数理解析研究所講究録
第1703巻 2010年 69-90

On the variance-stabilizing multivariate nonparametric regression estimation

Kiheiji NISHIDA and Yuichiro J. KANAZAWA [1]

1 Introduction

The objective of this paper is to show the idea of the nonparametric regression estimator that produces constant estimator variance over all values of the regressor variable. To accomplish the purpose, we introduce the variable bandwidth matrix based on asymptotic considerations chosen so as to correct heteroscedasticity, employing the same principle of the Aitken estimator in linear regression. We call the bandwidth matrix for this objective the variance-stabilizing (henceforth VS) bandwidth matrix.

As past studies for the VS bandwidth, we know that Fan and Gijbel (1992) introduced the idea of the VS bandwidth for the univariate local linear estimator. However, they did not investigate the VS bandwidth in detail because the paper was aimed at studying the variable bandwidth that minimizes mean squared error (MSE). In contrast to this, Nishida and Kanazawa (2009, 2010)

[1]Kiheiji NISHIDA is a student at the Graduate School of Systems and Information Engineering, University of Tsukuba. Yuichiro J. KANAZAWA is Professor of Statistics at the Graduate School of Systems and Information Engineering, University of Tsukuba, 1-1-1 Ten-noh-dai, Tsukuba, Ibaraki 305-8573, Japan. Correspondence concerning this article should be addressed to Kiheiji NISHIDA. His e-mail address is kiheiji.nishida@gmail.com.
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.

introduced the VS bandwidth for the univariate Nadaraya-Watson (henceforth NW) estimator and compared it to the Mean Integrated Squared Error (MISE) minimizing fixed bandwidth in terms of global and local performance. These two studies focus on univariate nonparametric regression estimators.

In the light of these two studies, we introduce the VS bandwidth matrix for the multivariate NW estimator. In section 2, we show the brief summaries of the multivariate NW estimator to know the principle of how the VS bandwidth matrix is introduced. In section 3, we introduce the VS bandwidth matrix for the multivariate NW estimator. In each subsection of section 2, we show the performance of the univariate and the bivariate VS regression estimators. We also show the estimating procedure of the VS bandwidth matrix in the univariate case. In section 4, we give concluding remark.

2 Summaries of the multivariate NW estimator

2.1 Basic setup

Let us consider a $p+1$ row vector $(X_{i0}, ..., X_{i(p-1)}, Y_i)$ of random variables. We assume $\mathbf{x}_{i\cdot} = (x_{i0}, ..., x_{i(p-1)})$, $i = 1, ..., n$, are the realizations of random explanatory vector $\mathbf{X}_{i\cdot} = (X_{i0}, ..., X_{i(p-1)})$, i.i.d. with respect to i and whose joint density function is denoted as $f_{\mathbf{X}_{i\cdot}}(\mathbf{x}_{i\cdot})$ on support $I^p \in R^p$. The n sample realizations of $(X_{i0}, ..., X_{i(p-1)})$ can be written in matrix form as

$$\begin{pmatrix} x_{10} \dots x_{1(p-1)} \\ \vdots \quad \ddots \quad \vdots \\ x_{n0} \dots x_{n(p-1)} \end{pmatrix}$$

and we define the ith column of the matrix in (2.1) as $\mathbf{x}_{\cdot i}$. We assume that the jth random explanatory variable $\mathbf{X}_{\cdot j}$ may be correlated or not orthogonal

with the kth 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

$$Y_i = m(\mathbf{X}_{i\cdot}) + U_i, \tag{2.1}$$

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 additionally assume the first two conditional moments of $U_i|\mathbf{X}_{i\cdot}$ are

$$\begin{aligned} E_{U_i|\mathbf{X}_{i\cdot}}[U_i|\mathbf{X}_{i\cdot} = \mathbf{x}_{i\cdot}] &= 0, \\ E_{U_i|\mathbf{X}_{i\cdot}}[U_i^2|\mathbf{X}_i = \mathbf{x}_{i\cdot}] &= \sigma^2(\mathbf{x}_{i\cdot}). \end{aligned} \tag{2.2}$$

Then, from the assumptions (2.1) and (2.2), we obtain

$$\begin{aligned} m(\mathbf{x}_{i\cdot}) &= E_{Y_i|\mathbf{X}_{i\cdot}}[Y_i|\mathbf{X}_{i\cdot} = \mathbf{x}_{i\cdot}] = \int y_i f_{Y_i|\mathbf{X}_{i\cdot}}(y_i|\mathbf{x}_{i\cdot})dy_i \\ &= \frac{\int y_i f_{\mathbf{X}_{i\cdot},Y_i}(\mathbf{x}_{i\cdot}, y_i)dy_i}{f_{\mathbf{X}_{i\cdot}}(\mathbf{x}_{i\cdot})}, \end{aligned}$$

where $f_{\mathbf{X}_{i\cdot},Y_i}(\mathbf{x}_{i\cdot}, y_i)$ is the multivariate density function of $(\mathbf{X}_{i\cdot}, Y_i)$. Replacing unknown density function $f_{\mathbf{X}_{i\cdot},Y_i}(\mathbf{x}_{i\cdot}, y_i)$ and $f_{\mathbf{X}_{i\cdot}}(\mathbf{x}_{i\cdot})$ with their estimator $\widehat{f_{\mathbf{X}_{i\cdot},Y_i}}(\mathbf{x}_{i\cdot}, y_i)$ and $\widehat{f_{\mathbf{X}_{i\cdot}}}(\mathbf{x}_{i\cdot})$, we obtain the nonparametric estimator $\hat{m}(\mathbf{x}_{i\cdot})$ written as

$$\hat{m}(\mathbf{x}_{i\cdot}) = \frac{\int y_i \widehat{f_{\mathbf{X}_{i\cdot},Y_i}}(\mathbf{x}_{i\cdot}, y_i)dy_i}{\widehat{f_{\mathbf{X}_{i\cdot}}}(\mathbf{x}_{i\cdot})}. \tag{2.3}$$

In estimating the denominator of (2.3), we employ the multivariate kernel density estimator with p dimensional kernel function $K_{\mathbf{X}}(\mathbf{x})$ at any point $\mathbf{x} = (x_0, ..., x_{(p-1)})$ on $I^p \in R^p$ and with multivariate symmetric bandwidth matrix,

$$\mathbf{H_X} = \begin{pmatrix} h_{00} & \cdots & h_{0(p-1)} \\ \vdots & \ddots & \vdots \\ h_{0(p-1)} & \cdots & h_{(p-1)(p-1)} \end{pmatrix}, \; h_{00} > 0, ..., h_{(p-1)(p-1)} > 0, \; |\mathbf{H_X}| \neq 0, \tag{2.4}$$

written as

$$\widehat{f_{\mathbf{H_X}}}(\mathbf{x}) = \frac{1}{n|\mathbf{H_X}|}\sum_{i=1}^{n} K_{\mathbf{X}}(\mathbf{H_X^{-1}}(\mathbf{x}-\mathbf{x}_{i\cdot})). \tag{2.5}$$

