数理解析研究所講究録
第1621巻 2009年 65-87

A note on the Variance-Stabilizing nonparametric regression estimation

Kiheiji Nishida and Yuichiro Kanazawa [1]

1 Introduction

Nonparametric regression models are often used to estimate regression function that needs to incorporate local information of the data into the estimation. There are many types of nonparametric regression models like the Nadaraya-Watson estimator (henceforce NW) (Nadaraya 1964, 1965, 1970, Watson 1963, 1964), the locally polynomial estimator (e.g. Härdle and Müller 2004 p94-), the k-Nearest-Neighbor estimator (e.g. Mack 1981) and so on. However it has been pointed out that these kinds of nonparametric regression estimators have heteroscedastic variance. This note introduces an idea to stabilize the variance of a nonparametric regression estimator, the NW estimator with one explanatory variable. With the idea of variance-stabilization,

[1]Kiheiji Nishida is a student at the Graduate School of Systems and Information Engineering, University of Tsukuba. Yuichiro 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 Yuichiro Kanazawa. His e-mail address is kanazawa@sk.tsukuba.ac.jp.

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

we show brief summaries of the NW estimator and related topics, especially the local and global bandwidth selection rules and its estimator.

2 Brief summaries of the past researches on the NW estimator.

2.1 The NW estimator and its bias and variance.

We employ the following setup. Suppose that we have n-pairs of random variables (X_i, Y_i), $i = 1, \cdots, n$. We assume x_i, $i = 1, \cdots, n$, are the realizations of i.i.d. random variable X whose density function is denoted as $f_X(x)$. Let u_i, $i = 1, \cdots, n$, be the realizations of the disturbance random variable U_i. We assume that $U_i|X_i$ are i.i.d. given X_i and that U_i are independent with X_j, $j \neq i$. We further assume that the conditional moments of $U|X$ as,

$$E_{U|X}[U|X = x] = 0, \quad E_{U|X}[U^2|X = x] = \sigma^2(x). \tag{2.1}$$

We assume that the response Y is influenced by the explanatory variable X as,

$$Y_i = m(X_i) + U_i = E_{Y|X}(Y_i|X_i) + U_i, \tag{2.2}$$

where $m(X_i) = E_{Y|X}(Y_i|X_i) = \int y f_{Y|X}(y|x)dy = \int y \frac{f_{X,Y}(x,y)}{f_X(x)}dy$. By replacing $f_{X,Y}(x, y)$ and $f_X(x)$ with the corresponding kernel bivariate and univariate estimates with the multiplicative kernel $K_{X,Y}(x, y) = K_X(x)K_Y(y)$ on the numerator and the kernel $K_X(x)$ on the denominator with the band-

width on X as h_x, we arrive at the NW estimator at $X = x$ as,

$$\hat{m}_{h_x}(x) = \frac{\sum_{i=1}^{n} K_X\left(\frac{x-X_i}{h_x}\right) Y_i}{\sum_{i=1}^{n} K_X\left(\frac{x-X_i}{h_x}\right)}. \tag{2.3}$$

With (2.3) and an additional assumption that the bandwidth is a decreasing function of sample size n written as,

$$h_x \to 0, \; nh_x \to \infty, \text{ as } n \text{ goes to infinity}, \tag{2.4}$$

bias and variance of the NW estimator were derived (see Pagan and Ullah 1991) as,

$$\begin{aligned} & E_{\mathbf{X},\mathbf{Y}}[\hat{m}_{h_x}(x)] - m(x) \\ =& \frac{h_x^2}{2f_X(x)} \left[\int t^2 K_X(t)dt\right] \left[2m^{(1)}(x)f_X^{(1)}(x) + m^{(2)}(x)f_X(x)\right] \\ & + O\left(\frac{1}{nh_x}\right) + o(h_x^2), \end{aligned} \tag{2.5}$$

and,

$$V_{\mathbf{X},\mathbf{Y}}[\hat{m}_{h_x}(x)] = \frac{1}{nh_x} \cdot \frac{\sigma^2(x)}{f_X(x)} \left[\int K_X^2(t)dt\right] + O\left(\frac{1}{n}\right) + o\left(\frac{1}{nh_x}\right). \tag{2.6}$$

Notice that the leading term in the variance (2.6) at a point x is a function of a term $\sigma^2(x)/f_X(x)$. Since both $\sigma^2(x)$ and $f_X(x)$ are unlikely to be proportional, the variances of the NW estimator at different points x_1 and x_2 differ in general unless properly controlled by the bandwidth h_x.

2.2 On local NW smoothing parameter selection

Since bias and variance are obtained, we can obtain locally optimal bandwidth in terms of Mean Squared Error (MSE) written as $E_{\mathbf{X},\mathbf{Y}}(\hat{m}_{h_x}(x) -$

$m(x))^2$. Locally balancing the leading-terms of the variance in (2.6) and the bias squared (2.5) for the NW estimator, the MSE-optimized bandwidth is obtained as,

$$h_x^{MSE}(x) = \left[\frac{\left[\int K_X^2(t)dt\right] \sigma^2(x)}{\left[\int t^2 K_X(t)dt\right]^2 \frac{\left(2m^{(1)}(x)f_X^{(1)}(x)+m^{(2)}(x)f_X(x)\right)^2}{f_X(x)}} \right]^{\frac{1}{5}} n^{-\frac{1}{5}}, \quad (2.7)$$

(See e.g. Chu and Marron 1991). Substituting (2.7) into h_x in (2.6), we obtain asymptotic variance with MSE-optimized bandwidth as,

$$V_{\mathbf{X},\mathbf{Y}}\left[\hat{m}_{h_x^{MSE}}(x)\right]$$

$$\approx \left[\frac{\left[\int K_X^2(t)dt\right]^{\frac{4}{5}} \sigma^{\frac{8}{5}}(x)}{\left[\int t^2 K_X(t)dt\right]^{-\frac{2}{5}} \left[\frac{\left(2m^{(1)}(x)f_X^{(1)}(x)+m^{(2)}(x)f_X(x)\right)^2}{f_X^6(x)} \right]^{-\frac{1}{5}}} \right] n^{-\frac{4}{5}}. \quad (2.8)$$

Notice that the variance (2.8) remains heteroscedastic.

