数理解析研究所講究録
第 1860 巻 2013 年 73-90

On Variance-Stabilizing Multivariate Nonparametric Regression Estimation — A Comparison Between the Two Variance-Stabilizing Bandwidth Matrices

Kiheiji NISHIDA
General Education Center, Hyogo University of Health Siences

1 Introduction

It is well-known that a nonparametric regression estimator does not produce constant estimator variance over domain. To obtain a homoscedastic nonparametric regression estimators especially for a kernel regression estimator, a bandwidth matrix that is designed to stabilize variance should be introduced. In this paper, we give an overview of the homoscedastic nonparametric regression estimators and make a comparison between the two possible variance-stabilizing (henceforth VS) bandwidth matrices.

The locally linear estimator (henceforth the LL estimator) as presented by Ruppert and Wand (1994) is one of the well-known nonparametric regression estimators to explore the association between a set of stochastic covariates $\mathbf{X}=(X_1,...,X_p)$ and the response Y. Let us consider a $p+1$-row vector $(\mathbf{X}_{i\cdot},Y_i)$ of random variables, where $\mathbf{X}_{i\cdot}=(X_{i1},...,X_{ip})$ is i.i.d. with respect to i and its joint density function $f_{\mathbf{X}}(\mathbf{x})$ is away from zero on compact support $I^p\in R^p$. The vector $\mathbf{x}_{i\cdot}=(x_{i1},...,x_{ip})$, $i=1,...,n$, is the realization of $\mathbf{X}_{i\cdot}$. The n sample realizations of $(X_{i1},...,X_{ip})$ can be written as the covariate matrix $(\mathbf{x}_{\cdot1},\mathbf{x}_{\cdot2},...,\mathbf{x}_{\cdot p})$, where $\mathbf{x}_{\cdot j}=(x_{1j},x_{2j},...,x_{nj})^T$, $i=1,...,n$. Then, the response Y_i, $i=1,...,n$, is written as

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

where $m(\cdot)$ is $m:R^p\rightarrow R$ 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 $\mathbf{X}_{j\cdot}$, $i\neq j$, with their means and variances to be zero and $\sigma^2(\mathbf{x}_{i\cdot})$ respectively. 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 assumption of second order kernel in Ruppert and Wand (1994). Let $\mathbf{H}$ be a p-dimensional symmetric positive definite-bandwidth matrix. All the entries h_{ij} in $\mathbf{H}$ converge to 0 as $n\rightarrow\infty$ and $n|\mathbf{H}|\rightarrow\infty$ as $n\rightarrow\infty$. Then, the LL estimator of $m(\cdot)$ is given by the solution for β_0 minimizing,

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

where

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

$\mathbf{W}(\mathbf{x}) = \mathrm{diag}\left(K_{\mathbf{X}}((\mathbf{x}_{1\cdot} - \mathbf{x})\mathbf{H}^{-1}), ..., K_{\mathbf{X}}((\mathbf{x}_{n\cdot} - \mathbf{x})\mathbf{H}^{-1})\right)$ is the weight matrix, $\boldsymbol{\beta} = (\beta_0, \beta_1, ..., \beta_p)^T$ is the coefficient vector, $\boldsymbol{Y} = (Y_1, ..., Y_n)^T$ is the vector of responses with length n. Solving the minimization problem (1) with respect to 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 0 for all other entries. Then, the theoretical conditional variance of the LL estimator is written as

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

where $R(K_{\mathbf{X}}) = \int \cdots \int K_{\mathbf{X}}^2(\mathbf{t})d\mathbf{t}$. The term $\sigma^2(\mathbf{x})/f_{\mathbf{X}}(\mathbf{x})$ in the leading term of (2) represents the heteroscedasticity of the LL estimator. Similarly, the theoretical conditional bias for the LL estimator at $\mathbf{x}$ is known to be

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

where $\mu_2(K_{\mathbf{X}})$ is the variance of the kernel and $\nabla^2 m(\mathbf{x})$ is the Hessian matrix,

$$\nabla^2 m(\mathbf{x}) = \begin{pmatrix} \frac{\partial^2 m(\mathbf{x})}{\partial x_1 \partial x_1} & \cdots & \frac{\partial^2 m(\mathbf{x})}{\partial x_1 \partial x_p} \\ \vdots & \ddots & \vdots \\ \frac{\partial^2 m(\mathbf{x})}{\partial x_p \partial x_1} & \cdots & \frac{\partial^2 m(\mathbf{x})}{\partial x_p \partial x_p} \end{pmatrix} = \begin{pmatrix} \alpha_{11}(\mathbf{x}) & \dots & \alpha_{1p}(\mathbf{x}) \\ \vdots & \ddots & \vdots \\ \alpha_{1p}(\mathbf{x}) & \dots & \alpha_{pp}(\mathbf{x}) \end{pmatrix}.$$

To obtain homoscedastic LL estimator, it is necessary to set the determinant of the local variable bandwidth matrix $|\mathbf{H}(\mathbf{x})|$ to be $\sigma^2(\mathbf{x})/f_{\mathbf{X}}(\mathbf{x})$ at every locational point $\mathbf{x}$. One such bandwidth estimator appears in Fan and Gijbels (1992). In the paper, they employ the global variable bandwidth $\sigma^2(X_i)h_0/f_X(X_i)$ for the univariate LL estimator and assign different weight to each observation in the kernel by $K((x - X_i)f_X(X_i)/(\sigma^2(X_i)h_0))$. The parameter h_0 is a global parameter that should be determined to minimize AMISE (Asymptotic Mean Integrated Squared Error). Nishida and Kanazawa (2011) also proposes the variance-stabilizing local variable bandwidth for the univariate Nadaraya-Watson estimator (Nadaraya, 1964, 1965, 1970; Watson, 1964; Watson and Leadbetter, 1963) and make a comparison with the Mean Integrated Squared Error (MISE) minimizing fixed bandwidth. Since the two VS bandwidths are so designed as to minimize MISE (Mean Integrated Squared Error) among the class of VS bandwidths, they cannot, by its definition, outperform the MSE minimizing local variable bandwidth in terms of MISE. In this sense, Fan and Gijibels (1992) are critical to the VS bandwidth.

In multivariate setting, Nishida and Kanazawa (2013) proposes the VS diagonal bandwidth matrix for the p-variate LL estimator. The proposed VS bandwidth matrix is of the form

$$\mathbf{H_{VS}}(\mathbf{x}) = h_0 \cdot \operatorname{diag}\left(\left[\frac{\sigma^2(\mathbf{x})}{f_{\mathbf{X}}(\mathbf{x})}\right]^{\eta_{11}^{Diag}(\mathbf{x})}, ..., \left[\frac{\sigma^2(\mathbf{x})}{f_{\mathbf{X}}(\mathbf{x})}\right]^{\eta_{pp}^{Diag}(\mathbf{x})} \right),$$

$$\sum_{i=1}^{p} \eta_{ii}^{Diag}(\mathbf{x}) = 1, \tag{3}$$

$$-\infty < \eta_{ii}^{Diag}(\mathbf{x}) < \infty, \tag{4}$$

and the global parameter h_0 and the local parameters $\eta_{ii}(\mathbf{x})$'s, $i = 1, ..., p$, are optimized to minimize AMISE under the constraints (3) and (4). Then, if we denote $w_{\mathbf{X}}(\mathbf{x})$ to be a weighting function, the optimal h_0^* and $\eta_{ii}^{Diag,*}(\mathbf{x})$, $i = 1, ..., p$, are respectively given by

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

where

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

and

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

if $\alpha_{ii}(\mathbf{x}) > 0$, $i = 1, ..., p$, or $\alpha_{ii}(\mathbf{x}) < 0$, $i = 1, ..., p$. If $\alpha_{ii}(\mathbf{x}) = 0$, $i = 1, ..., p$, any set of values $\eta_{ii}^{Diag,*}(\mathbf{x})$ satisfying $\sum_{i=1}^{p} \eta_{ii}^{Diag,*}(\mathbf{x}) = 1$ are available. If $\alpha_{ii}(\mathbf{x})$'s, $i = 1, ..., p$, are not of the same sign when $p \geq 3$, the optimal set of parameters $\eta_{ii}^{Diag,*}(\mathbf{x})$, $i = 1, ..., p$, is given by any set of values satisfying

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

If $\alpha_{qq}(\mathbf{x}) = 0$, $\alpha_{-ii}(\mathbf{x})$'s, $i = 1, ..., p, i \neq q$, are non-zero, we consider the $p-1$ dimensional minimization problem with the q-th variable left out of the AMISE minimization problem. This proposed VS bandwidth matrix is called the VS diagonal bandwidth matrix.