Similarly, in estimating the numerator of (2.3), let $K_{\mathbf{X},Y}(\mathbf{x},y)$ be $(p+1)$-dimensional kernel function at the point $(\mathbf{x},y)=(x_0,...,x_{(p-1)},y)$, and $\mathbf{H_{X,Y}}$ be the $(p+1)\times(p+1)$ dimensional symmetric smoothing parameter matrix,

$$\mathbf{H_{X,Y}}=\begin{pmatrix} h_{00} & \dots & h_{0(p-1)} & h_{0y} \\ \vdots & \ddots & \vdots & \vdots \\ h_{0(p-1)} & \dots & h_{(p-1)(p-1)} & h_{(p-1)y} \\ h_{0y} & \dots & h_{(p-1)y} & h_{yy} \end{pmatrix} = \begin{pmatrix} \mathbf{H_X} & \mathbf{h}_y^T \\ \mathbf{h}_y & h_{yy} \end{pmatrix},$$

where $\mathbf{h}_y=(h_{0y},...,h_{(p-1)y})$. With this notation, $(p+1)$ dimensional kernel density estimator that appears in the numerator of (2.3) is written as

$$\widehat{f_{\mathbf{H_{X,Y}}}}(\mathbf{x},y) = \frac{1}{n|\mathbf{H_{X,Y}}|}\sum_{i=1}^{n} K_{\mathbf{X},Y}(\mathbf{H_{X,Y}^{-1}}(\mathbf{x}-\mathbf{x}_{i\cdot},y-Y_i)). \tag{2.6}$$

If we additionally assume that the kernel $K_{\mathbf{X},Y}(\mathbf{x},y)$ is multiplicative in terms of $\mathbf{X}$ and Y, the smoothing parameter matrix is written as

$$\mathbf{H_{X,Y}} = \begin{pmatrix} \mathbf{H_X} & \mathbf{0} \\ \mathbf{0} & h_{yy} \end{pmatrix}, \quad |\mathbf{H_{X,Y}}| = |\mathbf{H_X}|h_{yy}$$

and the numerator of (2.3) is rewritten by the symmetry of the kernel as

$$\begin{aligned} &\int \frac{y}{n|\mathbf{H_{X,Y}}|}\sum_{i=1}^{n} K_{\mathbf{X},Y}\left(\mathbf{H_{X,Y}^{-1}}(\mathbf{x}-\mathbf{x}_{i\cdot},y-Y_i)\right)dy \\ =&\frac{1}{n|\mathbf{H_X}|h_{yy}}\sum_{i=1}^{n}\int yK_{\mathbf{X}}\left(\mathbf{H_X^{-1}}(\mathbf{x}-\mathbf{x}_{i\cdot})\right)K_Y\left(\frac{y-Y_i}{h_{yy}}\right)dy \\ =&\frac{1}{n|\mathbf{H_X}|}\sum_{i=1}^{n}\int (z+Y_i)K_{\mathbf{X}}\left(\mathbf{H_X^{-1}}(\mathbf{x}-\mathbf{x}_{i\cdot})\right)K_Y(z)dz \end{aligned}$$

$$=\frac{1}{n|\mathbf{H_X}|}\sum_{i=1}^{n} K_{\mathbf{X}}(\mathbf{H_X^{-1}}(\mathbf{x}-\mathbf{x}_{i\cdot}))Y_i. \tag{2.7}$$

Substituting the numerator and the denominator in (2.3) respectively for (2.5) and (2.7), the multivariate NW estimator at the point $\mathbf{x}$ is rewritten as

$$\widehat{m_{\mathbf{H_X}}}(\mathbf{x})=\frac{\sum_{i=1}^{n} K_{\mathbf{X}}(\mathbf{H_X^{-1}}(\mathbf{x}-\mathbf{x}_{i\cdot}))Y_i}{\sum_{i=1}^{n} K_{\mathbf{X}}(\mathbf{H_X^{-1}}(\mathbf{x}-\mathbf{x}_{i\cdot}))}.$$

(For univariate case, Nadaraya 1964, 1965, 1970; Watson 1964; Watson and Leadbetter 1964. For multivariate case, e.g. Härdle 2004). We write the kernel function $K_{\mathbf{X}}(\mathbf{x})$ as $K(\mathbf{x})$ for brevity.

2.2 A restriction on multivariate bandwidth matrix

We see the role of the off-diagonal elements of the bandwidth matrix in (2.4) in kernel density estimation. Observe the multivariate gaussian kernel,

$$K(\mathbf{H_X^{-1}x}) = \frac{1}{2\pi|\mathbf{H_X}|}\exp\left[-\frac{1}{2}\mathbf{x}(\mathbf{H_X^{-1}})^2\mathbf{x}^T\right],$$

and we find that $\mathbf{H_X^2}$ corresponds to the variance covariance matrix Σ of the normal distribution. This means that the off-diagonal elements of the bandwidth matrix $\mathbf{H_X}$ in (2.4) reflect the correlation between the variables in kernel function because the off-diagonal elements of $\mathbf{H_X^2}$ are zero if $h_{(k)(j)}$, $k \neq j$, are all set to be zero. This property is true of the symmetric kernel functions defined by the quadratic form $\mathbf{x}(\mathbf{H_X^{-1}})^2\mathbf{x}^T$ other than Gaussian.

Considering this property with the off-diagonal elements of bandwidth matrix in (2.4), we notice that it is better to employ a full bandwidth matrix for kernel density estimation if it is considered that $\mathbf{X}_{\cdot i}$ is correlated with $\mathbf{X}_{\cdot j}$. In this point, Wand and Jones (1993) demonstrated in the bivariate density estimation situation that MISE of the kernel density estimator is deteriorated if an orthogonal kernel is employed for a correlated data. In their seminal

ρ	0.0	0.3	0.6	0.9	1.0
the ratio (2.8)	1.00	0.93	0.74	0.37	0.00

Table 1: The ratio (2.8) for different values of correlation coefficient ρ from Wand and Jones (1993).

paper, they derived the theoretical AMISE of the bivariate kernel density estimator for the bivariate normal distribution with its correlation coefficient ρ. They also calculated the ratio,

$$\left[\frac{\inf_{\mathbf{H}\in\mathbf{H}_{full}} AMISE(\mathbf{H})}{\inf_{\mathbf{H}\in\mathbf{H}_{diag}} AMISE(\mathbf{H})}\right]^{\frac{3}{2}} = \left[\frac{2(1-\rho^2)}{\rho^2+2}\right]^{\frac{1}{2}}, \tag{2.8}$$

for different values of ρ, where $\mathbf{H}_{diag}$ and $\mathbf{H}_{full}$ are respectively the classes of the bivariate diagonal and the bivariate full bandwidth matrix. This ratio can be interpreted as the costs from not using full bandwidth matrix for the correlated data in kernel density estimation. The result in table 1 indicates that there surely exist the costs.

However, when we employ a full bandwidth matrix in (2.4), the number of parameters to be estimated is $p(p+1)/2$. If we take the curse of dimensionality into consideration, this is too many. To reduce the number of parameters to be estimated, the sphering approach, as in e.g. Fukunaga (1972) or Wand and Jones (1993), is available if it is probable that $\mathbf{X}_{i\cdot}$ is distributed as normal distribution with variance covariance matrix Σ. That is, linearly transforming the data so that the sample variance covariance matrix is diagonal, we compute the density by the diagonal bandwidth matrix and finally retransform the estimated pdf back to the original scale. In the following, we assume that the data is approximately distributed as normal and the sphering approach is available. Then we can search for the form of VS bandwidth matrix within the limit of the diagonal one, $\mathbf{H}_{\mathbf{X}} = \mathrm{diag}(h_{00}, \cdots, h_{(p-1)(p-1)})$.

