectively written as

$$AMISE(m(\mathbf{x}),\ \widehat{m^{LL}_{\mathbf{H_{VS}}}}(\mathbf{x})) = C_1 \cdot n^{-\frac{4}{p+4}} \cdot \left[\int \cdots \int_{I^p} f_{\mathbf{X}}(\mathbf{x}) \gamma^{\frac{4}{p}}(\mathbf{x}) \left[p \prod_{i=1}^{p} |\alpha_i(\mathbf{x})|^{\frac{1}{p}} \right]^2 d\mathbf{x} \right]^{\frac{p}{p+4}}, \quad (21)$$

$$AMISE(m(\mathbf{x}),\ \widehat{m^{LL}_{\mathbf{H}_{var}}}(\mathbf{x})) = C_1 \cdot n^{-\frac{4}{p+4}} \cdot \int \cdots \int_{I^p} f_{\mathbf{X}}(\mathbf{x}) \left[\gamma^{\frac{4}{p}}(\mathbf{x}) \left[\sum_{i=1}^{p} \alpha_i(\mathbf{x}) \right]^2 \right]^{\frac{p}{p+4}} d\mathbf{x}, \quad (22)$$

where $C_1 = \left(p^{-p/(p+4)} + p^{4/(p+4)}/4\right) \left[\int \cdots \int K^2_{\mathbf{X}}(\mathbf{x}) d\mathbf{x}\right]^{4/(p+4)} \left[\mu_2^2(K_{\mathbf{X}})\right]^{p/(p+4)}$ $\times \left[\int \cdots \int_{I^p} \sigma^2(\mathbf{x}) d\mathbf{x}\right]^{4/(p+4)} > 0$. We obtain the following proposition as to the magnitude relationship between (21) and (22).

Proposition 2. *Suppose that $\alpha_i(\mathbf{x}) > 0$, $i = 1, ..., p$, or $\alpha_i(\mathbf{x}) < 0$, $i = 1, ..., p$, holds at every $\mathbf{x}$. Then, the magnitude relationship of AMISE between the VS bandwidth matrix in* (10) *and the MSE minimizing variable bandwidth matrix in* (8) *is determined as follows.*

(i) *When $p = 1$, $AMISE(m(x),\ \widehat{m^{LL}_{h_{VS}}}(x))$ is always larger than $AMISE(m(x),\ \widehat{m^{LL}_{h_{var}}}(x))$.*

(ii) *When $p > 1$, a sufficient condition under which $AMISE(m(\mathbf{x}),\ \widehat{m^{LL}_{\mathbf{H_{VS}}}}(\mathbf{x}))$ is smaller than $AMISE(m(\mathbf{x}),\ \widehat{m^{LL}_{\mathbf{H}_{var}}}(\mathbf{x}))$ is*

$$\gamma^{\frac{4}{p}}(\mathbf{x}) \left[\sum_{i=1}^{p} \alpha_i(\mathbf{x}) \right]^2 = C, \quad \textit{at every } \mathbf{x}, \textit{ where } C > 0 \textit{ is any positive constant.} \quad (23)$$

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

$$AMISE(m(x),\ \widehat{m^{LL}_{h_{VS}}}(x)) - AMISE(m(x),\ \widehat{m^{LL}_{h_{var}}}(x))$$

$$= \frac{5}{4} \cdot n^{-\frac{4}{5}} \left[\int K_X^2(t)dt \right]^{\frac{4}{5}} \left[\mu_2^2(K_X) \right]^{\frac{1}{5}} \left[\int_I \sigma^2(x)dx \right]^{\frac{4}{5}}$$
$$\times \left[\int_I f_X^{\frac{4}{5}}(x) \left[f_X(x)\alpha^2(x)\gamma^4(x) \right]^{\frac{1}{5}} dx - \left[\int_I f_X(x)dx \right]^{\frac{4}{5}} \left[\int_I f_X(x)\alpha^2(x)\gamma^4(x)dx \right]^{\frac{1}{5}} \right] \leq 0.$$

□

Proof of (ii). When $p > 1$, if we employ (23), we obtain the relation

$$AMISE^{\frac{p+4}{p}}(m(\mathbf{x}),\ \widehat{m_{\mathbf{H_{VS}}}^{LL}}(\mathbf{x})) - AMISE^{\frac{p+4}{p}}(m(\mathbf{x}),\ \widehat{m_{\mathbf{H}_{var}}^{LL}}(\mathbf{x}))$$
$$= [C_1]^{\frac{p+4}{p}} \cdot C^{\frac{p}{p+4}} \cdot n^{-\frac{4}{p}} \cdot \int \cdots \int_{I^p} f_{\mathbf{X}}(\mathbf{x}) \left[\frac{\left[\sum_{i=1}^p \alpha_i(\mathbf{x})\right]^2 - \left[p \prod_{i=1}^p |\alpha_i(\mathbf{x})|^{\frac{1}{p}} \right]^2}{\left[\sum_{i=1}^p \alpha_i(\mathbf{x})\right]^2} \right] d\mathbf{x}. \quad (24)$$

Since $\left[\sum_{i=1}^p \alpha_i(\mathbf{x})\right]^2 - \left[p \prod_{i=1}^p |\alpha_i(\mathbf{x})|^{1/p} \right]^2 \geq 0$ always holds at every $\mathbf{x}$, the equation (24) is always greater than or equal to zero under the sufficient condition (23). □

Remark 6. Proposition 2-(ii) is brought about by the fact that the VS bandwidth matrix has more flexibility in its matrix form than the diagonal MSE minimizing bandwidth in (8). 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 bandwidth has one local parameter at the same given point $\mathbf{x}$. However, in univariate setting, the VS bandwidth is reduced to have one global parameter, while the MSE minimizing bandwidth remains to have one local parameter at a given point $\mathbf{x}$. As a result, the VS bandwidth will not be able to outperform the MSE minimizing variable bandwidth by definition as in Proposition 2-(i).