The VS diagonal bandwidth matrix has an advantage that, under a sufficient condition,

$$\gamma^{\frac{4}{p}}(\mathbf{x}) \left[\sum_{i=1}^{p} \alpha_{ii}(\mathbf{x})\right]^2 = \text{Const.}, \quad \gamma(\mathbf{x}) = \frac{\sigma^2(\mathbf{x})}{\int_{I^p} \sigma^2(\mathbf{x}) d\mathbf{x}} \Bigg/ \frac{f_{\mathbf{X}}(\mathbf{x})}{\int_{I^p} f_{\mathbf{X}}(\mathbf{x}) d\mathbf{x}},$$

our proposed VS bandwidth outperforms the MSE minimizing local variable scalar bandwidth matrix (henceforth the MSE-minimizing scalar bandwidth matrix),

$$\mathbf{H}_{\mathbf{var}}(\mathbf{x}) \;=\; \left[\frac{R(K_{\mathbf{X}})\sigma^2(\mathbf{x})}{\mu_2^2(K_{\mathbf{X}})f_{\mathbf{X}}(\mathbf{x})\left[\sum_{i=1}^{p}\alpha_{ii}(\mathbf{x})\right]^2}\right]^{\frac{1}{p+4}} p^{\frac{1}{p+4}}\cdot n^{-\frac{1}{p+4}}\cdot \mathbf{I}_p, \qquad (5)$$

which minimizes AMSE at every $\mathbf{x}$ among the class of local variable scalar bandwidth matrices $\mathbf{H}_{\mathbf{var}}(\mathbf{x}) = h_{00}(\mathbf{x})\cdot\mathbf{I}_p$. This result reveals that the VS bandwidth can outperform the MSE-minimizing bandwidth matrix if the dimensionality p is greater than one.

However, the proposed VS diagonal bandwidth matrix may be inadequate under a complex data structure. It is because we put a zero value at each off-diagonal element of the VS matrix instead of the terms that would be necessary to estimate a complex regression function. If we employ a full-bandwidth matrix $\mathbf{H_2}$ in bivariate setting, the leading term of the squared bias is written as

$$\frac{\mu_2^2(K_{\mathbf{X}})}{4}\left[(h_{11}^2+h_{12}^2)\alpha_{11}(x_1,x_2)+2h_{12}(h_{11}+h_{22})\alpha_{12}(x_1,x_2)+(h_{22}^2+h_{12}^2)\alpha_{22}(x_1,x_2)\right]^2. \quad (6)$$

If $h_{12} = 0$, the term that contains $\alpha_{12}(x_1, x_2)$ in (6) disappears and information about the term is overlooked.

We also expect that the term h_{12} reflects the correlation between $\mathbf{X}_{.1}$ and $\mathbf{X}_{.2}$. This is conceivable from that the squared of the bandwidth matrix $\mathbf{H}^2$ crresponds to the variance-covariance matrix of the data $\mathbf{X}_{.i}$ when Gaussian kernel is employed. In bivariate setting, for example, the off-diagonal elements of $\mathbf{H_2^2}$ are $h_{12}(h_{11}+h_{22})$, and $\mathbf{H_2^2}$ has no correlation if $h_{12} = 0$.

In this sense, under the data such as the mixed derivative functions of $m(\mathbf{x})$ are not zero and / or the correlations between explanatory variables are observed, more flexible VS bandwidth matrix such as a full-bandwidth matrix is motivated. The VS full-bandwidth matrix in multivariate setting is of the form

$$\mathbf{H}_{\mathbf{VS}^{++}}(\mathbf{x}) = \begin{pmatrix} h_{11}^{Full}\left[\frac{\sigma^2(\mathbf{x})}{f_{\mathbf{X}}(\mathbf{x})}\right]^{\eta_{11}^{Full}(\mathbf{x})} & h_{12}^{Full}\left[\frac{\sigma^2(\mathbf{x})}{f_{\mathbf{X}}(\mathbf{x})}\right]^{\eta_{12}^{Full}(\mathbf{x})} & \cdots & h_{1p}^{Full}\left[\frac{\sigma^2(\mathbf{x})}{f_{\mathbf{X}}(\mathbf{x})}\right]^{\eta_{1p}^{Full}(\mathbf{x})} \\ h_{12}^{Full}\left[\frac{\sigma^2(\mathbf{x})}{f_{\mathbf{X}}(\mathbf{x})}\right]^{\eta_{12}^{Full}(\mathbf{x})} & h_{22}^{Full}\left[\frac{\sigma^2(\mathbf{x})}{f_{\mathbf{X}}(\mathbf{x})}\right]^{\eta_{22}^{Full}(\mathbf{x})} & \cdots & \\ \vdots & \vdots & \ddots & \\ h_{1p}^{Full}\left[\frac{\sigma^2(\mathbf{x})}{f_{\mathbf{X}}(\mathbf{x})}\right]^{\eta_{1p}^{Full}(\mathbf{x})} & h_{2p}^{Full}\left[\frac{\sigma^2(\mathbf{x})}{f_{\mathbf{X}}(\mathbf{x})}\right]^{\eta_{2p}^{Full}(\mathbf{x})} & \cdots & h_{pp}^{Full}\left[\frac{\sigma^2(\mathbf{x})}{f_{\mathbf{X}}(\mathbf{x})}\right]^{\eta_{pp}^{Full}(\mathbf{x})} \end{pmatrix}, \quad (7)$$

$$\sum_{s\in S_p}\operatorname{sgn}(s)\prod_{i=1}^{p}h_{i,s_i}^{Full} > 0, \quad h_{ij}^{Full} = h_{ji}^{Full}, \qquad (8)$$

$$\sum_{i=1}^{p}\eta_{i,s_i}^{Full}(\mathbf{x}) = 1, \quad \text{for all } s\in S_p, \quad \eta_{ij}^{Full}(\mathbf{x}) = \eta_{ji}^{Full}(\mathbf{x}), \qquad (9)$$

$$h_{ii}^{Full} > 0, \quad -\infty < \eta_{ij}^{Full}(\mathbf{x}) < \infty, \quad \text{for } i,\ j = 1,\ldots,p, \qquad (10)$$

where S_p is the set of all permutations $s = \{s_1, s_2, ..., s_p\}$ of the set $\{1, 2, ..., p\}$ and $\text{sgn}(s)$ denotes the signature of each s; it is $+1$ for even s and -1 for odd s. The conditions (8), (9) and (10) assure us the positive definiteness of the bandwidth matrix by Sylvester's criterion.