2.3 On the conventional bandwidth matrices

Under the assumption of the diagonal bandwidth matrix, the bias and the variance of the multivariate NW estimator are derived at every point $\mathbf{x}$ (e.g. Härdle 2004) under the standard set of assumptions on appendix A. The bias so obtained is

$$\begin{aligned}&E_{\mathbf{Y},\mathbf{X}}\left[\widehat{m_{\mathbf{H_X}}}(\mathbf{x})\right] - m(\mathbf{x})\\ =&\frac{\mu_2(K_\mathbf{X})}{2f_\mathbf{X}(\mathbf{x})}\left[2(\nabla m(\mathbf{x}))^T\mathbf{H_X}\mathbf{H_X^T}\nabla f_\mathbf{X}(\mathbf{x}) + f_\mathbf{X}(\mathbf{x})\mathrm{tr}\left[\mathbf{H_X^T}\nabla^2 m(\mathbf{x})\mathbf{H_X}\right]\right] + o\left(1\right),\end{aligned}$$

where

$$\nabla f_\mathbf{X}(\mathbf{x}) = \begin{pmatrix}\frac{\partial f_\mathbf{X}(\mathbf{x})}{\partial x_0}\\ \vdots \\ \frac{\partial f_\mathbf{X}(\mathbf{x})}{\partial x_{p-1}}\end{pmatrix}, \nabla^2 f_\mathbf{X}(\mathbf{x}) = \begin{pmatrix}\frac{\partial^2 f_\mathbf{X}(\mathbf{x})}{\partial x_0 \partial x_0} & \cdots & \frac{\partial^2 f_\mathbf{X}(\mathbf{x})}{\partial x_0 \partial x_{p-1}}\\ \vdots & \ddots & \vdots \\ \frac{\partial^2 f_\mathbf{X}(\mathbf{x})}{\partial x_{p-1} \partial x_0} & \cdots & \frac{\partial^2 f_\mathbf{X}(\mathbf{x})}{\partial x_{p-1} \partial x_{p-1}}\end{pmatrix},$$

and

$$\mu_2(K_\mathbf{X}) = \int\cdots\int \mathbf{t}\mathbf{t}^T K(\mathbf{t})d\mathbf{t} \text{ and } \mathbf{t} = (t_0, ..., t_{p-1}).$$

Similarly, the variance so obtained is

$$V_{\mathbf{Y},\mathbf{X}}\left[\widehat{m_{\mathbf{H_X}}}(\mathbf{x})\right] = \frac{1}{n|\mathbf{H_X}|}\cdot\frac{\sigma^2(\mathbf{x})}{f_\mathbf{X}(\mathbf{x})}\left[\int\cdots\int K^2(\mathbf{t})d\mathbf{t}\right] + o(1). \quad (2.9)$$

When $h_{00} = h_{11} = \cdots = h_{(p-1)(p-1)}$, we can obtain the asymptotic MISE (AMISE) and derive the diagonal fixed bandwidth matrix that balances the leading terms of the integrated variance and the integrated bias squared written as

$$\mathbf{H}_{fixed} = \left[\frac{\left[\int\cdots\int K^2(\mathbf{t})d\mathbf{t}\right]\left[\int\cdots\int \sigma^2(\mathbf{x})d\mathbf{x}\right]}{\mu_2^2(K)T_{fixed}}\right]^{\frac{1}{p+4}} p^{\frac{1}{p+4}}\cdot n^{-\frac{1}{p+4}}\cdot\mathbf{I}_p, \quad (2.10)$$

where

$$T_{fixed} = \int\cdots\int \frac{1}{f_\mathbf{X}(\mathbf{x})}\left[\sum_{i=0}^{p-1}\alpha_i(\mathbf{x})\right]^2 d\mathbf{x}, \quad (2.11)$$

and

$$\alpha_i(\mathbf{x})=2\left(\frac{\partial m(\mathbf{x})}{\partial x_i}\right)\left(\frac{\partial f_{\mathbf{X}}(\mathbf{x})}{\partial x_i}\right)+f_{\mathbf{X}}(\mathbf{x})\left(\frac{\partial^2 m(\mathbf{x})}{\partial x_i^2}\right), i=0,\cdots p-1.$$

Hence the AMISE minimized by (2.10) is

$$\left(p^{\frac{-p}{p+4}}+\frac{p^{\frac{4}{p+4}}}{4}\right)\left[\frac{\left[\int\cdots\int K^2(\mathbf{t})d\mathbf{t}\right]^{\frac{4}{p+4}}\left[\int\cdots\int\sigma^2(\mathbf{x})d\mathbf{x}\right]^{\frac{4}{p+4}}}{\left[\mu_2^2(K)\right]^{-\frac{p}{p+4}}\left[T_{fixed}\right]^{-\frac{p}{p+4}}}\right]\cdot n^{-\frac{4}{p+4}}.$$

In the same manner to the fixed one in (2.10), the variable bandwidth matrix that minimizes mean squared error MSE defined at each $\mathbf{x}$ can be obtained as

$$\mathbf{H}_{var}(\mathbf{x})=\left[\frac{\left[\int\cdots\int K^2(\mathbf{t})d\mathbf{t}\right]\sigma^2(\mathbf{x})f_{\mathbf{X}}(\mathbf{x})}{\mu_2^2(K)\left[\sum_{i=0}^{p-1}\alpha_i(\mathbf{x})\right]^2}\right]^{\frac{1}{p+4}}p^{\frac{1}{p+4}}\cdot n^{-\frac{1}{p+4}}\cdot\mathbf{I}_p$$

and the AMISE so obtained by using (2.12) is

$$\left(p^{\frac{-p}{p+4}}+\frac{p^{\frac{4}{p+4}}}{4}\right)\left[\frac{\left[\int\cdots\int K^2(\mathbf{t})d\mathbf{t}\right]^{\frac{4}{p+4}}T_{var}}{\left[\mu_2^2(K)\right]^{-\frac{p}{p+4}}}\right]\cdot n^{-\frac{4}{p+4}},$$

where

$$T_{var}=\int\cdots\int\left[\sigma^2(\mathbf{x})\right]^{\frac{4}{p+4}}\left[\frac{1}{f_{\mathbf{X}}(\mathbf{x})}\left[\sum_{i=0}^{p-1}\alpha_i(\mathbf{x})\right]^2\right]^{\frac{p}{p+4}}d\mathbf{x}.$$

When $h_{00}=h_{11}=\cdots=h_{(p-1)(p-1)}$ does not necessarily hold, it is difficult to generalize both the fixed and the variable bandwidth matrices.

2.4 The problem with the variance of the multivariate NW estimator

We adderess the problem with the variance of the NW estimator. The leading term in the variance (2.9) at the point $\mathbf{x}$ is a function of the term $\sigma^2(\mathbf{x})/f_{\mathbf{X}}(\mathbf{x})$.

Both $f_{\mathbf{X}}(\mathbf{x})$ and $\sigma^2(\mathbf{x})$ are very unlikely to be uniform across different $\mathbf{x}$ in general. As a result, variances of the NW estimator at different points $\mathbf{x}_1$ and $\mathbf{x}_2$ differ in general unless properly controlled by the bandwidth matrix $\mathbf{H}_{\mathbf{X}}$. As examples of the heteroscedasticity of the NW estimator's variance, see Nishida and Kanazawa (2009, 2010).