3 Estimation of the variance-stabilizing bandwidth

To estimate the VS bandwidth matrix, the global parameter h_0^* in (11), the local parameters $\eta_i^*(\mathbf{x})$, $i = 1, ..., p$, in (13), $\sigma^2(\mathbf{x})$ and $f_{\mathbf{X}}(\mathbf{x})$ must be estimated. Basic idea is to individually estimate components in (10), (11) and (13), $\widehat{f_{\mathbf{X}}}(\mathbf{x})$, $\widehat{\alpha_i}(\mathbf{x}) = \widehat{\partial^2 m(\mathbf{x})/\partial x_i^2}$, $i = 1, ..., p$, $\widehat{\sigma^2}(\mathbf{x})$, and plug these estimators into (10), (11) and (13). This idea guarantees weak consisitency of the LL estimator with the VS bandwidth matrix, while simultaneously achieving homoscedasticity of $\widehat{m^{LL}_{\widehat{\mathbf{H_{VS}}}}}(\mathbf{x})$, 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 example of the plug-in algorithms along with some details for bivariate setting in Appendix 2.

In the algorithm, we use the quartic polynomial estimator proposed by Fan and Gijbels (1995) to allow for flexibility in estimating the second derivative of $m(\mathbf{x})$, or $\alpha_i(\mathbf{x})$, $i = 1, ..., p$, in (11) and (13). As for estimating $\widehat{\sigma^2}(\mathbf{x})$, we employ "residual-based", estimator explained in Fan and Yao (1998) which smoothes squared residuals $(Y_i - \widehat{m}(\mathbf{x}))^2$ by the Nadaraya-Watson regression estimator. The bivariate extention of residual-based estimator is to appear in (30) and we compute its bandwidth estimator so as to minimize cross-validation statistics to appear in (32) among the class of diagonal bandwidth matrix to appear in (31). To calculate the squared residuals, we estimate $m(\mathbf{x})$ by LL estimator with its bandwidth estimated so as to minimize cross-validation statistics to appear in (29) among the class of diagonal bandwidth matrix to appear in (28).

Simulation study

In the simulation study, we would first like to see if the proposed algorithm obtains $\widehat{h^*_0}$ close to h^*_0 in (11). We would also like to see if the proposed estimator of the VS bandwidth matrix in (33) stabilizes the variances of the LL estimator in general. We also evaluate our proposed estimator of the VS bandwidth relative to the *best possible*, that is, theoretically MSE minimizing variable bandwidth. So, for the latter, we employ multivariate extension of Fan and Gijbels (1992) local variable bandwidth in (8).

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 $N((0,0)^T, \text{diag}(0.25^2, 0.25^2))$ truncated on bounded domain $[-0.5, 0.5] \times [-0.5, 0.5]$. In this setting, 91.1% of the data points distributed as $N((0,0)^T, \text{diag}(0.5^2, 0.5^2))$ is included in the domain $[-0.5, 0.5] \times [-0.5, 0.5]$. The true regression function and the conditional variance function are respectively set to be $m(x_1, x_2) = 1 - x_1^2 - x_2^2$ as in the left panel of Figure 1 and $\sigma^2(x_1, x_2) = 0.25 + 0.5x_1^2 + 0.5x_2^2$ as in the right panel of Figure 1. In this setup, variance measured in terms of (3) grows large near boundaries. Also, this setup assures us that there are no points of $\mathbf{x}$ satisfying

$\left[\sum_{i=1}^{p} \alpha_i(\mathbf{x})\right]^2 = 0$ on $[-0.5, 0.5] \times [-0.5, 0.5]$ and the MSE minimizing variable bandwidth in (8) does not produce discontinuous points as mentioned in section 1. As kernel, we employ bivariate Gaussian.

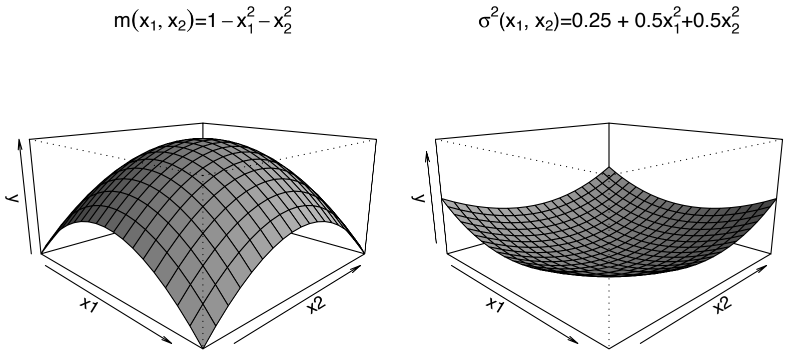

Figure 1: Graphics of the true regression function on the left and the conditional variance function on the right in our simulation study.