In bivariate setting, the matrix (7) is written as

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

where

$$h_{11}^{Full},\ h_{22}^{Full} > 0, \quad h_{11}^{Full} h_{22}^{Full} - (h_{12}^{Full})^2 > 0,$$
$$\eta_{11}^{Full}(\mathbf{x}) + \eta_{22}^{Full}(\mathbf{x}) = 1, \quad -\infty < \eta_{11}^{Full}(\mathbf{x}) < \infty.$$

To make the problem simpler, we additionally assume $h_{11}^{Full} = h_{22}^{Full}$ in (11) and obtain AMISE written as

$$\begin{aligned} & AMISE\left(m(x_1, x_2), \widehat{m_{\mathbf{H}_{\mathbf{VS}^{++}}}}(\mathbf{x})\right) \\ = \; & \frac{R(K_{\mathbf{X}})}{n\left[h_{11}^2 - h_{12}^2\right]} + \frac{\mu_2^2}{4} \cdot T_{Full}(h_{11}^{Full}, h_{12}^{Full}, \eta_{11}^{Full}(x_1, x_2)), \end{aligned} \tag{12}$$

where

$$\begin{aligned} & T_{Full}(h_{11}^{Full}, h_{12}^{Full}, \eta_{11}^{Full}(x_1, x_2)) \\ = \; & \int\int_{I^2} \left[\left[(h_{11}^{Full})^2 \left[\frac{\sigma^2(x_1, x_2)}{f_{X_1,X_2}(x_1, x_2)}\right]^{2\eta_{11}^{Full}(x_1,x_2)} + (h_{12}^{Full})^2 \left[\frac{\sigma^2(x_1, x_2)}{f_{X_1,X_2}(x_1, x_2)}\right]\right] \alpha_{11}(x_1, x_2)\right. \\ & + 2h_{12}^{Full} \left[h_{11}^{Full} \left[\frac{\sigma^2(x_1, x_2)}{f_{X_1,X_2}(x_1, x_2)}\right]^{\eta_{11}^{Full}(x_1,x_2)+\frac{1}{2}} + h_{11}^{Full} \left[\frac{\sigma^2(x_1, x_2)}{f_{X_1,X_2}(x_1, x_2)}\right]^{\frac{3}{2}-\eta_{11}^{Full}(x_1,x_2)}\right] \alpha_{12}(x_1, x_2) \\ & \left. + \left[(h_{11}^{Full})^2 \left[\frac{\sigma^2(x_1, x_2)}{f_{X_1,X_2}(x_1, x_2)}\right]^{2(1-\eta_{11}^{Full}(x_1,x_2))} + (h_{12}^{Full})^2 \left[\frac{\sigma^2(x_1, x_2)}{f_{X_1,X_2}(x_1, x_2)}\right]\right] \alpha_{22}(x_1, x_2)\right]^2 \\ \times \; & f_{X_1,X_2}(x_1, x_2) dx_1 dx_2. \end{aligned} \tag{13}$$

Even in a bivariate setting, it is hard to obtain the optimal parameters $h_{11}^{Full.*}$, $h_{12}^{Full,*}$, $\eta_{11}^{Full,*}(x_1, x_2)$ explicitly in terms of AMISE, so we have to resort to numerical calculation. If the domain is large, it is also practically inevitable to employ universal parameters η_{ii}^{Full}, $i = 1, 2$, instead of local ones, to reduce computational burden. In section 2, we mention under what situation the VS full-bandwidth matrix is advisable in terms of AMISE in bivariate setting.

Although the two VS bandwidth matrices are so designed as to stabilize the *asymptotic* variance of the LL estimator, we do not know to what degree they stabilize the variance when they are practically used for a complex data. Especially, we are interested in the cases where the mixed derivative of the true regression function is nonzero and/or the explanatory variables are correlated. We are also interested in the case where the sphering approach is not applicable, e.g. a multimodal density setting. To validate this, we run Monte-Carlo simulations with the theoretical VS bandwidth matrices in bivariate setting and present the results in Section 3. In the simulation, the MSE-minimizing local variable bandwidth in (5) is employed as a competitor for the heteroscedastic LL estimator. In section 2, we give some remarks on the VS-full bandwidth matrix and its estimator. In Section 4, we give Discussion.

2 On the VS full-bandwidth matrix

It is impossible to obtain the VS-full bandwidth matrix explicitly even in bivariate setting. To compute the VS-full bandwidth matrix numerically, we need to know the AMISE function has minimum values with respect to h_{11}^{Full} and h_{12}^{Full}, as well as η_{11}^{Full}. The following two remarks give us a sketch about the existence of minimum value of AMISE function when the VS full-bandwidth matrix is employed.

Remark 1. The function $AMISE(h_{11}^{Full}, h_{12}^{Full}, \eta_{11}^{Full})$ has at least one minimum value with respect to h_{11}^{Full} and h_{12}^{Full}. To know this, we expand (13) and obtain,

$$\begin{aligned}
&T_{Full}(h_{11}^{Full}, h_{12}^{Full}, \eta_{11}^{Full}) \\
&= (h_{11}^{Full})^4 \int\int_{I^2} \Big[V^{4\eta_{11}^{Full}}(x_1,x_2)\alpha_{11}^2(x_1,x_2) + V^{4(1-\eta_{11}^{Full})}(x_1,x_2)\alpha_{22}^2(x_1,x_2) \\
&\qquad +V^2(x_1,x_2)\alpha_{11}(x_1,x_2)\alpha_{22}(x_1,x_2)\Big] f_{X_1,X_2}(x_1,x_2)dx_1dx_2 \\
&+ (h_{12}^{Full})^4 \int\int_{I^2} V^2(x_1,x_2)\Big[\alpha_{11}^2(x_1,x_2) + \alpha_{22}^2(x_1,x_2) + \alpha_{11}(x_1,x_2)\alpha_{22}(x_1,x_2)\Big] f_{X_1,X_2}(x_1,x_2)dx_1dx_2 \\
&+ 2(h_{11}^{Full})^3(h_{12}^{Full}) \int\int_{I^2} \alpha_{12}(x_1,x_2)\Big[V^{\eta_{11}^{Full}+\frac{1}{2}}(x_1,x_2) + V^{\frac{3}{2}-\eta_{11}^{Full}}(x_1,x_2)\Big] \\
&\qquad \times\Big[V^{2\eta_{11}^{Full}}(x_1,x_2)\alpha_{11}(x_1,x_2) + V^{2(1-\eta_{11}^{Full})}(x_1,x_2)\alpha_{22}(x_1,x_2)\Big] f_{X_1,X_2}(x_1,x_2)dx_1dx_2 \\
&+ 2(h_{11}^{Full})(h_{12}^{Full})^3 \int\int_{I^2} \alpha_{12}(x_1,x_2)V(x_1,x_2)\Big[V^{\eta_{11}^{Full}+\frac{1}{2}}(x_1,x_2) + V^{\frac{3}{2}-\eta_{11}^{Full}}(x_1,x_2)\Big] \\
&\qquad \times\Big[\alpha_{11}(x_1,x_2) + \alpha_{22}(x_1,x_2)\Big] f_{X_1,X_2}(x_1,x_2)dx_1dx_2 \\
&+ (h_{11}^{Full})^2(h_{12}^{Full})^2 \int\int_{I^2} \Big[4\alpha_{12}^2(x_1,x_2)\Big[V^{\eta_{11}^{Full}+\frac{1}{2}}(x_1,x_2) + V^{\frac{3}{2}-\eta_{11}^{Full}}(x_1,x_2)\Big]^2 \\
&\qquad +\alpha_{11}(x_1,x_2)\alpha_{22}(x_1,x_2)\Big[V^{2\eta_{11}^{Full}+1}(x_1,x_2) + V^{3-2\eta_{11}^{Full}}(x_1,x_2)\Big] \\
&\qquad +2V^{2\eta_{11}^{Full}+1}(x_1,x_2)\alpha_{11}^2(x_1,x_2) + 2V^{3-2\eta_{11}^{Full}}(x_1,x_2)\alpha_{22}^2(x_1,x_2)\Big] f_{X_1,X_2}(x_1,x_2)dx_1dx_2, \qquad (14)
\end{aligned}$$

where $V(x_1, x_2) = \sigma^2(x_1, x_2)/f_{X_1,X_2}(x_1, x_2)$. From $h_{11}^{Full} > F_{12}^{Full}$, we know that the first term is of the greatest order of magnitude in terms of h_{11}^{Full} and h_{12}^{Full}. We also know

$$\begin{aligned}
&(h_{12}^{Full})^4 \int\int_{I^2} \Big[V^{4\eta_{11}^{Full}}(x_1, x_2)\alpha_{11}^2(x_1, x_2) + V^{4(1-\eta_{11}^{Full})}(x_1, x_2)\alpha_{22}^2(x_1, x_2) \\
&\qquad +V^2(x_1, x_2)\alpha_{11}(x_1, x_2)\alpha_{22}(x_1, x_2)\Big] f_{X_1,X_2}(x_1, x_2)dx_1dx_2 \\
&\quad \geq (h_{11}^{Full})^4 \int\int_{I^2} V^2(x_1, x_2)\Big[2\left|\alpha_{11}(x_1, x_2)\alpha_{22}(x_1, x_2)\right| + \alpha_{11}(x_1, x_2)\alpha_{22}(x_1, x_2)\Big] \\
&\qquad \times f_{X_1,X_2}(x_1, x_2)dx_1dx_2 \geq 0.
\end{aligned}$$