3 Introduction of the multivariate VS bandwidth matrix

We consider the form of the bandwidth matrix that counters heteroscedasticity. Employing the same principle of the Aitken estimator in linear regression, we propose the VS bandwidth matrix at the point $\mathbf{x}$,

$$\mathbf{H}_{VS}(\mathbf{x}) = h_0 \begin{pmatrix} \left[\frac{\sigma^2(\mathbf{x})}{f_{\mathbf{X}}(\mathbf{x})}\right]^{\eta_0(\mathbf{x})} & 0 & \cdots & 0 \\ 0 & \left[\frac{\sigma^2(\mathbf{x})}{f_{\mathbf{X}}(\mathbf{x})}\right]^{\eta_1(\mathbf{x})} & \cdots & \vdots \\ \vdots & \vdots & \ddots & \vdots \\ 0 & \cdots & \cdots & \left[\frac{\sigma^2(\mathbf{x})}{f_{\mathbf{X}}(\mathbf{x})}\right]^{\eta_{p-1}(\mathbf{x})} \end{pmatrix}, \tag{3.1}$$

where $h_0 > 0$ is a global constant and $\eta_i(\mathbf{x})$'s, $i = 0, ..., p-1$, are the local constants that are defined at each point $\mathbf{x}$ and satisfying

$$\sum_{i=0}^{p-1} \eta_i(\mathbf{x}) = 1, \quad -\infty < \eta_i(\mathbf{x}) < \infty. \tag{3.2}$$

Among the class of bandwidth matrix in the form of (3.1) and (3.2), we optimize global parameter h_0 so as to minimize AMISE given $\eta_i(\mathbf{x})$. The bandwidth matrix so obtained given $\eta_i(\mathbf{x})$ is

$$\mathbf{H}_{VS}(\mathbf{x}) = \left[\frac{[\int \cdots \int K^2(\mathbf{t})d\mathbf{t}]}{\mu_2^2(K) T_{VS}(\eta_0(\mathbf{x}), ..., \eta_{p-1}(\mathbf{x}))}\right]^{\frac{1}{p+4}} p^{\frac{1}{p+4}} \cdot n^{-\frac{1}{p+4}}$$

$$\times \operatorname{diag}\left(\left[\frac{\sigma^2(\mathbf{x})}{f_{\mathbf{X}}(\mathbf{x})}\right]^{\eta_0(\mathbf{x})}, \ldots, \left[\frac{\sigma^2(\mathbf{x})}{f_{\mathbf{X}}(\mathbf{x})}\right]^{\eta_{p-1}(\mathbf{x})}\right), \tag{3.3}$$

where

$$T_{VS}(\eta_0(\mathbf{x}), \ldots, \eta_{p-1}(\mathbf{x})) = \int \cdots \int \frac{1}{f_{\mathbf{X}}(\mathbf{x})} \left[\sum_{i=0}^{p-1} \alpha_i(\mathbf{x}) \left[\frac{\sigma^2(\mathbf{x})}{f_{\mathbf{X}}(\mathbf{x})}\right]^{2\eta_i(\mathbf{x})}\right]^2 d\mathbf{x}. \tag{3.4}$$

The asymptotic homoscedastic variance so obtained given $\eta_i(\mathbf{x})$ is

$$\left[\frac{\left[\int \cdots \int K^2(\mathbf{t})d\mathbf{t}\right]^{\frac{4}{p+4}}}{[\mu_2^2(K)]^{-\frac{p}{p+4}} \left[T_{VS}(\eta_0(\mathbf{x}), \ldots, \eta_{p-1}(\mathbf{x}))\right]^{-\frac{p}{p+4}}}\right] \cdot p^{-\frac{p}{p+4}} \cdot n^{-\frac{4}{p+4}}$$

and the AMISE given $\eta_i(\mathbf{x})$ is

$$\left(p^{\frac{-p}{p+4}} + \frac{p^{\frac{4}{p+4}}}{4}\right) \left[\frac{\left[\int \cdots \int K^2(\mathbf{t})d\mathbf{t}\right]^{\frac{4}{p+4}}}{[\mu_2^2(K)]^{-\frac{p}{p+4}} \left[T_{VS}(\eta_0(\mathbf{x}), \ldots, \eta_{p-1}(\mathbf{x}))\right]^{-\frac{p}{p+4}}}\right] \cdot n^{-\frac{4}{p+4}}. \tag{3.5}$$

Subsequently, to minimize (3.5), the term $T_{VS}(\eta_0(\mathbf{x}), \ldots, \eta_{p-1}(\mathbf{x}))$ in (3.5) must also be minimized in terms of $\eta_i(\mathbf{x})$ $i = 0, 1, \ldots, p-1$. For such $\eta_i(\mathbf{x})$, we solve the following minimization problem in terms of $\eta_i(\mathbf{x})$ at every $\mathbf{x}$,

$$\min_{\eta(\mathbf{x})} \left[\sum_{i=0}^{p-1} \alpha_i(\mathbf{x}) \left[\frac{\sigma^2(\mathbf{x})}{f_{\mathbf{X}}(\mathbf{x})}\right]^{2\eta_i(\mathbf{x})}\right], \quad \text{s.t.} \sum_{i=0}^{p-1} \eta_i(\mathbf{x}) = 1. \tag{3.6}$$

We have so far derived the optimal parameters $\eta_i^*(\mathbf{x})$ up to $p = 2$. In the following, we show the property of VS bandwidth matrix compared with the heteroscedastic bandwidth matrix in the cases of $p = 1$ and $p = 2$. We also show the estimating procedure in the case of $p = 1$.

3.1 Univariate VS bandwidth : $p = 1$

We write the point x_0 as x for brevity in the case of $p = 1$. Setting $p = 1$, we obtain $\eta_0^*(x) = 1$ at every x. Then, we obtain univariate VS bandwidth,

$$h_{VS}(x) = h_0 \cdot \frac{\sigma^2(x)}{f_X(x)}$$

$$= \left[\frac{\left[\int K^2(t)dt\right]}{\left[\int t^2 K(t)dt\right]^2 \left[\int \frac{\sigma^8(x)\alpha^2(x)}{f_X^5(x)}dx\right]} \right]^{\frac{1}{5}} n^{-\frac{1}{5}} \cdot \frac{\sigma^2(x)}{f_X(x)}, \qquad (3.7)$$

where $\alpha(x) = 2f_X^{(1)}(x)m^{(1)}(x) + f_X(x)m^{(2)}(x)$. For this, $\hat{m}_{h_{VS}}(x)$ is homoscedastic up to order $n^{-\frac{4}{5}}$,

$$AV_{\mathbf{X},\mathbf{Y}}\left[\widehat{m_{h_{VS}}}(x)\right] = \left[\frac{\left[\int K^2(t)dt\right]^{\frac{4}{5}}}{\left[\int t^2 K(t)dt\right]^{-\frac{2}{5}} \left[\int \frac{\sigma^8(x)\alpha^2(x)}{f_X^5(x)}dx\right]^{-\frac{1}{5}}} \right] n^{-\frac{4}{5}}. \qquad (3.8)$$

Let the MISE minimizing fixed bandwidth and the MSE minimizing variable bandwidth for the NW estimator respectively be h_{fixed} and $h_{var}(x)$. Fan and Gijbel (1992) compared the VS bandwidth for the univariate local linear estimator with the MSE minimizing variable bandwidth in terms of AMISE. Nishida and Kanazawa (2010) compared the VS bandwidth for the univariate NW estimator with the fixed bandwidth in terms of AMISE and local variance. In the following, we compare three bandwidths h_{fixed}, $h_{var}(x)$ and $h_{VS}(x)$ for the univariate NW estimator in terms of AMISE and local variance.