Procedure for simulation

For $n = 500$, 1000, $5,000$, $10,000$ and $15,000$:

1. Generate (X_{i1}, X_{i2}) of sample size n distributed as $f_{X_1,X_2}(x_1, x_2)$.
2. Generate $U_i|\{(X_{i1}, X_{i2}) = (x_{i1}, x_{i2})\}$ of sample size n distributed as $N(0, \sigma^2(x_{i1}, x_{i2}))$.
3. Obtain $((X_{i1}, X_{i2}), Y_i)$ of sample size n, where $Y_i = m(x_{i1}, x_{i2}) + U_i|\{(X_{i1}, X_{i2}) = (x_{i1}, x_{i2})\}$.
4. Estimate $\widehat{\mathbf{H_{VS}}}(x_1, x_2)$ at every point $(-0.5+\epsilon_1 \cdot j,\ -0.5+\epsilon_1 \cdot k)$, $\epsilon_1 = 0.01$, $j = 1, ..., 100$, $k = 1, ..., 100$, using the sample $((X_{i1}, X_{i2}), Y_i)$, $i = 1, ..., n$, obtained in $1 \sim 3$ above.
5. Construct VS LL estimator $\widehat{m^{LL}_{\widehat{\mathbf{H_{VS}}}}}(x_1, x_2)$. Similarly, construct LL estimator with the MSE minimizing variable bandwidth (8) written as $\widehat{m^{LL}_{\mathbf{H}_{var}}}(x_1, x_2)$.
6. Repeat $1 \sim 5$ $M = 100$ times.
7. Obtain the mean and the standard deviation of $\widehat{h_0^*}$ calculated $M = 100$ times in $1 \sim 5$, and by numerically integrate $\widehat{MISE_{VS}}$ given by

$$\widehat{MISE_{VS}} = \widehat{MISE}(m(x_1, x_2),\ \widehat{m^{LL}_{\widehat{\mathbf{H_{VS}}}}}(x_1, x_2))$$

$$= \frac{1}{M}\sum_{\mathbf{T}=1}^{M}\left[\int\int_{I^2} f_{X_1,X_2}(x_1,x_2)\left[m(x_1,x_2) - \widehat{m_{\widehat{\mathbf{H_{VS}}}}^{LL}}^{(\mathbf{T})}(x_1,x_2)\right]^2 dx_1dx_2\right], \quad (25)$$

where $\widehat{m_{\widehat{\mathbf{H_{VS}}}}^{LL}}^{(\mathbf{T})}(x_1,x_2)$ is the LL estimator calculated $(\mathbf{T})$ th generated sample of size n. Replacing $\widehat{m_{\widehat{\mathbf{H_{VS}}}}^{LL}}(x_1,x_2)$ in (25) with $\widehat{m_{\mathbf{H}_{var}}^{LL}}(x_1,x_2)$, we calculate $\widehat{MISE_{var}}$ as well. The ratio $\widehat{MISE_{VS}}/\widehat{MISE_{var}}$ is also calculated here.

8. At every point $(-0.5+\epsilon_1\cdot j,\ -0.5+\epsilon_1\cdot k)$, $\epsilon_1 = 0.01$, $j = 1,...,100$, $k = 1,...,100$, compute the sample variances of $\widehat{m_{\widehat{\mathbf{H_{VS}}}}^{LL}}^{(\mathbf{T})}(x_1,x_2)$ and $\widehat{m_{\mathbf{H}_{var}}^{LL}}^{(\mathbf{T})}(x_1,x_2)$, $\mathbf{T} = 1,...,M$, that are respectively calculated in 1 ∼ 5.
9. Obtain distributions of the sample variances of $\widehat{m_{\widehat{\mathbf{H_{VS}}}}^{LL}}^{(\mathbf{T})}(x_1,x_2)$ and $\widehat{m_{\mathbf{H}_{var}}^{LL}}^{(\mathbf{T})}(x_1,x_2)$, $\mathbf{T} = 1,...,M$, calculated at $10,000 = 100\times 100$ points in 8. Calculate means, standard deviations and the coefficients of variations of the sample variances of $\widehat{m_{\widehat{\mathbf{H_{VS}}}}^{LL}}^{(\mathbf{T})}(x_1,x_2)$ and $\widehat{m_{\mathbf{H}_{var}}^{LL}}^{(\mathbf{T})}(x_1,x_2)$, $\mathbf{T} = 1,...,M$.

The simulation result

Table 1 shows the result of simulation using the procedure 1 ∼ 7. In the table, we present means and standard deviations and the ratios $\widehat{h_0^*}/h_0^*$ of $M = 100$ simulated $\widehat{h_0^*}$ for $n = 500, 1,000, 5,000, 10,000, 15,000$. These numbers show that the estimator $\widehat{h_0^*}$ converges to h_0^* and is stable. We also present the estimates of $\widehat{MISE_{VS}}$, $\widehat{MISE_{var}}$ and the ratio $\widehat{MISE_{VS}}/\widehat{MISE_{var}}$. From these numbers on $\widehat{MISE_{VS}}$, we notice that $\widehat{MISE_{VS}}$'s approach to zero as n gets larger and, thus, $\widehat{MSE_{VS}}$'s approach to zero as well as n gets larger except for the countable number of points (x_1,x_2). Therefore, pointwise convergence of $\widehat{m_{\widehat{\mathbf{H_{VS}}}}^{LL}}(x_1,x_2)$ to $m(x_1,x_2)$ in the sense of mean square and thus weak consistency of $\widehat{m_{\widehat{\mathbf{H_{VS}}}}^{LL}}(x_1,x_2)$ to $m(x_1,x_2)$ are confirmed. Also the ratios $\widehat{MISE_{VS}}/\widehat{MISE_{var}}$ in Table 1 show that the price for homoscedasticity of the estimate decreases considerbly as the sample size n increases.