Thus, we know $\lim_{||\mathbf{h}|| \to \infty} T_{Full}(h_{11}^{Full}, h_{12}^{Full}, \eta_{11}^{Full}) = \infty$ and $\lim_{||\mathbf{h}|| \to 0} T_{Full}(h_{11}^{Full}, h_{12}^{Full}, \eta_{11}^{Full}) = 0$, where $\mathbf{h} = (h_{11}^{Full}, h_{12}^{Full})$, $h_{11}^{Full} > h_{12}^{Full} > 0$. On the other hand, it is verified that $\lim_{||\mathbf{h}|| \to \infty} R(K_{\mathbf{X}})/\left[n\left[(h_{11}^{Full})^2 - (h_{12}^{Full})^2\right]\right] = 0$ and $\lim_{||\mathbf{h}|| \to 0} R(K_{\mathbf{X}})/\left[n\left[(h_{11}^{Full})^2 - (h_{12}^{Full})^2\right]\right] = \infty$. Since $AMISE(h_{11}^{Full}, h_{12}^{Full}, \eta_{11}^{Full})$ is bounded below by zero, there exists at least one minimum value with respect to h_{11}^{Full} and h_{12}^{Full}.

Remark 2. There exists the optimal η_{11} that minimizes $T_{Full}(h_{11}^{Full}, h_{12}^{Full}, \eta_{11}^{Full})$. Without loss of generality, we assume $[\sigma^2(x_1, x_2)/f_{X_1,X_2}(x_1, x_2)] > 1$. Then, we can choose a constant term v_0 that satisfies

$$v_0 \left[\frac{\sigma^2(x_1, x_2)}{f_{X_1,X_2}(x_1, x_2)}\right]^{\eta_{11}^{Full}} < T_{Full}(h_{11}^0, h_{12}^0, \eta_{11}^{Full}), \tag{15}$$

where h_{11}^0 and h_{12}^0 are arbitrary constants. Since $\lim_{\eta \to \infty} v_0 \left[\sigma^2(x_1, x_2)/f_{X_1,X_2}(x_1, x_2)\right]^{\eta_{11}^{Full}} = \infty$, $T_{Full}(h_{11}^0, h_{12}^0, \eta_{11}^{Full})$ also goes to ∞, as $\eta_{11}^{Full} \to \infty$. On the other hand, we can choose another constant term v_1 that satisfies

$$v_1 \left[\frac{\sigma^2(x_1, x_2)}{f_{X_1,X_2}(x_1, x_2)}\right]^{-\eta_{11}^{Full}} < T_{Full}(h_{11}^0, h_{12}^0, \eta_{11}^{Full}). \tag{16}$$

Since $\lim_{\eta \to -\infty} v_1 \left[\sigma^2(x_1, x_2)/f_{X_1,X_2}(x_1, x_2)\right]^{-\eta_{11}^{Full}} = \infty$, $T_{Full}(h_{11}^0, h_{12}^0, \eta_{11}^{Full})$ also goes to ∞, as $\eta_{11}^{Full} \to -\infty$. Since the term $T_{Full}(h_{11}^{Full}, h_{12}^{Full}, \eta_{11}^{Full})$ is bounded below, we notice that there exists at least one $\eta_{11}^{Full,*}$.

We also present a sufficient condition under which the bandwidth parameter h_{12}^{Full} should be zero in terms of AMISE in bivariate setting. This condition claims that the VS diagonal bandwidth matrix is advisable over the VS full-bandwidth matrix in terms of AMISE under the situation.

Proposition 1. *In bivariate setting, the parameter h_{12}^{Full} in (11) should be zero to minimize AMISE under a sufficient condition,*

$$\begin{aligned} & \alpha_{11}(x_1,x_2) > 0, \quad \alpha_{12}(x_1,x_2) \geq 0, \quad \alpha_{22}(x_1,x_2) > 0, \\ \textit{or} \quad & \alpha_{11}(x_1,x_2) < 0, \quad \alpha_{12}(x_1,x_2) \leq 0, \quad \alpha_{22}(x_1,x_2) < 0, \end{aligned} \tag{17}$$

over the domain.

Proof. Let h_{11}^{Full} be fixed. The $AMISE(h_{11}^{Full}, h_{12}^{Full}, \eta_{11}^{Full})$ is minimized at $h_{12}^{Full} = 0$ if

$$\left.\frac{\partial AMISE(h_{12}^{Full})}{\partial(h_{12}^{Full})}\right|_{h_{12}^{Full}=0} > 0, \tag{18}$$

and

$$\frac{\partial^2 AMISE(h_{12}^{Full})}{\partial(h_{12}^{Full})^2} > 0, \tag{19}$$

on the support $0 \leq h_{12}^{Full} < h_{11}^{Full}$. Since the first and the second derivatives with respect to h_{12}^{Full} of the variance part in (12) are always positive, the signs of (18) and (19) are determined by all the signs of $\alpha_{11}(x_1,x_2)\alpha_{12}(x_1,x_2)$, $\alpha_{12}(x_1,x_2)\alpha_{22}(x_1,x_2)$ and $\alpha_{11}(x_1,x_2)\alpha_{22}(x_1,x_2)$, all of which appear in the first and the second partial derivatives with respect to h_{12}^{Full} of the function (14). As long as the condition (17) is satisfied, all the signs of $\alpha_{11}(x_1,x_2)\alpha_{12}(x_1,x_2)$, $\alpha_{12}(x_1,x_2)\alpha_{22}(x_1,x_2)$ and $\alpha_{11}(x_1,x_2)\alpha_{22}(x_1,x_2)$ are positive and the $AMISE(h_{12}^{Full})$ is minimized at $h_{12}^{Full} = 0$. □

A sketch on the estimator of the VS full-bandwidth matrix

Nishida and Kanazawa (2013) proposes an estimator of the VS diagonal bandwidth matrix. The idea is to estimate $f_{\mathbf{X}}(\mathbf{x})$, $\sigma^2(\mathbf{x})$, $\partial^2 m(\mathbf{x})/\partial \mathbf{x}^2$ and plug these estimators into the original VS diagonal bandwidth matrix. Nishida and Kanazawa (2013) employs the residual based estimator in Fan and Yao (1998) as an estimator of $\sigma^2(\mathbf{x})$, kernel density estimator as an estimator of $f_{\mathbf{X}}(\mathbf{x})$ and quartic polinomial fit for $\alpha_{ii}(\mathbf{x})$. For the bandwidths of these estimators, cross-validation statistics are employed. Since quartic polinomial fit for $\alpha_{ii}(\mathbf{x})$ is inconsistent estimator unless the true regression function is polynomial function, this estimator is ROT (Rule of Thumb). Although Nishida and Kanazawa (2013) points out that the VS regression estimation is a difficult task because of the difficulty in estimating $\sigma^2(\mathbf{x})$ and $\alpha_{ii}(\mathbf{x})$, some pieces of evidence that the proposed estimator produces homoscedastic nonparametric regression estimator is presented.

The problem is even more difficult when it comes to estimating the VS full-bandwidth matrix. Since it is impossible to obtain a theoretical bandwidth matrix explicitly, either plug-in (PI) approach or ROT is no longer accessible. Yang and Tschernig (1999) encounters the same difficulty to propose the estimator of the MISE-minimizing diagonal bandwidth matrix for the multivariate LL estimator. In their paper, the optimal bandwidth matrix in multivariate setting cannnot be obtained explicitly so they estimate the AMISE that depends on $\widehat{A}(\sigma^2(\cdot))$ and $\widehat{B}_{ij}(m(\cdot))$, where

$$A(\sigma^2(\cdot)) = \int \cdots \int_{I^p} w_{\mathbf{X}}(\mathbf{x})\sigma^2(\mathbf{x})d\mathbf{x},$$

$$B_{ij}(m(\cdot)) = \int \cdots \int_{I^p} w_{\mathbf{X}}(\mathbf{x})\alpha_{ii}(\mathbf{x})\alpha_{jj}(\mathbf{x})d\mathbf{x}, \quad i,j = 1, ..., p,$$

and minimize the AMISE estimated by numerical calculation in terms of $h_{11}, h_{22}, ..., h_{pp}$.