3.1.1 Comparison of the three bandwidths in terms of local variance.

We compare three bandwidths in terms of variance. The asymptotic variance when h_{fixed} is employed is

$$AV_{\mathbf{X},\mathbf{Y}}\left[\widehat{m_{h_{fixed}}}(x)\right] = \frac{\sigma^2(x)}{f_X(x)} \left[\frac{\left[\int K^2(t)dt\right]^{\frac{4}{5}} \left[\int_{I^1} \sigma^2(x)dx\right]^{-\frac{1}{5}}}{\left[\int t^2 K(t)dt\right]^{-\frac{2}{5}} \left[\int_{I^1} \frac{\alpha^2(x)}{f_X(x)}dx\right]^{-\frac{1}{5}}} \right] n^{-\frac{4}{5}}. \qquad (3.9)$$

Similarly, the asymptotic variance when $h_{var}(x)$ is employed is

$$AV_{\mathbf{X},\mathbf{Y}}\left[\widehat{m_{h_{var}}}(x)\right] = \left[\frac{\left[\int K^2(t)dt\right]^{\frac{4}{5}} \left[\sigma^2(x)\right]^{\frac{4}{5}}}{\left[\int t^2 K(t)dt\right]^{-\frac{2}{5}} \left[\frac{\alpha^2(x)}{f_X^6(x)}\right]^{-\frac{1}{5}}} \right] n^{-\frac{4}{5}}. \qquad (3.10)$$

Nishida and Kanazawa (2010) showed in proposition 1 that the univariate VS bandwidth produces the asymptotic variance smaller on some part of the support than the univariate fixed bandwidth,

$$\min_{x\in I^1} AV_{\mathbf{X},\mathbf{Y}}\left[\widehat{m_{h_{fixed}}}(x)\right] \leq AV_{\mathbf{X},\mathbf{Y}}\left[\widehat{m_{h_{VS}}}(x)\right] \leq \max_{x\in I^1} AV_{\mathbf{X},\mathbf{Y}}\left[\widehat{m_{h_{fixed}}}(x)\right].$$

Similarly, in comparison with the variable bandwidth, we obtain the same magnitude relation,

$$\min_{x\in I^1} AV_{\mathbf{X},\mathbf{Y}}\left[\widehat{m_{h_{var}}}(x)\right] \leq AV_{\mathbf{X},\mathbf{Y}}\left[\widehat{m_{h_{VS}}}(x)\right] \leq \max_{x\in I^1} AV_{\mathbf{X},\mathbf{Y}}\left[\widehat{m_{h_{var}}}(x)\right]. \quad (3.11)$$

The proof of (3.11) is in appendix B.

These two relations indicate that the worst case does not happen where the homoscedastic variance produced by (3.7) is larger over all regressor variables than the heteroscedastic variances produced by h_{fixed} or $h_{var}(x)$.

3.1.2 Comparison of the three bandwidths in terms of AMISE.

Nishida and Kanazawa (2010) showed in proposition 3 that the univariate VS bandwidth may in some cases achieve smaller AMISE than its fixed counterpart. This argument is characterized by the ratio of two density functions, $\beta(x) = \int_{I^1}\sigma^2(x)dx / \int_{I^1} f_X(x)dx$ because of the equation,

$$\begin{aligned} &AMISE^5(m(x), \widehat{m_{h_{fixed}}}(x)) - AMISE^5(m(x), \widehat{m_{h_{VS}}}(x)) \\ =&C_0\left(\frac{5}{4}\right)^5 \cdot n^{-4}\int_{I^1}\left[1-\beta^4(x)\right]\cdot\frac{\alpha^2(x)}{f_X(x)}dx. \end{aligned} \quad (3.12)$$

The term (3.12) takes positive value when the term $\alpha^2(x)/f_X(x)$ is large in the area where $\beta(x) < 1$, while the $\alpha^2(x)/f_X(x)$ is small in the area $\beta(x) > 1$.

To note this, we introduced the case where the VS bandwidth outperforms the fixed bandwidth in terms of AMISE in the illustrative example 1 in Nisihida and Kanazawa (2010). In the example, we set $f_X(x) = 2(2-x)/3$ and $m(x) = (x-2)^{-2}$ on the domain $I^1 = [0,1]$ to determine $\alpha^2(x)/f_X(x) =$

$(8/3)(2-x)^{-7}$ is a monotone increasing function. At the same time, we calculated the ratio of two AMISE's,

$$r(f_X(x), \sigma^2(x), m(x)) = \frac{AMISE(m(x), \widehat{m_{h_{VS}}}(x))}{AMISE(m(x), \widehat{m_{h_{fixed}}}(x))}, \tag{3.13}$$

by the four types of conditional variance functions, $\sigma^2(x) = 0.1(x+1)$, 0.1, $0.1(2-x)$ and $0.1(3-2x)$. We show in Table 2 the asymptotic homoscedastic variance in (3.8), the maximal and minimal values of the asymptotic heteroscedastic variance in (3.9), and the ratio of the two AMISE's. We observe the ratio r moves from 1.5133 to 0.8904. The practical implication obtained here is that VS bandwidth outperforms the fixed bandwidth in terms of AMISE if the conditional variance $\sigma^2(x)$ diminishes as x increases under the assumption that the regression function $m(x)$ is a monotone increasing.

$\sigma^2(x)$	homo.var. in (3.8)	hetero.var. in (3.9)		r in (3.13)
		min. (arg.min)	max. (arg.max)	
$0.1(x+1)$	0.0981	0.0324 ($x=0$)	0.1297 ($x=1$)	1.5133
0.1	0.0587	0.0351 ($x=0$)	0.0703 ($x=1$)	1.2525
$0.1(2-x)$	0.0648	0.0648	0.0648	1.0000
$0.1(3-2x)$	0.0727	0.0612 ($x=1$)	0.0918 ($x=0$)	0.8904

Table 2: The result of numerical calculation for illustrative example 1 in Nishida and Kanazawa (2010).

Similarly, in the illustrative example 2 in Nisihida and Kanazawa (2010), we investigated how fat-tailedness of the distribution of X's affects the ratio r of the AMISE's in (3.13). This is because we expect low density $f_X(x)$ at and near the boundaries of $I = [0,1]$ makes the proposed VS bandwidth very wide, rendering the bias not to be ignorable. The main result from the illustrative example 2 is that the VS bandwidth is more serviceable than fixed bandwidth when the distribution of X's are flatter.

On the other hand, in comparison with the MSE minimizing variable bandwidth, it is proved by the Hölder's inequality that the MSE minimizing variable bandwidth always achieve smaller AMISE than the VS bandwidth,

$$AMISE(m(x), \widehat{m_{h_{var}}}(x)) \leq AMISE(m(x), \widehat{m_{h_{VS}}}(x)).$$

3.1.3 Estimation of the univariate VS bandwidth

We show one estimating procedure of the univariate VS bandwidth. In view of (3.7), we estimate the univariate VS bandwidth by the plug-in approach,

$$\widehat{h_{VS}}(x) = \widehat{h_0} \cdot \frac{\widehat{\sigma^2}(x)}{\widehat{f_X}(x)}, \tag{3.14}$$

where $f_X(x)$ and $\sigma^2(x)$ are nonparametrically estimated and the constant term h_0 is estimated by cross-validation method as in Marron and Härdle (1986). As a candidate for $\widehat{f_X}(x)$, we employ kernel density estimator, where $\widehat{h_f}$ is chosen to minimize the least-squares cross-validation as in Rudemo (1982) and Bowman (1984). The $\widehat{h_f}$ so obtained is known asymptotically equivalent to the MISE optimized bandwidth of h_f as in Hall (1983). As a candidate for $\widehat{\sigma^2}(x)$, we employ the residual based estimator,

$$\frac{\sum_{i=1}^{n} K\left(\frac{x-X_i}{h_v}\right)(Y_i - \widehat{m_{h_v}}(X_i))^2}{\sum_{i=1}^{n} K\left(\frac{x-X_i}{h_v}\right)}, \tag{3.15}$$

as in, for instance, Fan and Yao (1998).