MSE-optimized bandwidth causes more serious problem in that the estimated regression with MSE-optimized bandwidth has discontinuities at some x. Let the x^J denote the point that satisfies,

$$\alpha(x^J) \equiv 2m^{(1)}(x^J)f_X^{(1)}(x^J) + m^{(2)}(x^J)f_X(x^J) = 0, \quad (2.9)$$

where the term in (2.9) appears in the denominator of (2.7). Then, MSE-optimized bandwidth will explode at the x^J. It also means that, as x approaches x^J, the NW estimator at the x^J converges weakly to the sample average $\bar{Y}$ because as h_x goes to infinity,

$$\hat{m}_{h_x}(x) = \frac{\sum_{i=1}^n K_X\left(\frac{x-X_i}{h_x}\right) Y_i}{\sum_{i=1}^n K_X\left(\frac{x-X_i}{h_x}\right)} \overset{w}{\rightsquigarrow} \frac{K_X(0)\sum_{i=1}^n Y_i}{nK_X(0)} = \frac{1}{n}\sum_{i=1}^n Y_i. \quad (2.10)$$

We observe how the NW estimator with the MSE-optimized bandwidth jumps by the example below.

Example

$m(x) = x^4\{x : -1.96 \leq x \leq 1.96\}$, $U_i|X_i \sim N(0, \sigma^2(X_i))$, $\sigma^2(x) = 1$, $X_i \sim N(0,1)$. $x^J = 0, \pm\frac{\sqrt{6}}{2}$. For illustration, a sample of size 1000 is generated to obtain X_i, U_i and then $Y_i = X_i^4 + U_i$. These (X_i, Y_i) $i = 1, \cdots, 1000$ are used to compute (2.3) with theoretically calculated $h_x^{MSE}(x)$ in (2.7). The upper panel in Figure 1 plots the MSE-optimized bandwidth and the lower panel describes how the NW estimator comes out in out of the 1000 samples with MSE-optimized bandwidth at every x. Notice from Figure 1 that the NW jumps at the x where the size of bandwidth is very large.

This is no coincidence. Suppose that $f_X(x)$ belongs to exponential family denoted as,

$$f_X(x) = \exp\left[\sum_{i=1}^{k} \eta_i(\theta_1, ..., \theta_k)T_i(x) - B(\theta_1, ..., \theta_k)\right] H(x), \qquad (2.11)$$

where $\eta_i(\theta_1, ..., \theta_k)$'s $i = 1, ..., n$ are its natural parameters. Then we obtain,

$$\begin{aligned} &2m^{(1)}(x)f_X^{(1)}(x) + m^{(2)}(x)f_X(x) \\ &= \left[2m^{(1)}(x)\left\{\frac{H^{(1)}(x)}{H(x)} + \sum_{i=1}^{k} \eta_i(\theta_1, ..., \theta_k)T_i^{(1)}(x)\right\} + m^{(2)}(x)\right] f_X(x). \end{aligned} \qquad (2.12)$$

If we suppose $f_X(x) \neq 0$, the expression in (2.12) is zero whenever,

$$\frac{H^{(1)}(x)}{H(x)} + \sum_{i=1}^{k} \eta_i(\theta_1, ..., \theta_k)T_i^{(1)}(x) = -\frac{m^{(2)}(x)}{2m^{(1)}(x)}, \qquad (2.13)$$

or,

$$m^{(1)}(x) = 0 \quad \text{and} \quad m^{(2)}(x) = 0. \qquad (2.14)$$

As for the lefthand side of (2.13), if we further assume $f_X(x)$ is the standard normal distribution as our example, then,

$$k = 1,\ H(x) = \frac{1}{\sqrt{2\pi}},\ \eta_1 = -\frac{1}{2},\ T_1(x) = x^2, \tag{2.15}$$

and thus (2.13) is written as,

$$x = \frac{m^{(2)}(x)}{2m^{(1)}(x)}. \tag{2.16}$$

If we additionally assume $m(x)$ is a nth-polynomial in (2.16) to be written as,

$$m(x) = \sum_{i=0}^{n} a_i x^i, \tag{2.17}$$

then x must satisfy nth-order equation with constant term,

$$\begin{aligned} &2na_n x^n + 2(n-1)a_{n-1}x^{n-1} \\ &+ \sum_{i=2}^{n-1} \left[2(n-i)a_{n-i} - (n-(i-2))(n-(i-1))a_{n-(i-2)}\right] x^{n-i} \\ &-2a_2 = 0. \end{aligned} \tag{2.18}$$

Let us exclude trivial cases where the coefficients of (2.18) are all zeroes. Under this assumption, if $a_2 = 0$, then $x = 0$ satisfies (2.18). If n is odd, then the polynomial always extend from $-\infty$ to ∞, thereby crossing $f(x) = 0$ somewhere in between. The points $x^J = \pm\sqrt{6}/2$ in the example satisfy (2.13).

On the other hand, (2.14) is satisfied at $x = 0$ of the example. Interpretation of (2.14) is as follows: The fact that $m^{(1)}(x) = 0$ means that the regression function has local extremum at the x : the fact that $m^{(2)}(x) = 0$ holds additionally means that the curvature is zero at the x. Since every line horizontal to x axis typically satisfies (2.14), the NW estimator at x

with MSE-optimized bandwidth of infinite width is just the sample average if a regression function at x has tangent line horizontal to the x axis as we already see.

2.3 On global NW smoothing parameter selection

Reflecting the past experiences on density estimation, Mean Integrated Squared Error (MISE),

$$\begin{aligned} &MISE(\hat{m}_{h_x}(x), m(x)) \\ &= \int_{-\infty}^{\infty} \left[\int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} f_{X_1,\cdots,X_n,Y_1,\cdots,Y_n}(x_1, \cdots, x_n, y_1, \cdots, y_n) \right. \\ &\left. \times [\hat{m}_{h_x}(x) - m(x)]^2 \, dx_1 \cdots dx_n dy_1 \cdots dy_n \right] f_X(x) dx, \end{aligned} \tag{2.19}$$

is often employed in the context of smoothing parameter selection of non-parametric regression model to obtain the global theoretically optimal h_x as,

$$h_{fixed}^{MISE} = \left[\frac{\left[\int K_X^2(t)dt\right] \int \sigma^2(x)dx}{\left[\int t^2 K_X(t)dt\right]^2 \int \frac{\left(2m^{(1)}(x)f_X^{(1)}(x) + m^{(2)}(x)f_X(x)\right)^2}{f_X(x)} dx} \right]^{\frac{1}{5}} n^{-\frac{1}{5}}. \tag{2.20}$$