To obtain $\widehat{A}(\sigma^2(\cdot))$ and $\widehat{B}_{ij}(m(\cdot))$, they employ two ways, ROT and PI approahes. For ROT approach, the use of a quartic Taylor expansion is employed as in Ruppert et.al. (1995). They separate the data into equalized blocks and use a quartic Taylor expansion on each block. If we denote the number of blocks in one direction, say j, to be N_j, the total number of blocks in the domain is $N = \prod_{j=1}^{p} N_j$. To determine the optimal N^*, they employ Mallow's C_p criterion,

$$C_p(\mathbf{N}) = \frac{RSS(\mathbf{N})\{n - k(p)\lfloor \frac{n}{4k(p)} \rfloor\}}{\min_{\mathbf{N}} RSS(\mathbf{N})} - (n - 2k(p)N),$$

where $RSS(\mathbf{N})$ denotes the residual sum of squares based on the quartic fit with blocking $\mathbf{N} = (N_1, N_2, ..., N_p)$,

$$k(p) = 1 + \sum_{i=1}^{4} \binom{p+i-1}{i}$$

is the maximum number of parameters in one block. Then, the corresponding estimator of $\widehat{B}_{ij}(m(\cdot))$ is the estimate of error variance: residual sum of squares of the function $\widetilde{m_{ROT,\mathbf{N}^*}}(\mathbf{x})$, a function estimated by a quartic Taylor expansion, divided by the number of degrees of freedom. The estimator of $\widehat{B_{ij}}(m(\cdot))$ is the sample average of $[\partial^2 \widehat{m_{ROT,\mathbf{N}^*}}(\mathbf{x})/\partial x_1^2]\,[\partial^2 \widehat{m_{ROT,\mathbf{N}^*}}(\mathbf{x})/\partial x_2^2]$ weighted by $\widehat{f_{\mathbf{X}}}(\mathbf{x})$.

For PI approach, Yang and Tschernig (1999) estimates the second derivative of the true regression function via partial local cubic estimator, with most cross terms left-out of full local cubic estimator, given by

$$\widehat{\alpha_{jj}}(\mathbf{x}) = (2!)\mathbf{e_j}\left[\mathbf{D_j}^T(\mathbf{x})\mathbf{W}(\mathbf{x})\mathbf{D_j}(\mathbf{x})\right]^{-1}\left[\mathbf{D_j}^T(\mathbf{x})\mathbf{W}(\mathbf{x})\mathbf{Y}\right],$$

where

$$\mathbf{D_j}(\mathbf{x}) = \Big(\mathbf{D_{j,1}}(\mathbf{x}), \mathbf{D_{j,2}}(\mathbf{x}), \mathbf{D_{j,3}}(\mathbf{x}), \mathbf{D_{j,4}}(\mathbf{x}), \mathbf{D_{j,5}}(\mathbf{x}), \mathbf{D_{j,6}}(\mathbf{x}), \mathbf{D_{j,7}}(\mathbf{x})\Big),$$

$$
\begin{aligned}
\mathbf{D_{j,1}(x)} &= \left\{1\right\}, \quad (n \times 1), \\
\mathbf{D_{j,2}(x)} &= \left\{(x_{is} - x_s)\right\}_{i=1,\ldots,n,\ s=1,\ldots,p}, \\
\mathbf{D_{j,3}(x)} &= \left\{(x_{is} - x_s)(x_{ij} - x_j)\right\}_{i=1,\ldots,n,\ s=1,\ldots,p,\ s\neq j}, \\
\mathbf{D_{j,4}(x)} &= \left\{(x_{is} - x_s)^2\right\}_{i=1,\ldots,n,\ s=1,\ldots,p}, \\
\mathbf{D_{j,5}(x)} &= \left\{(x_{is} - x_s)(x_{ij} - x_j)^2\right\}_{i=1,\ldots,n,\ s=1,\ldots,p,\ s\neq j}, \\
\mathbf{D_{j,6}(x)} &= \left\{(x_{is} - x_s)^2(x_{ij} - x_j)\right\}_{i=1,\ldots,n,\ s=1,\ldots,p,\ s\neq j}, \\
\mathbf{D_{j,7}(x)} &= \left\{(x_{ij} - x_j)^3\right\}_{i=1,\ldots,n},
\end{aligned}
$$

and $\mathbf{e_j}$ is a $1 \times (5p-1)$ row vector with 1 as the $2p+j$ $(= 1+p+p-1+j)$-th entry 0 for the other entries. To estimate the bandwidths for partial local cubic regression estimator, they derive the asymptotic bias and variance of the estimator and employ ROT approach.

In our setting, the similar approaches to estimate AMISE directly in Yang and Tschernig (1999) may be applicable. To estimate $T_{Full}(h_{11}^{Full}, h_{22}^{Full}, \eta_{11}^{Full})$, we need to estimate the mixed derivative function of $m(\mathbf{x})$ that appears in (13). If we employ partial local cubic estimator, the estimator of the mixed derivative function of $m(\mathbf{x})$ with respect to the variables x_j and x_k is given by

$$
\widehat{\alpha_{jk}}(\mathbf{x}) = \mathbf{e_{jk}} \left[\mathbf{D_j}^T(\mathbf{x})\mathbf{W}(\mathbf{x})\mathbf{D_j}(\mathbf{x})\right]^{-1} \left[\mathbf{D_j}^T(\mathbf{x})\mathbf{W}(\mathbf{x})\mathbf{Y}\right], \tag{20}
$$

where $\mathbf{e_{jk}}$ is a $1 \times (5p-1)$ row vector of 0s whose $(1+p+k)$ element is 1. To obtain the bandwidth matrix of the estimator (20), further study on the asymptotic bias and variance of (20) is needed.

3 Monte-Carlo Simulations with theoretical bandwidth matrices

Wand and Jones (1993) gives an extensive study about the choice of bandwidth matrix in bivariate density estimation. In the study, they employ scalar, diagonal and full bandwidth matrices for kernel density estimator. Then, they set up extensive numbers of simulation cases and calculate AMISE's theoretically for each simulation case. Following the practice in Wand and Jones (1993), we set up several simulation cases and run Monte-Carlo Simulations with theoretical bandwidth matrices in bivariate setting to know performances of the two VS bandwidth matrices. The procedure is as follows.

1. Generate (X_{i1}, X_{i2}) of sample size n distributed as $f_{X_1,X_2}(x_1, x_2)$. Generate $U_i|\{\mathbf{X}_{i\cdot} =$

$\mathbf{x}_{i\cdot}\}$ of sample size n distributed as $N(0,\sigma^2(x_{i1},x_{i2}))$. Obtain $(\mathbf{X}_{i\cdot},Y_i)$ of sample size n, where $Y_i = m(x_{i1},x_{i2}) + U_i|\{\mathbf{X}_i = \mathbf{x}_i\}$.

2. Construct LL estimators $\widehat{m_{\mathbf{H}_{\mathbf{VS}}}}(\mathbf{x})$, $\widehat{m_{\mathbf{H}_{\mathbf{VS++}}}}(\mathbf{x})$ and $\widehat{m_{\mathbf{H}_{\mathbf{var}}}}(\mathbf{x})$ at every grid point defined on the domain. The number of grid points in the domain is $G = 10,000$.
3. Repeat 1 $\sim$ 3 $M = 100$ times.
4. Obtain the estimator of MISE given by

$$\widehat{MISE}(m(x_1,x_2),\widehat{m_{\mathbf{H}}}(x_1,x_2)) = \frac{1}{M}\sum_{\mathbf{t}=1}^{M}\left[\int\int_{I^2} f_{\mathbf{X}}(x_1,x_2)\left[m(x_1,x_2)-\widehat{m_{\mathbf{H}}}^{(\mathbf{t})}(x_1,x_2)\right]^2 dx_1dx_2\right],$$

where $\widehat{m_{\mathbf{H}}}^{(\mathbf{t})}(x_1,x_2)$ is the LL estimator calculated $(\mathbf{t})$ th generated sample of size n.