However when estimating $\sigma^2(x)$, we generally need to estimate $m(x)$ beforehand, which needs an estimator of $\sigma^2(x)$ beforehand because we are in the variance-stabilizing setting. Hence an iterative method that enables us to estimate $\sigma^2(x)$ and $m(x)$ simultaneously, like $(\widehat{\sigma^2}^{(0)}(x) \to \widehat{m}^{(0)}(x) \to \widehat{\sigma^2}^{(1)}(x) \to \widehat{m}^{(1)}(x) \to ...)$, is required. The following is an idea of the iterative estimating procedure.

Stage 1: Estimation of $h_v^{(t)}$.

- Obtain $\bar{Y}$ as an initial value and compute squared residuals $r^{2,(0)}(X_i) = (Y_i - \bar{Y})^2$, $i = 1, ..., n$.
- Obtain bandwidth $h_v^{(0)}$ that minimizes the cross-validation statistic,

$$CV(h_v^{(0)}) = \frac{1}{n}\sum_{i=1}^{n}(r^{2,(0)}(X_i) - \widehat{\sigma^2_{-i,h_v^{(0)}}}(X_i))^2,$$

with respect to $h_v^{(0)}$, where

$$\widehat{\sigma^2_{-i,h_v^{(0)}}}(X_i) = \frac{\sum_{j=1;j\neq i}^{n} K\left(\frac{X_j - X_i}{h_v^{(0)}}\right) r^{2,(0)}(X_i)}{\sum_{j=1;j\neq i}^{n} K\left(\frac{X_j - X_i}{h_v^{(0)}}\right)}.$$

The estimated bandwidth is $\widehat{h_v}^{(0)}$.

- Obtain nonparametric estimator of conditional variance as

$$\widehat{\sigma^2_{\widehat{h_v}^{(0)}}}(x) = \frac{\sum_{i=1}^{n} K\left(\frac{x - X_i}{\widehat{h_v}^{(0)}}\right) r^{2,(0)}(X_i)}{\sum_{i=1}^{n} K\left(\frac{x - X_i}{\widehat{h_v}^{(0)}}\right)}.$$

Stage 2: Estimation of $h_0^{(t)}$.

- Obtain bandwidth $\widehat{h_0}^{(0)}$ that minimizes the cross-validation statistic,

$$CV(\widehat{h_0}^{(0)}) = \frac{1}{n}\sum_{i=1}^{n}(Y_i - \hat{m}_{-i,\widehat{h_{VS}}^{(0)}}(X_i))^2,$$

with respect to $\widehat{h_0}^{(0)}$, where

$$\widehat{h_{VS}}^{(0)}(X_i) = \frac{\widehat{\sigma^2_{\widehat{h_v}^{(0)}}}(X_i)}{\widehat{f_{h_f}}(X_i)} \cdot \widehat{h_0}^{(0)}, i = 1, ..., n.$$

- Obtain NW type estimator $\widehat{m_{\widehat{h_{VS}}^{(0)}}}(X_i) = \frac{\sum_{j=1}^{n} K\left(\frac{X_i - X_j}{\widehat{h_{VS}}^{(0)}(X_i)}\right) Y_j}{\sum_{j=1}^{n} K\left(\frac{X_i - X_j}{\widehat{h_{VS}}^{(0)}(X_i)}\right)}$, $i = 1, ..., n$.
- Compute squared residuals $r^{2,(1)}(X_i) = (Y_i - \widehat{m_{\widehat{h_{VS}}^{(0)}}}(X_i))^2$, $i = 1, ..., n$.

Stage 3

- Substitute $r^{2,(0)}(X_i)$ in **Stage 1** for $r^{2,(1)}(X_i)$ in **Stage 2** and repeat **Stage 1** and **Stage 2** until $\widehat{h_0}^{(t)}$ converges.

Output

We obtain $\widehat{h_v}$ and $\widehat{h_0}$ alternately.

Simulation

We performed the simulation of the estimating procedure above. We estimated two bandwidths $\widehat{h_v}$ and $\widehat{h_0}$ by $N = 50$ sets of sample size n randomly generated by the following functions $m(x) = 0.5x^2$, $\sigma^2(x) = 0.1(0.5|x| + 0.1)$ and $f_X(x) = N(0,4)\{x : -1 < x < 1\}$ on the domain $[-1,1]$. In the simulation, a data set is generated on $[-3.0, 3.0]$ to eliminate boundary effect. The result and the plots of the estimated regression function are respectively on Table 3 and Figure 1. The result on the table says that $\widehat{h_0}$ is well estimated while the estimation of $\widehat{h_v}$ is poor. We need further investigation on this point.

	bandwidth $\widehat{h_0}$				bandwidth $\widehat{h_v}$			
n	mean	median	std.dev.	$\widehat{h_0}/h_0$	mean	median	std.dev	$\widehat{h_v}/h_v$
500	5.9423	6.1234	1.5248	0.9203	0.6979	0.7195	0.2740	0.5285
1000	6.4420	6.4548	1.6702	0.9977	0.7359	0.7918	0.2334	0.5573
5000	6.4633	6.7947	1.0918	1.0001	0.6668	0.7240	0.1594	0.5050
Theoretical value : $h_v \cdot n^{1/5} = 1.3204$								
Theoretical value : $h_0 \cdot n^{1/5} = 6.4567$								
Theoretical value : $h_f \cdot n^{1/5} = 2.2391$								

Table 3: The result of the simulation. In the estimation, we employed Gaussian kernel.

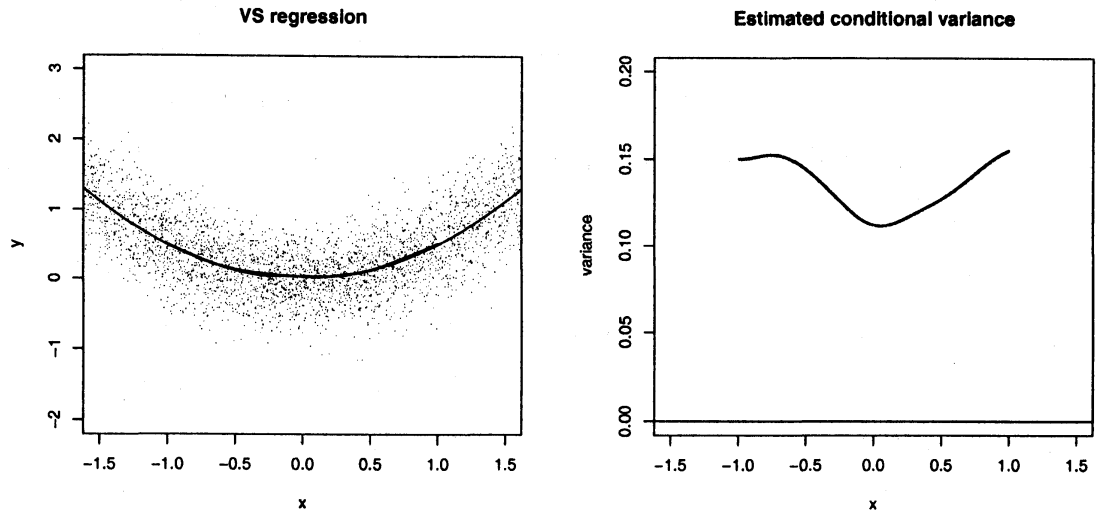

Figure 1: The left panel plots the variance-stabilizing regression obtained by the estimating procedure. The right panel is the nonparametric regression of the conditional variance function obtained in the process of the estimating procedure.