To see if the variance is stabilized by the proposed VS bandwidth, we present in Figure 2 boxplots of sample variances of $\widehat{m_{\widehat{\mathbf{H_{VS}}}}^{LL}}^{(\mathbf{T})}(x_1,x_2)$ and $\widehat{m_{\mathbf{H}_{var}}^{LL}}^{(\mathbf{T})}(x_1,x_2)$, $\mathbf{T} = 1,...,100$, by 0.05 intervals on 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 2 share the same scale in terms of y-axis, but the scale of y-axis is shrunk from top to bottom. Since $m(x_1, x_2) = 1 - x_1^2 - x_2^2$ 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 only along with x_1-axis. From Figure 2, we notice that comparatively smaller variances are achieved by the VS bandwidth near the boundaries with sample size greater than $5,000$, a piece of evidence that the estimator of VS bandwidth stabilizes variance of the LL estimator when a sample size is large enough. Table 2 summarizes Figure 2.

We notice from Table 2 that both sample means and standard deviations of the sample variances under both VS and MSE bandwidth diminish as sample size gets large. When sample size is greater than $5,000$, we also notice that the estimator of the VS bandwidth achieves smaller sample means and standard deviations of the sample variance relative to the theoretically MSE minimizing variable bandwidth.

n	$\widehat{h_0^*}$		$\widehat{h_0^*}/h_0^*$	$\widehat{MISE_{VS}}$	$\widehat{MISE_{var}}$	$\widehat{MISE_{VS}}/\widehat{MISE_{var}}$
	mean	std.dev	mean			
500	0.2223	0.0402	1.1344	$1.6683 \cdot 10^{-2}$	$6.0058 \cdot 10^{-3}$	2.7778
1,000	0.1963	0.0345	1.0951	$8.1323 \cdot 10^{-3}$	$4.0317 \cdot 10^{-3}$	2.0170
5,000	0.1489	0.0057	1.0866	$1.4029 \cdot 10^{-3}$	$1.1804 \cdot 10^{-3}$	1.1884
10,000	0.1322	0.0038	1.0821	$8.6218 \cdot 10^{-4}$	$7.6824 \cdot 10^{-4}$	1.1223
15,000	0.1239	0.0022	1.0861	$6.1840 \cdot 10^{-4}$	$5.4853 \cdot 10^{-4}$	1.1273

Table 1: Simulation result : Estimation of h_0^*, $\widehat{MISE_{VS}}$ and $\widehat{MISE_{var}}$.

Mean and std.dev. of sample variances of the LL at 100×100 points ($M = 100$)						
n	Estimator of VS bandwidth (33)			MSE minimizing variable bandwidth (8)		
	mean	std.dev.	coef.var.	mean	std.dev.	coef.var.
500	$1.7116 \cdot 10^{-2}$	$2.6967 \cdot 10^{-2}$	1.5755	$1.1799 \cdot 10^{-2}$	$1.1304 \cdot 10^{-2}$	0.9580
1,000	$1.1388 \cdot 10^{-2}$	$9.5131 \cdot 10^{-3}$	0.8353	$7.6741 \cdot 10^{-3}$	$7.2046 \cdot 10^{-3}$	0.9388
5,000	$1.9352 \cdot 10^{-3}$	$1.1351 \cdot 10^{-3}$	0.5865	$2.3556 \cdot 10^{-3}$	$2.6702 \cdot 10^{-3}$	1.1335
10,000	$1.1992 \cdot 10^{-3}$	$7.6798 \cdot 10^{-4}$	0.6404	$1.6076 \cdot 10^{-3}$	$2.0903 \cdot 10^{-3}$	1.3002
15,000	$8.2800 \cdot 10^{-4}$	$5.0645 \cdot 10^{-4}$	0.6116	$1.1818 \cdot 10^{-3}$	$1.6600 \cdot 10^{-3}$	1.4046

Table 2: Simulation result : Summary of Figure 2 to check if the variance is stabilized.