5. At every grid point, calculate the sample variances of $\widehat{m_{\mathbf{H}_{\mathbf{VS}}}}(\mathbf{x})$, $\widehat{m_{\mathbf{H}_{\mathbf{VS++}}}}(\mathbf{x})$ and $\widehat{m_{\mathbf{H}_{\mathbf{var}}}}(\mathbf{x})$ that are respectively calculated $M = 100$ times in 1 $\sim$ 3 for $n = 5,000$.
6. As measures to check if the variance is stabilized, we calculate the means, the standard deviations and the Gini-coefficients of the sample variances of $\widehat{m_{\mathbf{H}_{\mathbf{VS}}}}(\mathbf{x})$, $\widehat{m_{\mathbf{H}_{\mathbf{VS++}}}}(\mathbf{x})$ and $\widehat{m_{\mathbf{H}_{\mathbf{var}}}}(\mathbf{x})$ calculated at every grid point in 5.

In the simulation cases to be presented, the domain and the grid points are respectively defined to be $[-0.5,0.5]\times[-0.5,0.5]$ and $(-0.495+0.01\times(i-1),-0.495+0.01\times(j-1))$, $i=1,...,100$, $j=1,...,100$. The conditional variance function is $\sigma^2(x_1,x_2) = 0.5+0.25x_1^2+0.25x_2^2$ as illustrated in Figure 2. As densities, we employ a normal density $f_{\mathbf{X}}(x_1,x_2;\mu_1=0.0,\mu_2=0.0,\sigma_1^2=0.25^2,\sigma_2^2=0.25^2,\rho)$ truncated on $[-0.5,0.5]^2$ with its correlation coefficient $\rho = 0.0$, 0.25, 0.5 and 0.75. We also employ a bimodal density, a mixture of the two normal densities $N((0.25,0),\mathrm{diag}(0.15^2,0.15^2))$ and $N((-0.25,0.0),\mathrm{diag}(0.15^2,0.15^2))$ with its mixing ratio even. Figure 1 illustrates the five distributions of the sample (X_{i1},X_{i2}), $i=1,...,5,000$. Then, we suppose the following three true regression functions denoted as simulation 1, 2 and 3 respectively. The perspective plots and contour plots of the simulation cases are given in Figure 3 and Figure 4 respectively. As kernel, we employ bivariate Gaussian kernel.

In general, VS nonparameteric regression estimation requires a large sample size data as stated in Nishida and Kanazawa (2013). We consider that the sample size $5,000$ is enough to know the behavior of the VS bandwidth matrices.

Simulation 1. The true regression function is $m(x_1,x_2) = -2x_1^2 - x_2^2$. In this setup, we intend that $\alpha_{12}(x_1,x_2)$ is zero over the domain.

Simulation 2. The true regression function is $m(x_1,x_2) = -2x_1^2 + 1.5x_1x_2 - x_2^2$. In this setup, we intend that $\alpha_{12}(x_1,x_2)$ is nonzero constant over the domain.

Simulation 3. The true regression function is $m(x_1,x_2) = \sin(3x_1)\cos(3x_2)$. In this setup, we intend that $\alpha_{12}(x_1,x_2)$, which is $-9\cos(3x_1)\sin(3x_2)$, varies over the domain.

We show the result. The theoretical values of the parameteres in the two VS bandwidth matrices are given in Table 1. The result of the simulations, Gini-coefficients, MISE

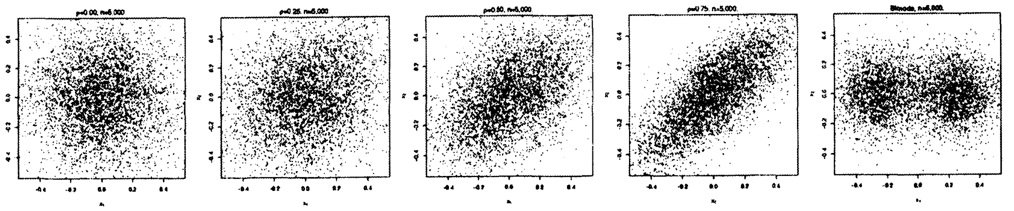

Figure 1: Distribution of the sample (X_{i1}, X_{i2}).

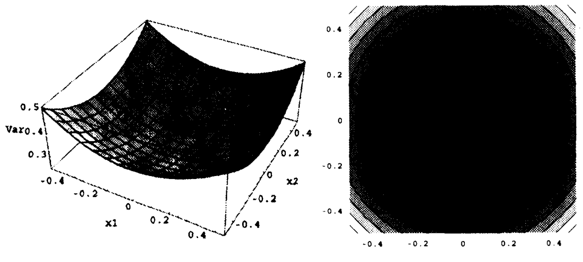

Figure 2: True conditional variance function : $\sigma^2(x_1, x_2) = 0.25 + 0.5x_1^2 + 0.5x_2^2$.

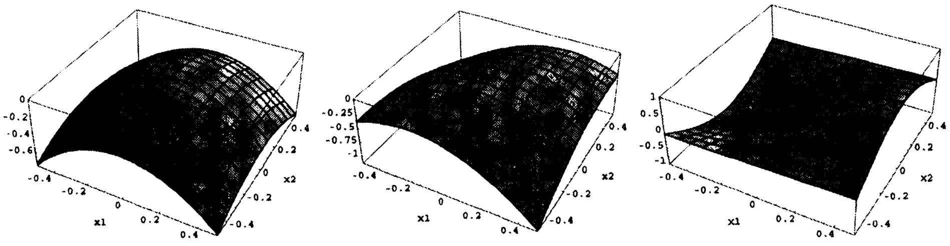

Figure 3: Perspective plots : Left=Simulation 1, Center=Simulation 2, Right=Simulation 3.

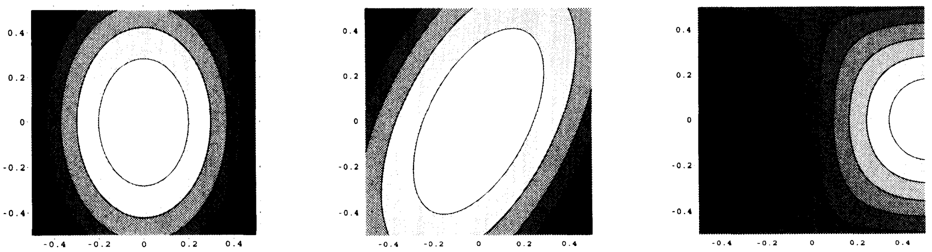

Figure 4: Contour plots : Left=Simulation 1, Center=Simulation 2, Right=Simulation 3.

estimated and standard deviation of sample variances, are given in Table 2. Figure 5 summarizes Table 2. Although it is natural that the three simulation cases do not represent all the data to happen, the result gives us some points of interest.

First, we examine the result of the theoretical bandwidth matrices. From Table 1, we notice that the size of bandwidth tends to diminish as the correlation coefficient ρ increases from 0.0 to 0.75. It seems that the smaller bandwidth is assigned when the data is highly correlated. The comparison between simulation 1 and 2 also gives us an interesting point of view. In simulation 1, the sign of the second derivative functions are $\alpha_{11}(x_1, x_2) < 0$ and $\alpha_{22}(x_1, x_2) < 0$, whereas in simulation 2, $\alpha_{11}(x_1, x_2) < 0$, $\alpha_{22}(x_1, x_2) < 0$ and $\alpha_{12}(x_1, x_2) > 0$. As a result, $h_{12}^{Full,*}$ comes out to be zero in simulation 1 whereas nonzero in simulation 2. It seems that the parameter $h_{12}^{Full,*}$ serves as an adjustment to control the impact of the mixed derivetive on AMISE. We also notice that the size of h_0^* and $h_{11}^{Full,*}$ are similar each other in simulation 1 whereas in simulation 2 dissimilar. It is because the mixed derivative of $m(\cdot)$ is zero in simulation 1 so there is no difference between the VS diagonal and the VS full-bandwidth. As for the bimodal density setting, we cannot find clear-cut features in the theoretical bandwidth matrices.