There are many unknown quantities $f_X(x)$, $f_X^{(1)}(x)$, $m^{(1)}(x)$, $m^{(2)}(x)$ and $\sigma^2(x)$ in (2.20). Especially problematic issue is the presence of $m^{(1)}(x)$ and $m^{(2)}(x)$ to estimate $m(x)$. As such, plug-in approaches do not seem to be encouraging. In its stead, the following cross-validation statistics to be minimized with-respect to h_x is proposed,

$$CV_{MISE}(h_x) = \frac{1}{n} \sum_{i=1}^{n} [Y_i - \hat{m}_{h_x,-i}(X_i)]^2 \hat{f}_{h_x}(X_i), \tag{2.21}$$

where $\hat{m}_{h_x,-i}(X_i)$ is the leave-one-out NW estimator,

$$\hat{m}_{h_x,-i}(X_i) = \frac{\sum_{j=1,j\neq i}^{n} K_X\left(\frac{X_i - X_j}{h_x}\right) Y_j}{\sum_{j=1,j\neq i}^{n} K_X\left(\frac{X_i - X_j}{h_x}\right)}. \tag{2.22}$$

Marron and Härdle (1986) showed that the obtained bandwidth so as to minimize (2.21) is asymptotically equivalent to the MISE-optimized bandwidth (2.20).

Because they employ a global measure of closeness of the estimate to the true curve, they obtain a global smoothing paramater. The $\hat{m}_{h^{MISE}_{fixed}}(x)$ is heteroscedastic across x as we see from (2.23) obtained by substituting (2.20) for (2.6).

$$\begin{aligned} & V_{\mathbf{X},\mathbf{Y}}[\hat{m}_{h^{MISE}_{fixed}}(x)] \\ &= \left[\frac{\left[\int K_X^2(t)dt\right]^{\frac{4}{5}} \left[\int \sigma^2(x)dx\right]^{-\frac{1}{5}}}{\left[\int t^2 K_x(t)dt\right]^{-\frac{2}{5}} \left[\int \frac{\left(2m^{(1)}(x)f_X^{(1)}(x) + m^{(2)}(x)f_X(x)\right)^2}{f_X(x)} dx\right]^{-\frac{1}{5}}} \right] n^{-\frac{4}{5}} \\ & \times \frac{\sigma^2(x)}{f_X(x)} + O\left(\frac{1}{n}\right) + o\left(\frac{1}{nh_x}\right). \end{aligned} \tag{2.23}$$

3 Variance-Stabilizing bandwidth

3.1 Motivation for homoscedastic NW estimator

We see the actual situation to use the NW estimator. We are in the situation to estimate Engel curve, which is the relation between net-income and food expense of households. The upper panel in Figure 2 is the scatter plot of the paired observations of net-income on the horizontal axis and food expense

on the vertical axis of U.K. households in 1983. The data is from Härdle and Müller (2004 p87) and is normalized to show that the point $(1, 1)$ is sample averages and sample size is $n = 6830$. The middle panel in Figure 2 is the strip box plots of the same data and the lower panel is the kernel density estimate of net-income.

Strip boxplots in Figure 2 illustrates that food expense of households increases at the gradually declining rate. In this sense, it seems to be appropriate to employ nonparametric regression model like NW that suits for incorporating local information into the estimation. However, the strip boxplots also illustrate the variability of food expense expands as net-income of households increases. In addition, the lower panel of Figure 2 says that distribution of net-income seems to be bell shaped. Since the functional forms of $\sigma^2(x)$ are unlikely to be proportional to that of $f_X(x)$, we see that the NW estimator is heteroscedastic if applied to the data.

3.2 Theoretical Variance-Stabilizing bandwidth

We propose a method that is applicable to the data above using a new bandwidth. A new bandwidth needs to be variable to correct heteroscedasticity that is being observed across x. At the same time, bandwidth must be able to be estimated from a sample. What about the bandwidth in the form,

$$h_x^{VS}(x) = h_0 \cdot \frac{\sigma^2(x)}{f_X(x)}. \tag{3.1}$$

With (3.1), h_0 can be estimated from a variation of cross-validation statistics (2.21), $f_X(x)$ and $\sigma^2(x)$ can be nonparametrically estimated. At the same time, we can avoid estimating $m^{(1)}(x)$ and/or $m^{(2)}(x)$ in order to esti-

mate $m(x)$. Estimating $f_X(x)$ nonparametrically can be done independently of estimating $m(x)$ nonparametrically, using only x_i, $i = 1, \cdots, n$. Estimating $\sigma^2(x)$ nonparametrically can be done without estimating $m(x)$, for instance, using difference sequences estimators of Müller and Stadtmüller (1987), Brown (2007) and so on.

Justification to use (3.1) are the following. First, the local information on the variance structure $\sigma^2(x)$ as well as on the density $f_X(x)$ is only reflected on the term $\sigma^2(x)/f_X(x)$, making the leading variance term in (2.6) asymptotically uniform over the domain and is only dependent on the type of kernel, on the over all shape of density, and on the overall variance structure. Second, in the domain where the density $f_X(x)$ is low, (3.1) forces one to choose wider bandwidth. The $f_X(x)$ being low means that there are relatively few data points on the region, so aggregating them to estimate the regression function $m(x)$ makes sense. Third, in the domain where the variance $\sigma^2(x)$ is low, the bandwidth should be narrower because small variance implies that the data points on that region are more trustworthy.

Optimization of the constant term h_0 in (3.1) is also necessary. There can be some measures to optimize this but we employ MISE to compare this with the conventional MISE-optimized fixed bandwidth. Then, Variance-Stabilizing bandwidth (henceforce VS) so optimized is,

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

$$= \left[\frac{\left[\int K_X^2(t)dt\right]}{\left[\int t^2 K_X(t)dt\right]^2 \left[\int \frac{\sigma^8(x)\left(2m^{(1)}(x)f_X^{(1)}(x) + m^{(2)}(x)f_X(x)\right)^2}{f_X^5(x)} dx\right]} \right]^{\frac{1}{5}} \cdot n^{-\frac{1}{5}} \cdot \frac{\sigma^2(x)}{f_X(x)}.$$

(3.2)

For this, $\hat{m}_{h^{VS}}(x)$ is homoscedastic up to order $n^{-\frac{4}{5}}$,

$$V_{\mathbf{X},\mathbf{Y}}\left[\hat{m}_{h^{VS}}(x)\right]$$
$$= \left[\frac{\left[\int K_X^2(t)dt\right]^{\frac{4}{5}}}{\left[\int t^2 K_X(t)dt\right]^{-\frac{2}{5}} \left[\int \frac{\sigma^8(x)\left(m^{(2)}(x)f_X(x)+2m^{(1)}(x)f_X^{(1)}(x)\right)^2}{f_X^5(x)}dx\right]^{-\frac{1}{5}}} \right] n^{-\frac{4}{5}}$$
$$+O\left(\frac{1}{n}\right)+o\left(\frac{1}{nh_x}\right). \tag{3.3}$$

Our VS bandwidth has another merit. As far as we assume $f_X(x) \neq 0$, the bandwidth in (3.2) does not become infinity for realistic regression situations. As a result, the NW estimator using bandwidth (3.2) does not jump, for instance, even for the previous example where the MSE-optimized bandwidth does give a discontinuities to $\hat{m}_{h_x^{MSE}}(x)$.