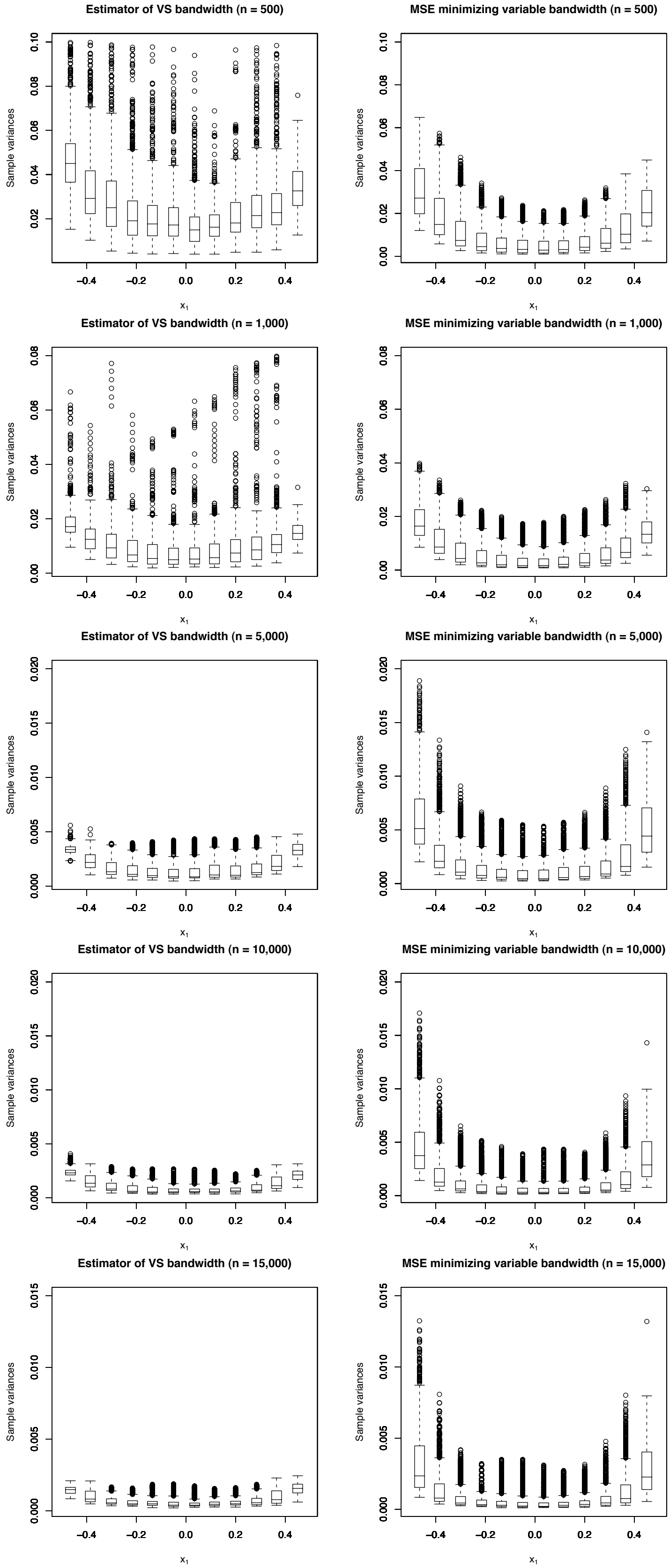

Figure 2: Simulation result : Distributions of sample variance at different point.

4 Discussion

In this paper, we first introduce the multivariate VS bandwidth matrix by simultaneously employing two types—global and local—of parameters and derive theoretically optimal parameter values in Proposition 1. In Proposition 2, we give a sufficient condition under which our proposed VS bandwidth theoretically outperforms the MSE minimizing variable bandwidth, a natural multivariate extension of Fan and Gijbels (1992). This proposition reveals that our VS bandwidth can outperform MSE minimizing variable bandwidth in terms of AMISE in *multivariate setting.* It also shows why Fan and Gijbels (1992) decided not to employ variance stabilizing approach in constructing LL estimate in their *univariate setting.*

In section 3 and in Appendix 2, we illustrate an idea and the corresponding algorithm to estimate VS bandwidth and perform a simulation study to find out that the global parameter h_0^* is successfully estimated using the algorithm. We also find that penalty incurred by employing VS bandwidth relative to the MSE minimizing variable bandwidth decreases from 2.7778 to 1.1273 as the sample size increases in terms of estimated MISE. This penalty, we find, is caused mainly by inflated bias for employing VS bandwidth.

Our proposed VS bandwidth is so designed as to negate the variable part in the variance of the LL estimator (3). The result given in Figure 2 and numerically summarized in Table 2 shows that, under the proposed VS bandwidth selection algorithm, the variance is stabilized as the sample size increases, and also in comparison with the MSE minimizing variable bandwidth.

Some may argue that another type of MSE minimizing variable bandwidth that minimizes AMSE among the class of diagonal bandwidth matrix,

$$\mathbf{H}_{var^+}(\mathbf{x}) = \mathrm{diag}\left(h_{11}(\mathbf{x}), ..., h_{pp}(\mathbf{x})\right), \tag{26}$$

should be compared with (5) along with (6) and (7). However, we feel that the class of VS bandwidth matrix that ought to be compared with (26) is the one that minimizes AMISE among the class of

$$\mathbf{H}_{\mathbf{VS}^{+}}(\mathrm{x}) = \mathrm{diag}\left(h_{11}\left[\frac{\sigma^2(\mathbf{x})}{f_{\mathbf{X}}(\mathbf{x})}\right]^{\eta_1(\mathbf{x})}, ..., h_{pp}\left[\frac{\sigma^2(\mathbf{x})}{f_{\mathbf{X}}(\mathbf{x})}\right]^{\eta_p(\mathbf{x})}\right) \tag{27}$$

because the number of parameters $h_{11}(\mathbf{x}), ..., h_{pp}(\mathbf{x})$ employed in (26) and the number of parameters $h_{11}, ..., h_{pp}$ in (27) are the same. In general, optimizing the class of VS bandwidth (27) in terms of AMISE is theoretically and computationally far more complex.

When we estimate the second derivative of $m(\mathbf{x})$ that appears in (11) and (13) in the estimation procedure, we employ the quartic polynomial estimator as in Fan and Gijbels (1992). The rule of thumb helps us estimate $\alpha_i(\mathbf{x})$, $i = 1, ..., p$, with comparatively smaller computational burden. However, it fails if the true regression curve shows a large degree of fluctuations over the domain. If so, more refined approach such as employing local polynomial estimator would be needed. In this paper, we focus mainly on the performance of the estimator of VS bandwidth, so we employ a true regression function receptive to the quartic polynomial estimator in simulation.