Second, we examine the achevement of variance-stabilization. From Table 2 as well as Figure 5, we find a clear result that either the VS diagonal or the VS full-bandwidth matrix outperforms the MSE-minimizing bandwidth matrix in terms of Gini-coefficients when ρ ranges from 0 to 0.75. This is a convincing evidence that either of the two VS bandwidth matrices can attain the variance-stabilization if the parameters in bandwidth are well-estimated. We also notice that the Gini-coefficients of the VS bandwidth matrices tend to increase as ρ increases form 0 to 0.75. The Gini-coefficients of the MSE-minimizing bandwidth matrix, on the other hand, tends to diminish as ρ increases. It seems that the MSE-minimizing bandwidth matrix tends to perform better than the two VS bandwidth matrices in terms of Gini-coefficient when the data is highly correlated. Similarly, we also notice that the VS diagonal bandwidth matrix tends to perform better than the VS-full bandwidth matrix in terms of Gini-coeff.. It is because the VS diagonal bandwidth matrix adjusts the size of $\eta_{ii}^{Full,*}(\mathbf{x})$ locally whereas the VS full-bandwidth matrix in our setting does $\eta_{ii}^{Full,*}$ globally. As for the bimodal density setting, the VS full-bandwidth matrix shows a good performance in simulation 3 in terms of Gini coeff., a piece of evidence that the VS-full bandwidth matrix is advisable in a multimodal density setting to achieve homoscedasticity.

Third, we examine the MISE estimated. From Table 2, we observe that the MISE's estimated diminishes as the correlation coefficient ρ increases from 0.0 to 0.75 in simulation 2 and 3 for all the three bandwidth matrices. On the other hand, the MISE's estimated tends to increase as the correlation coefficient ρ increases from 0.0 to 0.75 in simulation 1. It seems that the MISE tends to diminish as the correlation ρ increases from 0.0 to 0.75 when the mixed derivative of $m(\cdot)$ is nonzero over the domain. We also observe that the VS diagonal and full-bandwidth matrices can, in many cases, outperform the MSE-minimizing bandwidth matrix in terms of MISE estimated. This result supports the assertion in Nishida and Kanazawa (2013). As for the bimodal setting, it seems that the VS bandwidth matrices do not produce good results for all the simulation cases.

		$\rho = 0.00$	$\rho = 0.25$	$\rho = 0.50$	$\rho = 0.75$	*Bimode*
Simulation 1.	h_0^*	0.5049	0.4782	0.3719	0.1303	0.3736
	$h_{11}^{Full,*}$	0.4966	0.4726	0.3705	0.1302	0.3725
	$h_{12}^{Full,*}$	0.0	0.0	0.0	0.0	0.0
	$\eta_{11}^{Full,*}$	0.4433	0.4391	0.4621	0.4855	0.4492
Simulation 2.	h_0^*	0.5049	0.4782	0.3719	0.1303	0.3736
	$h_{11}^{Full,*}$	0.5417	0.5168	0.4093	0.1447	0.4093
	$h_{12}^{Full,*}$	0.1447	0.1405	0.1157	0.0413	0.1157
	$\eta_{11}^{Full,*}$	0.4407	0.4350	0.4590	0.4844	0.4451
Simulation 3.	h_0^*	0.4522	0.4370	0.3686	0.1537	0.4244
	$h_{11}^{Full,*}$	0.4523	0.4391	0.3791	0.1674	0.0980
	$h_{12}^{Full,*}$	0.0	0.0231	0.0475	0.0372	0.0
	$\eta_{11}^{Full,*}$	0.5	0.5	0.5	0.5	0.5

Table 1: Theoretical values of the parameters in the two VS bandwidth matrices. The sample size is set to be unity in this table.

		$\rho = 0.00$	$\rho = 0.25$	$\rho = 0.50$	$\rho = 0.75$	*Bimode*
Simulation 1.						
VS.Diag.	Gini coeff.	0.3155	0.3105	0.3502	0.4098	0.9511
	Var. Std.	0.0013	0.0014	0.0026	0.0125	31.0829
	$\widehat{MISE}$	0.1497	0.1500	0.1498	0.1765	14.1006
VS.Full.	Gini coeff.	0.3444	0.3467	0.3674	0.4105	0.9262
	Var. Std.	0.0012	0.0014	0.0025	0.0133	1.5859
	$\widehat{MISE}$	0.1514	0.1517	0.1503	0.1774	1.5550
MSE-min.	Gini coeff.	0.5847	0.5741	0.5377	0.4371	0.8134
	Var. Std.	0.0041	0.0039	0.0035	0.0020	0.1358
	$\widehat{MISE}$	0.1549	0.1551	0.1524	0.1585	0.3505
Simulation 2.						
VS.Diag.	Gini coeff.	0.3238	0.3333	0.3609	0.3826	0.9343
	Var. Std.	0.0013	0.0016	0.0026	0.0098	68.1049
	$\widehat{MISE}$	0.1710	0.1307	0.0941	0.0846	30.2487
VS.Full.	Gini coeff.	0.3205	0.3380	0.3524	0.4025	0.9268
	Var. Std.	0.0008	0.0009	0.0015	0.0072	0.7523
	$\widehat{MISE}$	0.1686	0.1283	0.0921	0.0747	0.8726
MSE-min.	Gini coeff.	0.5820	0.5809	0.5366	0.4198	0.8222
	Var. Std.	0.0040	0.0043	0.0035	0.0020	0.1399
	$\widehat{MISE}$	0.1770	0.1364	0.0967	0.0630	0.3694
Simulation 3.						
VS.Diag.	Gini coeff.	0.2872	0.3388	0.2833	0.3299	0.4115
	Var. Std.	0.0012	0.0236	0.0018	0.0020	0.0020
	$\widehat{MISE}$	0.4276	0.4217	0.4042	0.3695	0.0404
VS.Full.	Gini coeff.	0.2832	0.2570	0.2837	0.3346	0.4000
	Var. Std.	0.0012	0.0009	0.0016	0.0056	0.1399
	$\widehat{MISE}$	0.4276	0.4217	0.4039	0.3657	0.0532
MSE-min.	Gini coeff.	0.5730	0.5899	0.6138	0.5384	0.5083
	Var. Std.	0.0036	0.0039	0.0052	0.0032	0.0034
	$\widehat{MISE}$	0.4228	0.4159	0.3971	0.3500	0.0122

Table 2: The result of Monte-Carlo simulation with theoretical bandwidth matrices.