3.3 Numerical example

In the following, we demonstrate VS bandwidth by the example in page 5.

Example revisited

The upper panel in Figure 3 plots MISE optimized, MSE-optimized and VS theoretical bandwidths at each x. The lower panel in Figure 3 plots theoretical variances of the NW estimator with MISE optimized, MSE-optimized and VS bandwidths at each x. The two panels show trade-off between variance and bandwidth. That is, VS bandwidth is variable across x to stabilize variance while the MISE-optimized bandwidth is fixed across x.

The upper panel in Figure 4 plots one example of the NW with VS bandwidth for the first generated sample of $(X_i, Y_i), i = 1, \cdots, 1000$. In the panel, we see that the NW estimator with VS bandwidth has more jagged regions compared to the lower panel in Figure 1. This region corresponds to the area where the bandwidths are very narrow in the example. The NW estimator with VS bandwidth pays penalty in smoothness for stabilizing its variance. Wider (Narrower) bandwidths are assigned to the large (small) variance regions to deflate (inflate) variance relative to other regions. In examples, both tails correspond to large variance regions.

The lower panel in Figure 4 plots the sample variances of NW estimator of $n = 1000$ with MISE-optimized, MSE-optimized and VS bandwidth at each x (calculation is repeated $M = 1000$ times). Notice that the variance stabilization is achieved (failed) on the middle (tails) of the domain. We think that this is brought by the large value of $1/f_X(x)$ and/or $m^{(1)}(x)$ on the tail areas of the example.

4 Estimation of VS bandwidth

To estimate VS bandwidth, we employ combination of plug-in and cross-validation methods. With (3.2), h_0^{MISE} can be estimated from a variation of cross-validation statistics, $f_X(x)$ and $\sigma^2(x)$ can be nonparametrically estimated. There are some candidates for estimators of $f_X(x)$ or $\sigma^2(x)$.

Step1 : Estimation of $\hat{f}_X(x)$

To find sample based bandwidth for kernel density estimator of $f_X(x)$, we employ the well-known least-squares cross-validation method (Rudemo 1982

and Bowman 1984). The bandwidth so obtained is asymptotically equivalent to the MISE optimized bandwidth of kernel density estimator (Hall 1983).

Step2 : Estimation of $\widehat{\sigma^2}(x)$

Possible candidate is the residual based estimators,

$$\widehat{\sigma_R^2}(x)=\frac{\sum_{i=1}^{n} K_X\left(\frac{x-X_i}{h_x}\right)(Y_i-\hat{m}_{h_x}(X_i))^2}{\sum_{i=1}^{n} K_X\left(\frac{x-X_i}{h_x}\right)}, \tag{4.1}$$

or the direct estimator, which is the NW-like kernel estimator of conditional variance function,

$$\begin{aligned}\widehat{\sigma_D^2}(x)&=\widehat{E_{Y|X}[Y^2|X]}-\left(\widehat{E_{Y|X}[Y|X]}\right)^2\\&=\frac{\sum_{i=1}^{n} K_X\left(\frac{x-X_i}{h_x}\right)Y_i^2}{\sum_{i=1}^{n} K_X\left(\frac{x-X_i}{h_x}\right)}-\left[\frac{\sum_{i=1}^{n} K_X\left(\frac{x-X_i}{h_x}\right)Y_i}{\sum_{i=1}^{n} K_X\left(\frac{x-X_i}{h_x}\right)}\right]^2.\end{aligned} \tag{4.2}$$

Fan and Yao (1998) showed that the bias of direct estimator is larger than that of residual estimator. However, to obtain residual based estimator, it is necessary to estimate $\hat{m}_{h_x}(X_i)$ before estimating conditional variance. Therefore, we employ neither and instead difference sequences estimator.

Let $x_{(1)}, x_{(2)}, ..., x_{(n)}$ be ordered observations. Difference sequences estimator yields initial variance estimator $\widehat{\sigma_{DF}^2}(x_{(i)})$ at the data point $x_{(i)}$ by,

$$\widehat{\sigma_{DF}^2}(x_{(i)})=\left(\sum_{j=-r}^{r} d_j Y_{(i+j)}\right)^2, \tag{4.3}$$

where,

$$\sum_{j=-r}^{r} d_j=0, \tag{4.4}$$

$$\sum_{j=-r}^{r} d_j^2=1, \tag{4.5}$$

and,

$$d_j=0, \text{ for } j < -r \text{ or } r < j. \tag{4.6}$$

The quantity $r > 0$ is the order of difference sequences which depends on sample size n. Müller and Stadtmüller (1987) calculated the difference sequences that minimize variance of initial variance estimator of given x_i and order r but we have not found the exact relation between n and r so far. The condition (4.5) is needed to make asymptotically unbiased estimator of $\sigma^2(x_{(i)})$ (see Apendix).

Interpolating so obtained initial variance estimators $\widehat{\sigma^2_{DF}}(x_{(i)})$ $(i = 1, ..., n)$ by the NW estimator, we can estimate conditional variance function,

$$\widehat{\sigma^2_{h_v}(x)} = \frac{\sum_{i=r+1}^{n-r} K_X\left(\frac{x-x_{(i)}}{h_x}\right) \widehat{\sigma^2_{DF}}(x_{(i)})}{\sum_{i=r+1}^{n-r} K_X\left(\frac{x-x_{(i)}}{h_x}\right)}. \tag{4.7}$$

The bandwidth h_v in (4.7) can be estimated by the cross-validation method in (2.21).