To illustrate the issue of discontinuous MSE minimizing LL estimators we alluded to in introduction, we plot the LL estimators with both bandwidths, VS bandwidth and MSE minimizing bandwidth in univariate setting. We employ true regression function, $0.5x^4$. For the regression function, the denominator of the MSE minimizing variable bandwidth takes zero value at $x_* = 0$ that satisfies

$$\alpha_1^2(x_*) = 0 \quad \text{or,} \quad m^{(2)}(x_*) = 0,$$

where the curvature $|m^{(2)}(x)|/\left[1 + [m^{(1)}(x)]^2\right]^{3/2}$ of $m(x)$ is zero at this x_*. This univariate setup itself highlights the existence of discontinuous points, which we carefully avoided in our bivariate simulation setting in section 3. To numerically illustrate the problem, we calculate the VS bandwidth $h_{VS}(x)$ and the theoretically MSE minimizing variable bandwidth $h_{var}(x)$ respectively with $\sigma^2(x) = 0.1(|x| + 1)$, $f_X(x)$, a normal distribution with its mean 0 and standard deviation $\sqrt{2.5}$ truncated on $[-2, 2]$, and Gaussian kernel. The result is shown in two bottom panels in Figure 3. To illustrate what effects these choice of bandwidths have on the actual estimation of $m(x)$, we generate a data set (X_i, Y_i), $i = 1, ..., 1000$, from the true functions and calculate the univariate LL estimators using these bandwidths. The result is shown on two upper panels in Figure 3.

The upper left panel plots the LL estimators with VS bandwidth while the upper right panel with the theoretically MSE minimizing variable bandwidth, both calculated by 0.01 intervals. The bottom two panels on the figure plot the size of the corresponding bandwidth at every point x. We find the discontinuous point at $x_* = 0$ on the upper right panel, which we do not see on the upper left panel. Although MSE minimizing variable bandwidths generate small vertical fluctuations in the LL estimator in most of the support, one discontinuous point at $x_* = 0$ on the upper right panel shows that the LL estimator is off greatly in the neighborhood. On the other hand, although VS bandwidths generate large fulctuations in the LL estimator in most of the support, it does not have a single discontinuous point.

Appendix 1

Set of assumptions

S 1 Random explanatory vectors $\mathbf{X}_{i\cdot} = (X_{i1}, ..., X_{ip})$ are i.i.d. with respect to i.

S 2 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 $\mathbf{X}_{j\cdot}$, $i \neq j$.

S 3 Pairs of random variables $(\mathbf{X}_{i\cdot}, Y_i)$ are independent with respect to i.

S 4 Column vectors in the covariate matrix (1) are not necessarily independent or orthogonal with respect to j.

A 1 The data $\mathbf{X}_{i\cdot}$'s are distributed as approximately multivariate normal so that the p dimensional bandwith matrix $\mathbf{H_X}$ is assumed to be diagonal, $\mathbf{H_X} = \text{diag}(h_{11}, h_{22}, ..., h_{pp})$.

A 2 $\mathbf{H_X} \to \mathbf{O}$ as $n \to \infty$.

A 3 $n|\mathbf{H_X}| \to \infty$ as $n \to \infty$.

A 4 The density of $\mathbf{X}$ is $0 < C_f \leq f_{\mathbf{X}}(\mathrm{x}) \leq C^f$ on bounded support I^p.