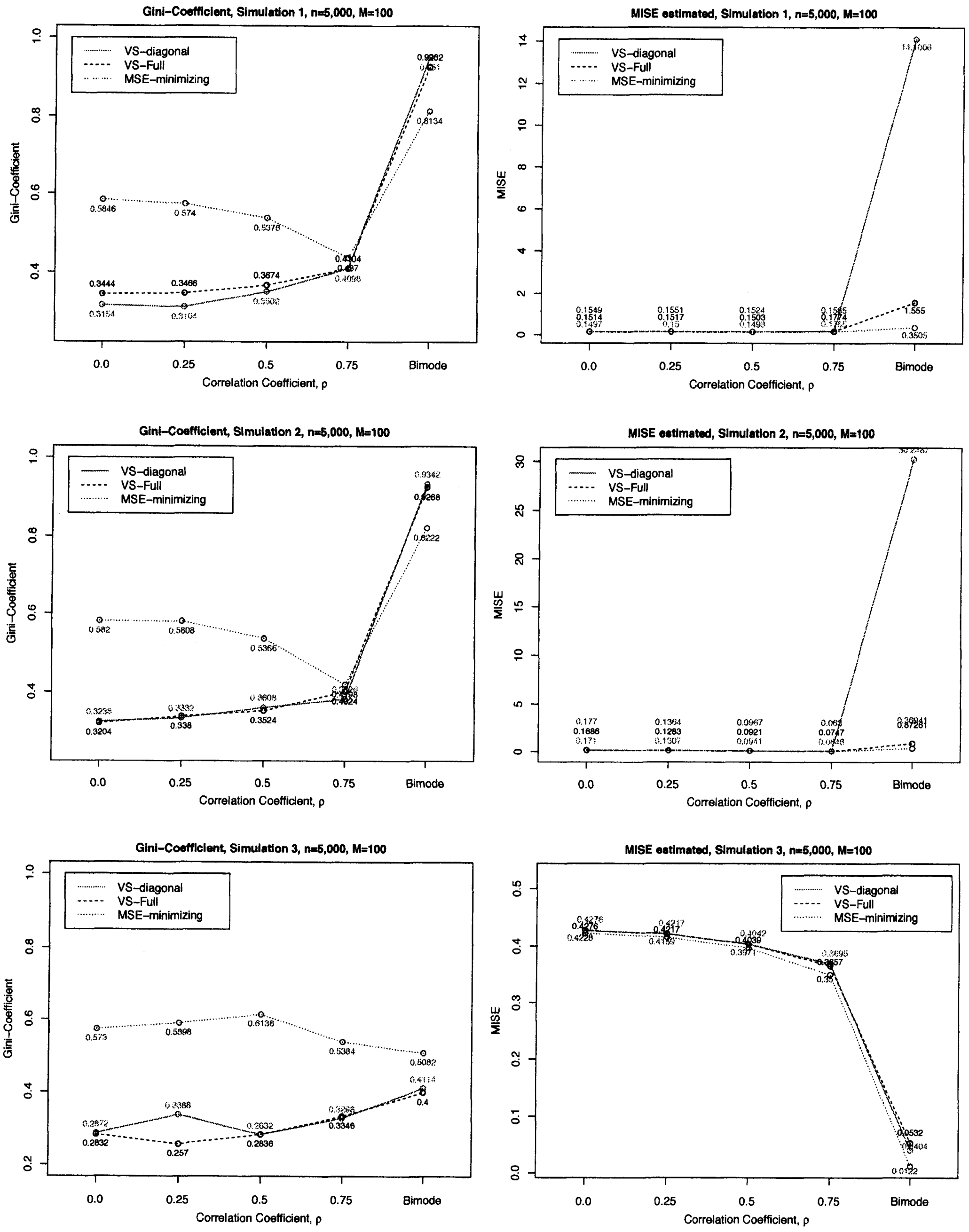

Figure 5: Summary of the simulation results.

4 Discussion

In this paper, we propose the VS full-bandwidth matrix for the multivariate LL estimator to complement the VS diagonal bandwidth matrix proposed by Nishida and Kanazawa (2013). The derivation of the optimal VS full-bandwidth matrix in multivariate setting is so intensive that we consider the problem in bivariate setting with an additional assumption $h_{11}^{Full} = h_{22}^{Full}$. However, even in this tempered setting, it is impossible to explicitly obtain the optimal parameters, h_{11}^{Full}, h_{12}^{Full} and $\eta_{11}^{Full}(x_1, x_2)$, so we resort to numerical calculation. Although the parameter $\eta_{11}^{Full}(x_1, x_2)$, which is arranged to negate the variance term, should be locally determined by nature, we are obliged to use it as an universal parameter over the domain to ease the computational burden.

Our main concern in this paper is to make a comparison between the two VS bandwidth matrices in terms of MISE and the stability of the variance. To validate this, we run Monte-Carlo simulations with theoretical bandwidth matrices, $\mathbf{H}_{\mathbf{VS}}(x_1, x_2)$, $\mathbf{H}_{\mathbf{VS}^{++}}(x_1, x_2)$ and $\mathbf{H}_{\mathbf{var}}(x_1, x_2)$. Through the simulation, we confirm that either $\mathbf{H}_{\mathbf{VS}}(x_1, x_2)$ or $\mathbf{H}_{\mathbf{VS}^{++}}(x_1, x_2)$ is superior to $\mathbf{H}_{\mathbf{var}}(x_1, x_2)$ in terms of stability of variance over the domain. We also notice that the mixed derivative of $m(x_1, x_2)$ surely influences the result in terms of MISE. As for the correlation between covariates, we observe that the theoretical parameters, h_0^*, $h_{11}^{Full,*}$ and $h_{12}^{Full,*}$ as well as MISE tend to diminish as ρ increases by the presented simulation cases. We also observe that variance-stabilization is difficult for both of the two VS bandwidth matrices when covariates are highly correlated. As for the multimodality of the density function, we obtain neither a tendency nor clear-cut explanations.

In our Monte-Carlo simulation study, we present two measures, the Gini-coefficient and the MISE estimated. Since these two measures are two different things, we can choose the type of the bandwidth matrix that optimizes, for example, the following performance function,

$$\zeta \cdot \text{Gini-coefficient} + (1 - \zeta) \cdot \text{MISE}, \tag{21}$$

where ζ denotes the ratio representing the level of importance between the stability of variance and Error. In Table 3, we revalue simulation 2 by the performance function (21). From Table 3, we notice that the VS full-bandwidth matrix is well-balanced between stability of variance and error in this simulation setting.

To obtain the estimator of the VS full-bandwidth matrix, we need to obtain the estimator of the mixed derivative function of $m(\cdot)$, as well as $f_{X_1,X_2}(\cdot)$ and $\sigma^2(\cdot)$. The estimation of the mixed derivative function of $m(\cdot)$ employing partial local cubic estimator is difficult in general and requires us to estimate its pilot bandwidth beforehand via ROT approach or cross-validation. After that, we resort to numerical calculation to obtain h_{11}^{Full}, h_{12}^{Full} and η_{11}^{Full}. It is expected that the calculation is far more intensive.

		$\rho = 0.00$	$\rho = 0.25$	$\rho = 0.50$	$\rho = 0.75$	*Bimode*
Simulation 2.						
$\zeta = 0.00$	VS.Diag.	0.1711	0.1308	0.0942	0.0846	30.2488
	VS.Full.	0.1687	0.1284	0.0921	0.0747	0.8726
	MSE-min.	0.1770	0.1365	0.0967	0.0630	0.3694
$\zeta = 0.25$	VS.Diag.	0.2093	0.1814	0.1615	0.1591	22.9202
	VS.Full.	0.2066	0.1808	0.1572	0.1567	0.8862
	MSE-min.	0.2783	0.2476	0.2067	0.1522	0.4826
$\zeta = 0.50$	VS.Diag.	0.2475	0.2321	0.2289	0.2337	15.5916
	VS.Full.	0.2446	0.2332	0.2223	0.2386	0.8998
	MSE-min.	0.3795	0.3587	0.3167	0.2414	0.5958
$\zeta = 0.75$	VS.Diag.	0.2857	0.2827	0.2962	0.3082	8.2630
	VS.Full.	0.2826	0.2857	0.2874	0.3206	0.9133
	MSE-min.	0.4808	0.4698	0.4267	0.3306	0.7091
$\zeta = 1.00$	VS.Diag.	0.3239	0.3334	0.3636	0.3827	0.9344
	VS.Full.	0.3206	0.3381	0.3525	0.4025	0.9269
	MSE-min.	0.5820	0.5809	0.5367	0.4199	0.8223

Table 3: The result of Simulation 2 revalued by the performance function (21).

Acknowledgements

The author thanks Professor Yuichiro KANAZAWA at University of Tsukuba and Visiting Associate Professor Kazuaki NAKANE at Osaka university for their helpful comments. The author also thanks the financial support from the Japan Society for the Promotion of Science under Grant-in-Aid for Research Activity Start-up 24830048.

References

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

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

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

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

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

[6] 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.

[7] Nishida, K. and Kanazawa, Y. (2013). On Variance-Stabilizing Multivariate Nonparametric Regression Estimation. Communications in Statistics — Theory and Methods, in press.

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

[9] Ruppert, D., Sheather, S.J. and Wand, M.P. (1995). An Effective Bandwidth Selector for Local Least Squares Regression. Journal of the American Statistical Association 90:1257-1270.

[10] 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.

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

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

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

General Education Center,
Hyogo University of Health Sciences,
1-3-6, Minatojima, Chuo-ku, Kobe, Hyogo, 650-8530, JAPAN.
E-mail address: `kiheiji.nishida@gmail.com`

兵庫医療大学, 西田 喜平次