Step3 : Estimation of $\widehat{h_0^{MISE}}$

We form the following cross-validation statistics,

$$CV_{VS}(h_0)=\frac{1}{n}\sum_{i=1}^{n}\left(Y_i - \hat{m}_{h^{VS},-i}(X_i)\right)^2 \hat{f}_{h_f}(X_i),$$

$$\text{where} \quad h^{VS}=\frac{\widehat{\sigma^2}_{h_v}(X_i)}{\hat{f}_{h_f}(X_i)}h_0 \text{ given } \widehat{\sigma^2}_{h_v}(x) \text{ and } \hat{f}_{h_f}(x). \tag{4.8}$$

and minimize this function with respect to h_0 to find $\widehat{h_0^{MISE}}$. If we conjecture the true $f_X(x)$ and $\sigma^2(x)$, then we know that the minimizer $\hat{h}_0$ of (4.8) is asymptotically equal to theoretically optimal h_0^{MISE}. We probably be able to prove this by extending the proof by Marron and Härdle (1986). So far,

we have one numerical example, which is presented in Figure 5. The setting of the example is $f_X(x) = (1/\sqrt{2\pi})\exp(-x^2/2)\{x : -1.96 < x < 1.96\}$, $m(x) = x^2\{x : -1.96 < x < 1.96\}$ and $\sigma^2(x) = 1\{x : -1.96 < x < 1.96\}$. When $f_X(x)$ and $\sigma^2(x)$ are estimated, we do not have simulation results.

5 Concluding remarks

We have derived theoretical VS bandwidth. We also showed that the obtained bandwidth is estimable by combination of plug-in and cross-validation methods. However, VS bandwidth does not necessarily generate smaller MISE than MISE fixed bandwidth. In the example on page 5, theoretical MISE with VS bandwidth is 10.2717, which is larger than that with MISE fixed bandwidth by 128%. This is brought by the fact that VS bandwidth achieves variance-stabilization at the cost of larger bias. In line with the above, we are going to find conditions of $m(x)$, $f_X(x)$, and/or $\sigma^2(x)$ under which our VS bandwidth is superior to the conventional one at MISE.

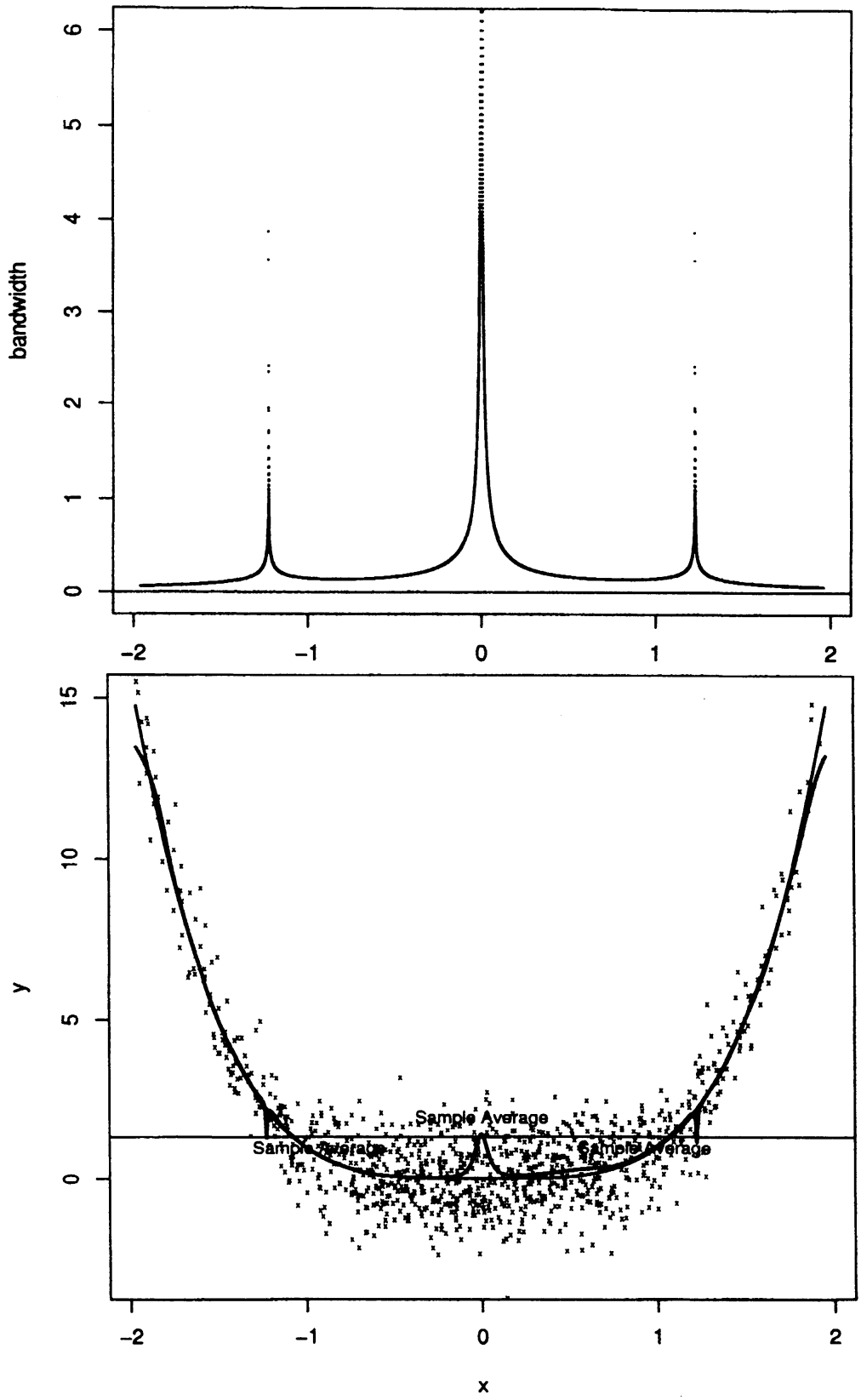


Figure 1: The upper panel plots the MSE-optimized bandwidth and the lower panel describes how the NW estimator comes out in out of the 1000 samples with MSE-optimized bandwidth at every x.