3.2 Bivariate VS bandwidth matrix: $p = 2$

When $p = 2$, we can derive the optimal parameters $\eta_0^*(\mathbf{x})$ and $\eta_1^*(\mathbf{x}) = 1 - \eta_0^*(\mathbf{x})$ that minimize (3.6) at every $\mathbf{x}$. The optimal $\eta_0^*(\mathbf{x})$ so obtained is

$$\eta_0^*(\mathbf{x}) = \begin{cases} \dfrac{\log\left[\left|\frac{\alpha_1(\mathbf{x})}{\alpha_0(\mathbf{x})}\right|\left[\frac{\sigma^2(\mathbf{x})}{f_{\mathbf{X}}(\mathbf{x})}\right]^2\right]}{4\log\left[\frac{\sigma^2(\mathbf{x})}{f_{\mathbf{X}}(\mathbf{x})}\right]}, & \text{if } \alpha_i(\mathbf{x}) \neq 0, \quad i = 1,2, \\ \text{any value satisfying } \eta_0^*(\mathbf{x}) + \eta_1^*(\mathbf{x}) = 1, & \text{if } \alpha_i(\mathbf{x}) = 0, \quad i = 1,2. \end{cases} \tag{3.16}$$

We give an interpretation to the set of parameters $\eta_i(\mathbf{x})$, $i = 1,2$. Compare the two terms (2.11) and (3.4) that appear in the integrated squared bias terms. Since the term $[\sigma^2(\mathbf{x})/f_{\mathbf{X}}(\mathbf{x})]^{2\eta_i(\mathbf{x})}$ that appears in (3.4) can be considered as the penalty for the variance stabilization, the parameter $\eta_i(\mathbf{x})$ can be interpretted as the fractional rate of the penalty for variance stabi-

lization between the axes. For example, if $\eta_0(\mathbf{x}) = 1/2$ and $\eta_1(\mathbf{x}) = 1/2$, the penalty for variance stabilization is equally distributed to each axis at the point $\mathbf{x}$. Similarly, if $\eta_0(\mathbf{x}) = 0$ and $\eta_1(\mathbf{x}) = 1$, all the penalty for variance stabilization is incurred to the axis x_1 and no penalty is incurred to the x_0 axis at the point $\mathbf{x}$ and vice versa.

In the following, we arrange the numerical cases given $m(x_0, x_1)$, $\sigma^2(x_0, x_1)$ and $f_{X_0,X_1}(x_0, x_1)$, and calculate theoretical AMISE's by the different bandwidth matrices. The one is calculated by the fixed bandwidth matrix in (2.10) denoted by $AMISE_{fixed}$. The another is calculated by the VS bandwidth matrix in (3.3) denoted by $AMISE_{VS}$. As for the local parameters $\eta_i(\mathbf{x})$ of the VS bandwidth matrix, we employ two types, the optimal one in (3.16) and the universal one determined arbitrarily. The results are on Table 4.

Case1: $m(x_0, x_1) = -3x_0^2 - 3x_1^2$, $\sigma^2(x_0, x_1) = 0.25$, $f_{X_0,X_1}(x_0, x_1) \sim N(0, 0, \sigma_{X_0}, \sigma_{X_1}, \rho)$ trucated on $I^2 = [-0.5, 0.5] \times [-0.5, 0.5]$ and $\sigma_{X_0} = \sigma_{X_1}$, $\rho = 0.0$.

Case2: $m(x_0, x_1) = -3x_0^2 + 2x_0x_1 - 3x_1^2$, $\sigma^2(x_0, x_1) = 0.25$, $f_{X_0,X_1}(x_0, x_1) \sim N(0, 0, \sigma_{X_0}, \sigma_{X_1}, \rho)$ trucated on $I^2 = [-0.5, 0.5] \times [-0.5, 0.5]$ and $\sigma_{X_0} = \sigma_{X_1}$, $\rho = 0.0$.

Case3: $m(x_0, x_1) = \sin(2\pi x_0) + x_1$, $\sigma^2(x_0, x_1) = 0.25$, $f_{X_0,X_1}(x_0, x_1) \sim N(0, 0, \sigma_{X_0}, \sigma_{X_1}, \rho)$ trucated on $I^2 = [-0.5, 0.5] \times [-0.5, 0.5]$ and $\sigma_{X_0} = \sigma_{X_1}$, $\rho = 0.0$.

One main result obtained from the table is that $AMISE_{VS}$ with optimal local parameter, $\eta_i^*(\mathbf{x})$, is always smaller than that with universal one determined arbitrarily. This result is a matter of course because we optimized the local parameter so as to minimize AMISE. Another obtained result is that $AMISE_{VS}$ can sometimes outperform $AMISE_{fixed}$. See the cases 1,2 when $\sigma_{X_0} = 1$ and the case 3 when $\sigma_{X_0} = 0.5, 1$. In the case 3, we observe that $AMISE_{VS}$ with universal local parameter can outperform $AMISE_{fixed}$. We still need to investigate in what situation such preferable cases happen.

	Case 1		Case 2		Case 3	
	$\sigma_{X_0} = 0.5$	$\sigma_{X_0} = 1.0$	$\sigma_{X_0} = 0.5$	$\sigma_{X_0} = 1.0$	$\sigma_{X_0} = 0.5$	$\sigma_{X_0} = 1.0$
η_0	$AMISE_{VS}$					
-1.00	19.0226	59.2533	20.3504	63.5712	34.6694	95.7336
-0.75	12.3613	37.8475	13.2567	40.6537	21.8813	61.1660
-0.5	8.1146	24.2137	8.7331	26.0450	14.0122	39.1085
-0.25	5.4044	15.5581	5.8481	16.7690	9.1076	25.0236
0.00	3.7008	10.1466	4.0424	10.9824	6.0092	16.0240
0.25	2.7188	7.0098	3.0153	7.6588	4.0280	10.2776
0.5	2.3831	5.9114	2.6710	6.5139	2.7714	6.6803
0.75	2.7188	7.0098	3.0153	7.6588	2.1334	4.9928
1.00	2.7008	10.1466	4.0424	10.9824	2.3486	5.9398
1.25	5.4044	15.5581	5.8481	16.7690	3.4085	9.0462
1.5	8.1146	24.2137	8.7331	26.0450	5.2671	14.1915
1.75	12.3613	37.8475	13.2567	40.6537	8.2748	22.3276
2.00	19.0226	59.2533	20.3504	63.5712	13.1392	35.1611
Opt $\eta_0^*(\mathbf{x})$	2.1353	5.2776	2.5214	6.0976	2.0456	4.9428
	$AMISE_{fixed}$					
	2.0066	5.6909	2.2366	6.2596	2.4230	6.4551

Table 4: The result of numerical calculation. We calculate theoretical AMISE's for different bandwidth matrices, $\mathbf{H}_{fixed}$ and $\mathbf{H}_{VS}(\mathbf{x})$.