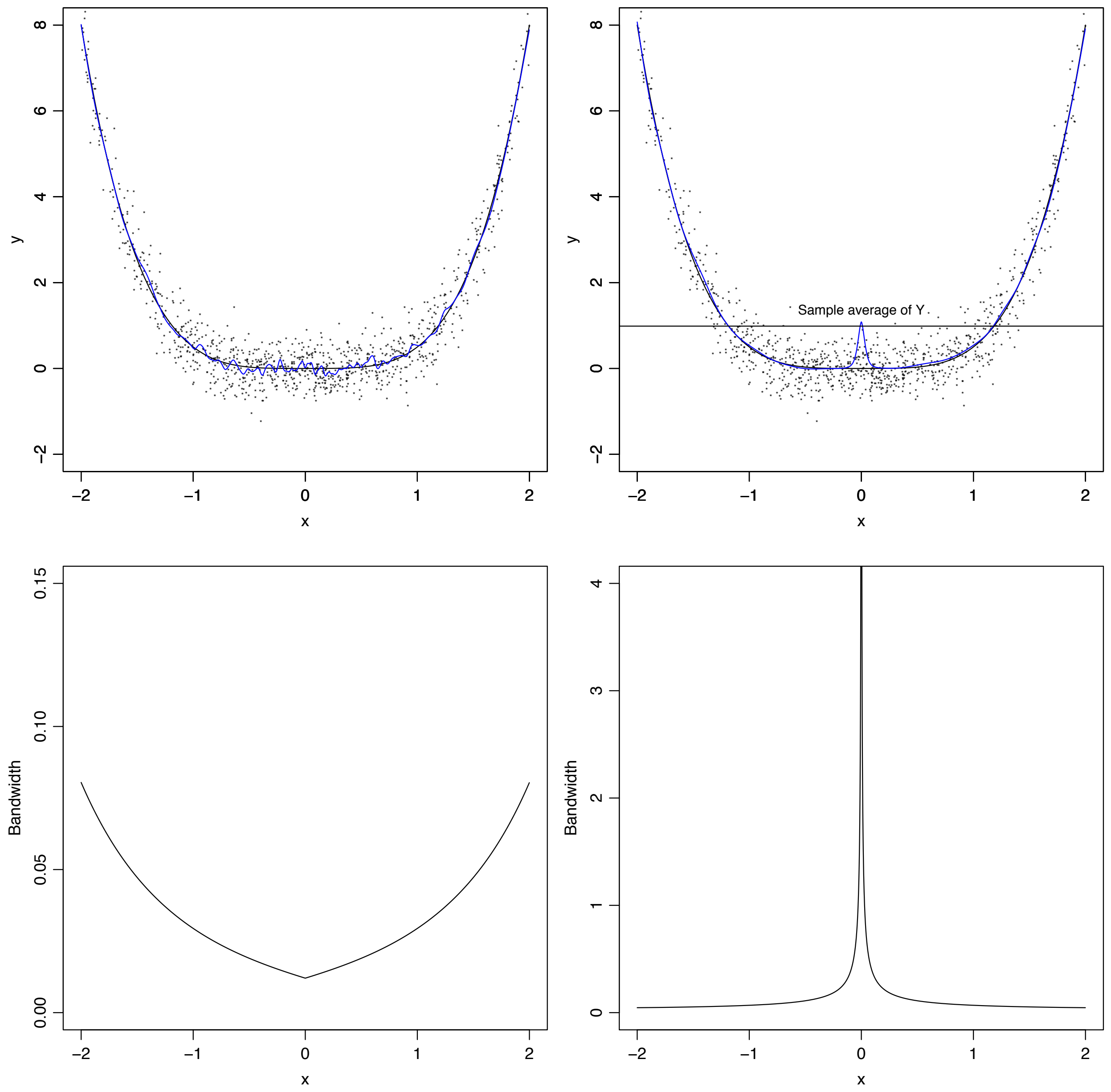

Figure 3: The upper left panel plots the LL estimators with VS bandwidth while the upper right with the MSE minimizing variable bandwidth, both by 0.01 intervals in the case of $m(x) = 0.5x^4$. The lower left and right panels plot the corresponding size of VS bandwidth and the MSE minimizing variable bandwidth at every point x respectively.

A 5 $f_{\mathbf{X}}(\mathbf{x})$ is bounded continuously twice differentiable with respect to x_i, $i = 1, ..., p$.

A 6 $m(\mathbf{x})$ is bounded continuously twice differenciable with respect to x_i, $i = 1, ..., p$.

A 7 $\sigma^2(\mathbf{x})$, $0 < C_{\sigma^2} \leq \sigma^2(\mathbf{x}) \leq C^{\sigma^2}$, is bounded continuously twice differentiable with respect to x_i, $i = 1, ..., p$.

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

(i) $K_{\mathbf{X}}(\mathbf{t})$ is symmetric and $\int \cdots \int K_{\mathbf{X}}(\mathbf{t}) d\mathbf{t} = 1$,
(ii) $\mu_2(K_{\mathbf{X}}) = \int \cdots \int \mathbf{t}\mathbf{t}^T K_{\mathbf{X}}(\mathbf{t}) d\mathbf{t} < \infty$,
(iii) $\int \cdots \int K_{\mathbf{X}}^2(\mathbf{t}) d\mathbf{t} < \infty$,
(iv) $\int \cdots \int |K_{\mathbf{X}}(\mathbf{t})| d\mathbf{t} < +\infty$,
(v) $|\mathbf{t}||K_{\mathbf{X}}(\mathbf{t})| \to 0$ as $|\mathbf{t}| \to \infty$,
(vi) $\sup |K_{\mathbf{X}}(\mathbf{t})| < \infty$,
(vii) $\int \int K_{\mathbf{X}}(t_i, t_j) t_i t_j dt_i dt_j = 0$, $\forall i, j = 1, ..., p, j \neq i$.

Appendix 2

We present an algorithm to estimate the multivariate LL estimater with the VS bandwidth matrix. For illustrative purpose, we consider bivariate situation.

Stage 1. Estimation of $f_{\mathbf{X}}(x_1, ..., x_p)$, $i = 1, ..., p$.
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 $\widehat{\mathbf{H_F}}$ is the estimator of bandwidth matrix donoted as

$$\mathbf{H_F} = \text{diag}(h_{f11}, h_{f22})$$

by the assumption **A 1** and $K_{X_1,X_2}(\,\cdot\,,\,\cdot\,)$ is bivariate Gaussian. Assuming **A 1**, we employ the Scott's rule (Scott 1992 p.152) 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 $\partial^2 m(x_1,...,x_p)/\partial x_i \partial x_j$, $i,j=1,...,p$, $i \neq j$.

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,

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

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

$$\hat{r}^2(X_{i1},X_{i2}) = (Y_i - \widehat{m^{LL}_{\widehat{\mathbf{H_M}}}}(X_{i1},X_{i2}))^2, \quad i=1,...,n,$$

where $\widehat{\mathbf{H_M}}$ is the estimator of the diagonal bandwidth matrix defined to be

$$\mathbf{H_M} = \mathrm{diag}(h_m,h_m). \tag{28}$$

The estimator of $\mathbf{H_M}$ is selected so as to minimize the cross-validation statistics in terms of $\widehat{h_m}$ written as,

$$CV(\widehat{h_m}) = \frac{1}{n}\sum_{i=1}^{n}\left[Y_i - \widehat{m^{LL}_{-i,\widehat{\mathbf{H_M}}}}(X_{i1},X_{i2})\right]^2, \tag{29}$$

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 of sample left out. Then, we construct residual-based variance estimator,

$$\widehat{\sigma^2_{\widehat{\mathbf{H_V}}}}(x_1, x_2) = \frac{\sum_{i=1}^n K_{X_1,X_2}\left((x_1 - X_{i1}, x_2 - X_{i2})\widehat{\mathbf{H_V}}^{-1}\right)\widehat{r}^2(X_{i1}, X_{i2})}{\sum_{i=1}^n K_{X_1,X_2}\left((x_1 - X_{i1}, x_2 - X_{i2})\widehat{\mathbf{H_V}}^{-1}\right)}, \tag{30}$$

where $\widehat{\mathbf{H_V}}$ is the estimator of the diagonal bandwidth matrix defined to be

$$\mathbf{H_V} = \text{diag}(h_v, h_v). \tag{31}$$

As an estimator of (31), we employ the following bandwidth that minimizes the cross-validation statistics with respect to $\widehat{h_v}$,

$$CV(\widehat{h_v}) = \frac{1}{n}\sum_{i=1}^n \left[\widehat{r}^2(X_{i1}, X_{i2}) - \widehat{\sigma^2_{-i,\widehat{\mathbf{H_V}}}}(X_{i1}, X_{i2})\right]^2, \tag{32}$$

where

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

is leave-one-out residual based estimator with its i-th element of sample left out.