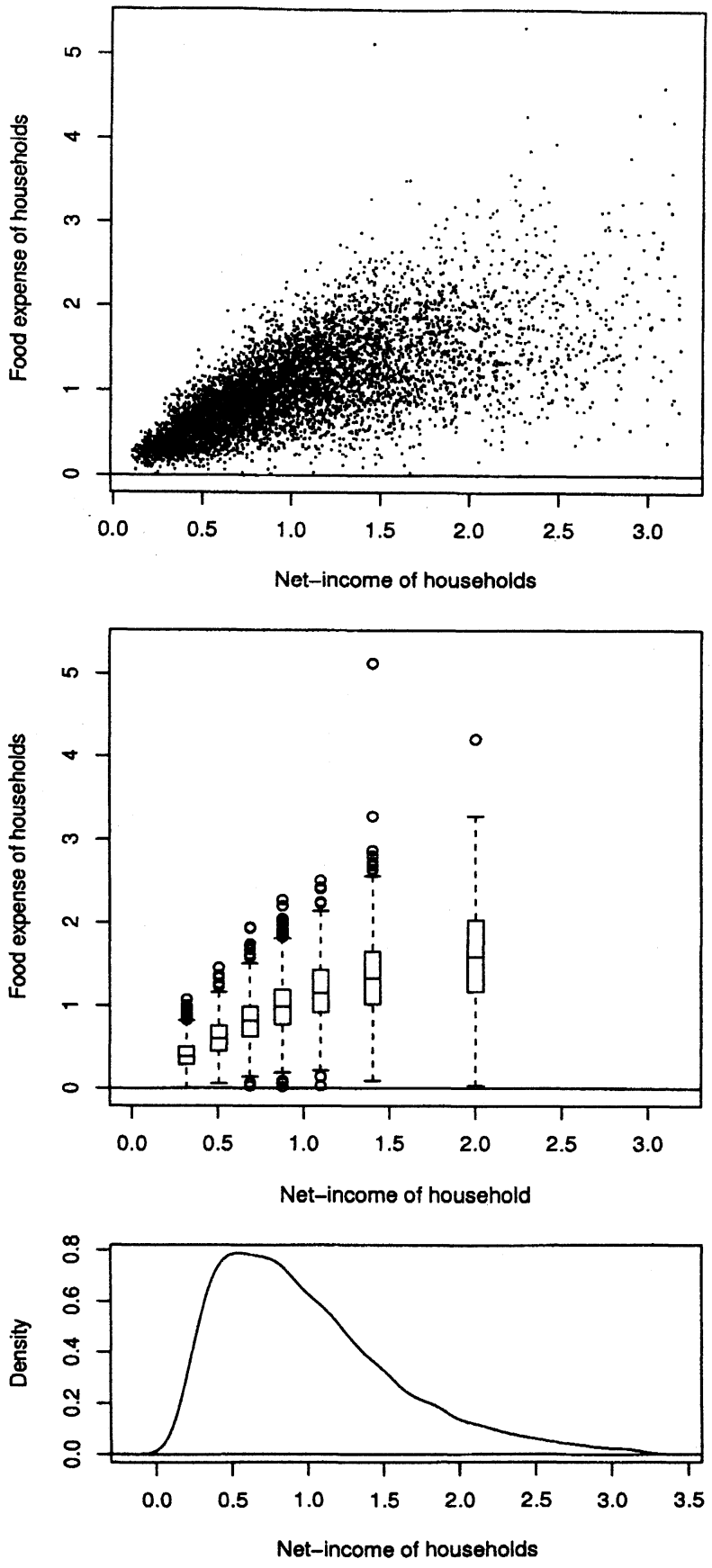
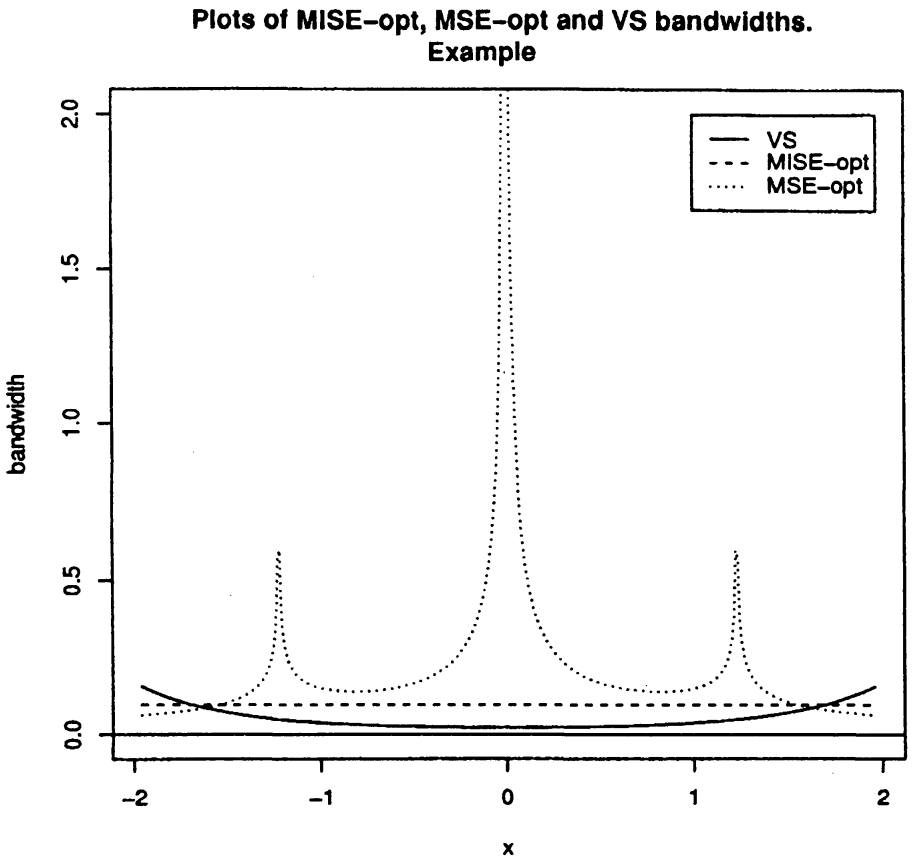

Food expense of households
0
1
2
3
4
5
0.0
0.5
1.0
1.5
2.0
2.5
3.0
Net–income of households
Food expense of households
0
1
2
3
4
5
0.0
0.5
1.0
1.5
2.0
2.5
3.0
Net–income of household
Density
0.0
0.2
0.4
0.6
0.8
0.0
0.5
1.0
1.5
2.0
2.5
3.0
3.5
Net–income of households

Figure 2:


Plots of MISE–opt, MSE–opt and VS bandwidths.
Example
VS
MISE–opt
MSE–opt
bandwidth
0.0
0.5
1.0
1.5
2.0
–2
–1
0
1
2
x