4 Concluding remark

In this paper, we limited the VS bandwidth matrix to the diagonal one by the sphering approach and derived the theoretical form of the VS bandwidth defined by two types of parameters h_0, global one, and $\eta_i(\mathbf{x})$, local one.

In the univariate case, we compared three bandwidths h_{fixed}, $h_{var}(x)$ and $h_{VS}(x)$ in terms of local variance and AMISE. The implication obtained here is that, in terms of local variance, $h_{VS}(x)$ can outperform both h_{fixed} and $h_{var}(x)$ on some part of the domain. In terms of AMISE, $h_{VS}(x)$ can sometimes outperform h_{fixed}, but never outperform $h_{var}(x)$.

However, we observe a case where $h_{VS}(x)$ can be superior to the variable bandwidth $h_{var}(x)$ in another point of view. At the point x^* that satisfies $\alpha^2(x^*) = 0$, the bandwidth $h_{var}(x^*)$ takes infinitely large value. If $h_{var}(x^*)$ approaches infinity, $\widehat{m_{h_{var}}}(x^*)$ approaches to $\bar{Y}$. Then, if $m(x^*)$ does not happen to agree with $\bar{Y}$, $\widehat{m_{h_{var}}}(x^*)$ has unconformity at the point x^*. As for $h_{VS}(x)$ in (3.7), such unconformity does not happen as long as $f_X(x) \neq 0$. If we consider this case, VS bandwidth can be serviceable. Examples of this kind of problem with variable bandwidth are illustrated in Nishida and Kanawaza (2009) and can also be observed in the case of $p \geq 2$.

In the bivariate case, we derived the optimal local parameters $\eta_i^*(\mathbf{x})$ and give an interpretation. As a future study, we need to give further investigations to the bivariate case and develop a estimating procedure that can be compared to the existing MSE based estimator as in Doksum (2000). We also need to investigate the behavior of VS regression of $p \geq 3$.

Appendix A

Suppose that the kernel function satisfy:

K1 Kernel is a density function $K(\mathbf{t})$ symmetric about zero.

K2 $\int \cdots \int \mathbf{t}\mathbf{t}^T K(\mathbf{t})d\mathbf{t} < \infty$.
K3 $\int \cdots \int K^2(\mathbf{t})d\mathbf{t} < \infty$.
K4 $\int \cdots \int |K(\mathbf{t})|d\mathbf{t} < \infty$.
K5 $\lim_{\mathbf{t}\to\pm\infty} \mathbf{t}K(\mathbf{t}) \to 0$.

We place the following standard set of assumptions:

A1 $\mathbf{H}_\mathbf{X} \to 0$ as n goes to infinity.
A2 $n\mathbf{H}_\mathbf{X} \to \infty$ as n goes to infinity.
A3 The density of $\mathbf{X}$ is $0 < f_\mathbf{X}(\mathbf{x}) < \infty$ on a compact support I^p.
A4 The $f_\mathbf{X}(\mathbf{x})$ is bounded continuously differentiable on I^p.
A5 The conditional variance $\sigma^2(\mathbf{x})$ is continuous with respect to $\mathbf{x}$ and $0 < \sigma^2(\mathbf{x}) < \infty$ on I^p.
A6 The regression function $m(\mathbf{x})$ is twice bounded continuously differentiable and is bounded and not identically constant on I^p.

Appendix B

We obtain the following relation between the two variances (3.9) and (3.10),

$$AV^5_{\mathbf{X},\mathbf{Y}}\left[\widehat{m_{h_{var}}}(x)\right] - AV^5_{\mathbf{X},\mathbf{Y}}\left[\widehat{m_{h_{VS}}}(x)\right]$$
$$= \frac{\left[\int K^2(t)dt\right]^4}{\left[\int t^2K(t)dt\right]^{-2}} \frac{\int_{I^1} \frac{\sigma^8(x)\alpha^2(x)}{f_X^5(x)}dx}{fx(x)} \left[\frac{\frac{\sigma^8(x)\alpha^2(x)}{f_X^5(x)}}{\int_{I^1} \frac{\sigma^8(x)\alpha^2(x)}{f_X^5(x)}dx} - f_X(x)\right]. \qquad \text{(B.1)}$$

This is enough to show (3.11) because (B.1) is a subtraction of the two density functions defined on the common domain I^1.

References

Bowman, A, W. (1984). An alternative method of cross-validation for the smoothing of density estimates. *Biometroka*, **71**, pp353-360.

Doksum, K., Peterson, D., Samarov, A. (2000). On variable bandwidth selection in local polynomial regression. *Journal of the Royal Statistical Society B*, **62**(3), pp431-448.

Fan, J., Gijbels, I. (1992). Variable bandwidth and local linear regression smoothers. *Annals of statistics*, **20**(4), pp2008-2036.

Fan, J., Yao, Q. (1998). Efficient estimation of conditional variance functions in stochastic regression. *Biometrika*, **85**(3), pp645-660.

Fukunaga, K. (1972). Introduction to statistical pattern recognition. Academic press, New York.

Hall, P. (1983). Large sample optimality of least square cross-validation in density estimation. *Annals of statistics*, **11**, pp1156-1174.

Härdle, W., Müller, M., Sperlich, S., Werwatz, A. (2004). Nonparametric and Semiparametric Models. Springer, Berlin and Heidelberg.

Marron, J.S., Härdle, W. (1986). Random approximations to some measures of accuracy in nonparametric curve estimation. *Journal of multivariate Analysis*, **20**, pp91-113.

Nadaraya, E.A. (1964). On estimating regression. *Theory of probability and its applications*, **9**, pp141-142.

Nadaraya, E.A. (1965). On nonparametric estimation of density functions and regression curves. *Theory of probability and its applications*, **10**, pp186-190.

Nadaraya, E.A. (1970). (Translated by B.Seckler from Russian) Remarks on nonparametric estimates for density functions and regression curves. *Theory of probability and its applications*, **15**, pp134-137.

Nishida, K., and Kanazawa, J.Y. (2009). A note on the variance-stabilizing nonparametric regression. *RIMS Kôkyûroku*, **1621**, pp65-87. Research Institute for Mathematical Sciences, Kyoto University, Japan.

Nishida, K., and Kanazawa, J.Y. (2010). On variable variance-stabilizing bandwidth for the Nadaraya-Watson regression estimator. *Department of Social Systems and Management Discussion Paper Series*, **1257**, The Graduate School of Systems and Information Engineering, University of Tsukuba, Japan.

Rudemo, M. (1982). Empirical choice of histgrams and kernel density estimators. *Scandinavian Journal of Statistics*, **9**, pp65-78.

Wand, M.P., Jones, M.C. (1993). Comparison of smoothing parametrizations in bivariate kernel density estimation. *Journal of the American Statistical Association*, **88**(422), pp520-528.

Watson, G.S. (1964). Smooth regression analysis. *Sankhyā, Series A*, **26**, pp359-372.

Watson, G.S., Leadbetter, M.R. (1963). On the estimation of probability density, I. *Annals of Mathematical statistics*, **34**, pp480-491.