The bandwidth minimized in terms of cross-validation statistics is equivalent to the one minimizing averaged squared error (henceforth ASE) on average. Also, ASE and MISE lead asymptotically to the same level of smoothing as in Marron and Härdle (1986). This is the reason to employ cross-validation statistics.

Remark 7. When $\sigma^2(x_1, x_2)$ is considered to be homoscedastic, we employ $[1/(n-2)]\sum_{i=1}^n \widehat{r}^2(X_{i1}, X_{i2})$ for the estimator of $\sigma^2(x_1, x_2)$ instead of "residual based" estimator.

Remark 8. At the end of Stage 3, we are able to estimate $\widehat{\eta_i^*}(x_1, x_2)$, $i = 1, 2$. If $\widehat{\alpha_1}(x_1, x_2) = \widehat{\alpha_2}(x_1, x_2) = 0$ happens at the point (x_1, x_2) as indicated in Remark 1, we set $\widehat{\eta_i^*}(x_1, x_2) =$

$1/2, i = 1, 2$. Similarly, if $\widehat{\alpha_1}(x_1, x_2) \neq 0$, $\widehat{\alpha_2}(x_1, x_2) = 0$, or $\widehat{\alpha_1}(x_1, x_2) = 0$, $\widehat{\alpha_2}(x_1, x_2) \neq 0$, happens at the point (x_1, x_2) as pointed out in Remark 3, we set $\widehat{\eta_1^*}(x_1, x_2) = 1, \widehat{\eta_2^*}(x_1, x_2) = 0$, or $\widehat{\eta_1^*}(x_1, x_2) = 0$, $\widehat{\eta_2^*}(x_1, x_2) = 1$, respectively.

Stage 4. Compute $\widehat{h_0^*}$.

Now that we have obtained $\widehat{f_{\widehat{\mathbf{H_F}}}}(x_1, x_2)$, $\widehat{\sigma^2_{\widehat{\mathbf{H_V}}}}(x_1, x_2)$, $\widehat{\alpha_i}(x_1, x_2)$, $\widehat{\eta_i^*}(x_1, x_2)$, $i = 1, 2$, in **Stages 1-3**, we can obtain the universal,

$$\widehat{h_0^*} = \left[\frac{[\int\int K^2(t_1, t_2)dt_1dt_2]}{\mu_2^2(K)\widehat{T_{VS}}(\widehat{\eta_1^*}(x_1, x_2), \widehat{\eta_2^*}(x_1, x_2))}\right]^{\frac{1}{6}} 2^{\frac{1}{6}} \cdot n^{-\frac{1}{6}},$$

by numerically integrate the function of the form,

$$\widehat{T_{VS}}(\widehat{\eta_1^*}(x_1, x_2), \widehat{\eta_2^*}(x_1, x_2)) = \int\int_{I^2} \widehat{f_{\widehat{\mathbf{H_F}}}}(x_1, x_2) \left[\sum_{i=1}^{2} \widehat{\alpha_i}(x_1, x_2) \left[\frac{\widehat{\sigma^2_{\widehat{\mathbf{H_V}}}}(x_1, x_2)}{\widehat{f_{\widehat{\mathbf{H_F}}}}(x_1, x_2)}\right]^{2\widehat{\eta_i^*}(x_1, x_2)}\right]^2 dx_1dx_2.$$

Stage 5. So far, one universal $\widehat{h_0^*}$ and, at every point (x_1, x_2), $\left[\widehat{\sigma^2_{\widehat{\mathbf{H_V}}}}(x_1, x_2)/\widehat{f_{\widehat{\mathbf{H_F}}}}(x_1, x_2)\right]$ and $\widehat{\eta_i^*}(x_1, x_2)$, $i = 1, 2$, are obtained. With the estimated VS bandwidth matrix,

$$\widehat{\mathbf{H_{VS}}}(x_1, x_2) = \widehat{h_0^*} \cdot \text{diag}\left(\left[\frac{\widehat{\sigma^2_{\widehat{\mathbf{H_V}}}}(x_1, x_2)}{\widehat{f_{\widehat{\mathbf{H_F}}}}(x_1, x_2)}\right]^{\widehat{\eta_1^*}(x_1, x_2)}, \left[\frac{\widehat{\sigma^2_{\widehat{\mathbf{H_V}}}}(x_1, x_2)}{\widehat{f_{\widehat{\mathbf{H_F}}}}(x_1, x_2)}\right]^{\widehat{\eta_2^*}(x_1, x_2)}\right) \cdot 2^{\frac{1}{6}} \cdot n^{-\frac{1}{6}}, \quad (33)$$

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

Acknowledgements

This research is supported in part by the Grant-in-Aid for Scientific Research (C)(2) 12680310, (C)(2) 16510103 and (B) 20310081 from the Japan Society for the Promotion of Science.

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.

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.

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.