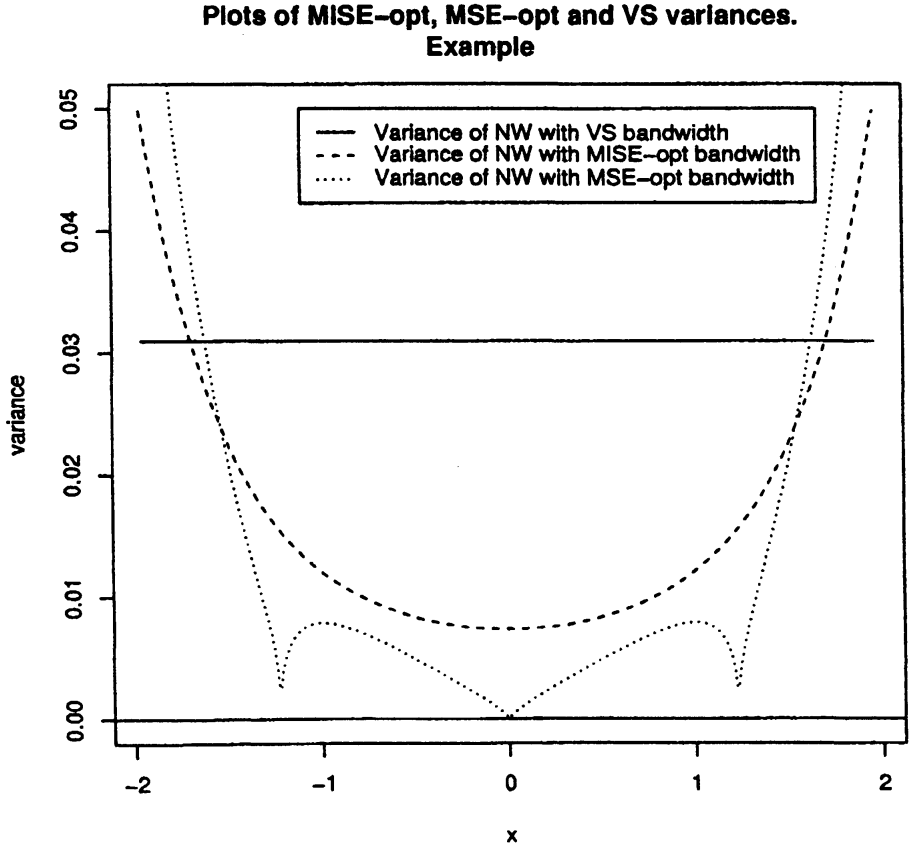

Plots of MISE–opt, MSE–opt and VS variances.
Example
Variance of NW with VS bandwidth
Variance of NW with MISE–opt bandwidth
Variance of NW with MSE–opt bandwidth
variance
0.00
0.01
0.02
0.03
0.04
0.05
–2
–1
0
1
2
x

Figure 3:

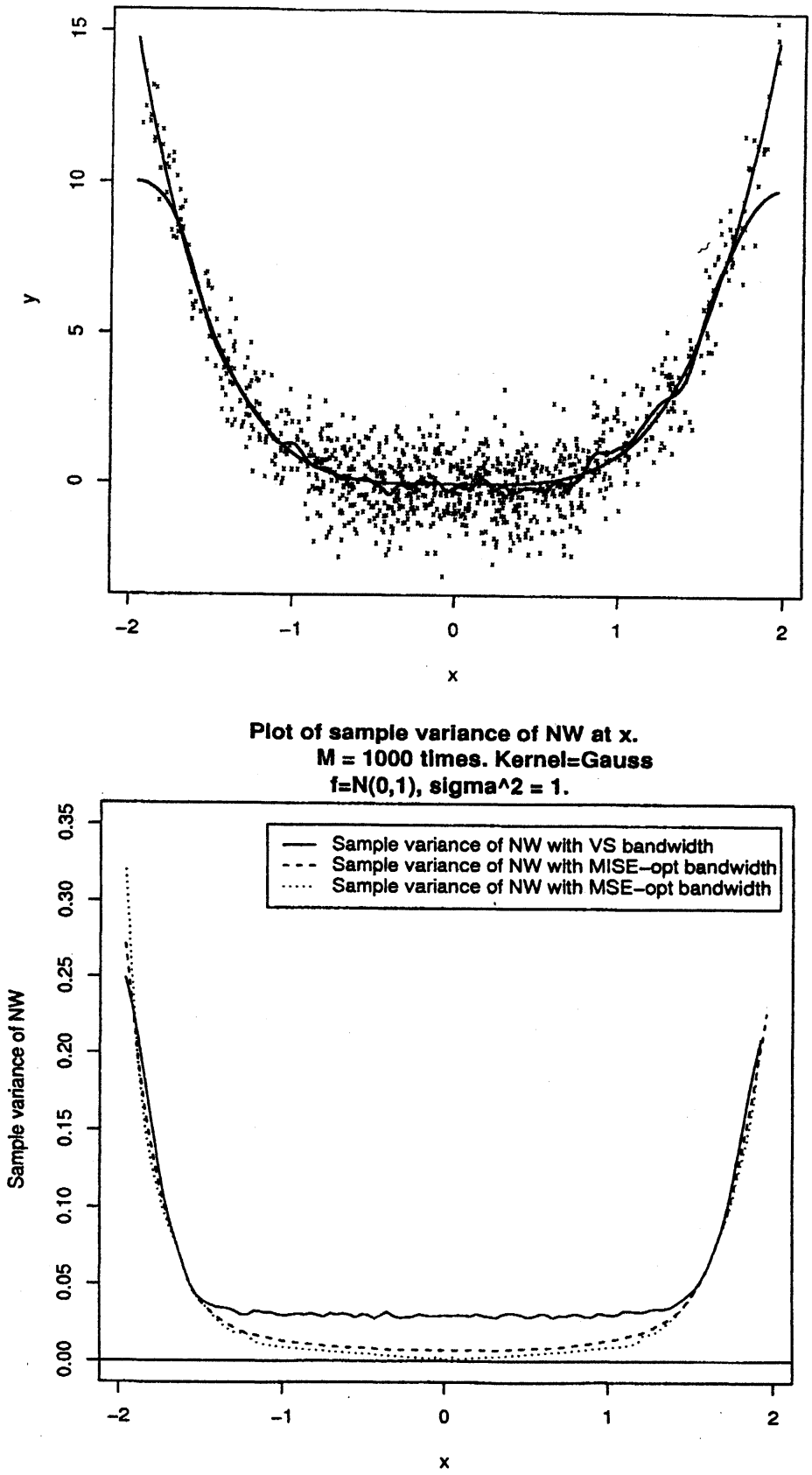
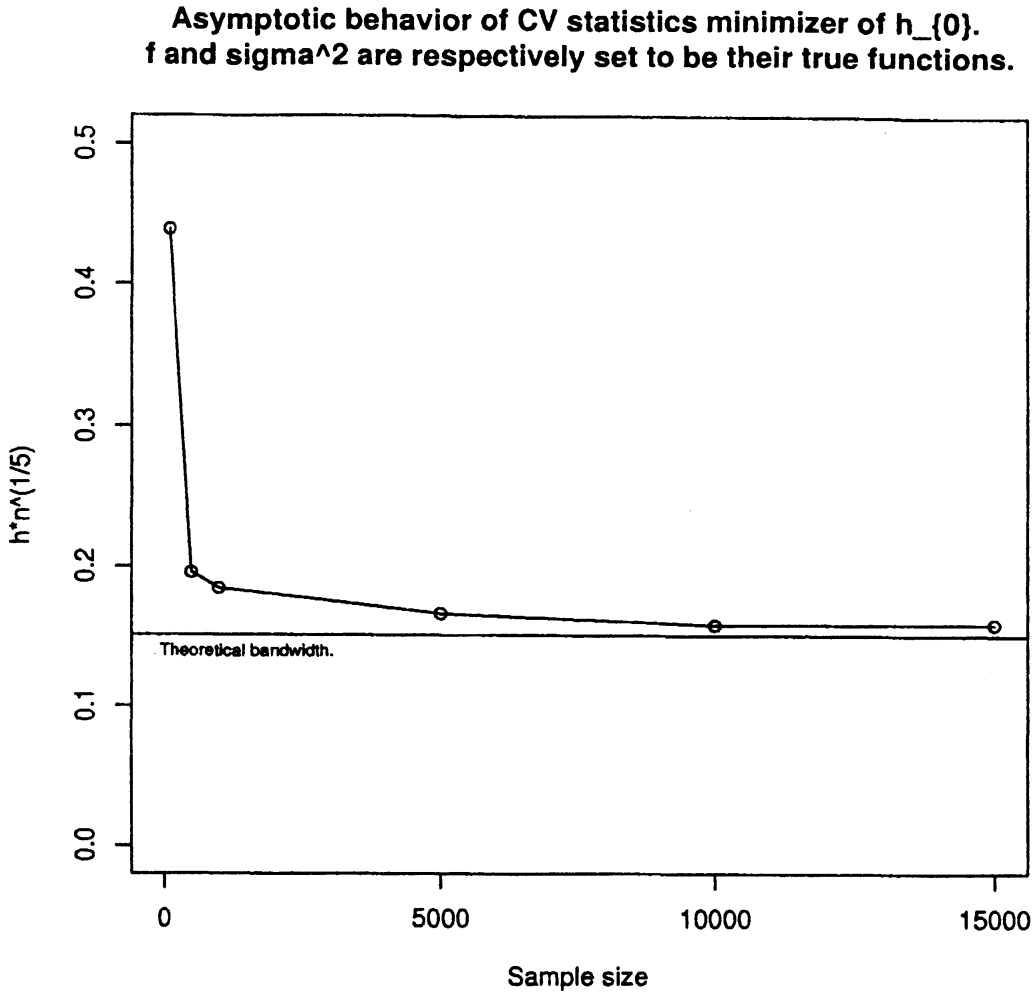


Figure 4: The upper panel plots the NW estimator with VS bandwidth of the example.



n	$\hat{h}_0 n^{1/5}$	$V[\hat{h}_0 n^{1/5}]$
100	0.4396	0.1808
500	0.1956	0.0024
1000	0.1845	0.0015
5000	0.1666	0.0011
10000	0.1588	0.0017
15000	0.1593	0.0018
$h_0^{MISE} = 0.1512$		

Figure 5: Asymptotic behavior of cross-validation statistics based on MISE with VS bandwidth.

Appendix

In Müller and Stadtmüller (1987), the following assumptions in addition to (4.4) (4.5) and (4.6) are employed.

$$E_{U|X}\left[U_i^4|X_i=x\right]=\mu_4(x)<\infty,$$

$$\max_{1\le i\le n-1}(x_{(i+1)}-x_{(i)})=O\left(\frac{1}{n}\right).$$

On $\sigma^2(x)$, $f_X(x)$, $m(x)$ and $\mu(x)$, the following conditions are placed,

$$\begin{aligned} f &\in \mathrm{Lip}_1([a,b]),\\ m(x) &\in \mathrm{Lip}_\alpha([a,b]),\ 0<\alpha<1,\\ \mu(x) &\in \mathrm{Lip}_\beta([a,b]),\ 0<\beta<1,\\ \sigma^2(x) &\in \mathrm{Lip}_\gamma([a,b]),\ 0<\gamma<1,\\ f &\text{ is bounded away from } 0, \end{aligned}$$

where a and b are real numbers. Let $U_{(i)}$ denote conditional random variable $U|X=x_{(i)}$. Then, expectation of difference sequences estimator is,

$$\begin{aligned} &E_{\mathbf{Y}}\left[\widehat{\sigma^2}_{DF}(x_{(i)})\right]\\ =&E_{\mathbf{Y}}\left[\left(\sum_{j=-r}^{r}d_j(U_{(i+j)}+m(x_{(i+j)}))\right)^2\right]\\ =&E_{\mathbf{Y}}\left[\left(\sum_{j=-r}^{r}d_jU_{(i+j)}+m(x_{(i)})\sum_{j=-r}^{r}d_j+O\left(\frac{1}{n^\alpha}\right)\right)^2\right]\\ =&E_{\mathbf{Y}}\left[\left(\sum_{j=-r}^{r}d_jU_{(i+j)}+O\left(\frac{1}{n^\alpha}\right)\right)^2\right] \end{aligned}$$

$$
\begin{aligned}
&= E_{\mathbf{Y}}\left[\sum_{j=-r}^{r} d_j^2 U_{(i+j)}^2 + \sum_{j=-r}^{r}\sum_{k=-r;k\neq j}^{r} d_j d_k U_{(i+j)} U_{(i+k)}\right] + O\left(\frac{1}{n^{2\alpha}}\right) \\
&= \sum_{j=-r}^{r} d_j^2 E_{\mathbf{Y}}\left[U_{(i+j)}^2\right] + O\left(\frac{1}{n^{2\alpha}}\right) \\
&= \sum_{j=-r}^{r} d_j^2 \sigma^2(x_{(i+j)}) + O\left(\frac{1}{n^{2\alpha}}\right) \\
&= \sigma^2(x_{(i)}) \sum_{j=-r}^{r} d_j^2 + O\left(\frac{1}{n^{\gamma}}\right) + O\left(\frac{1}{n^{2\alpha}}\right).
\end{aligned}
$$

Notice that the difference sequences estimator is asymptotically unbiased estimator of $\sigma^2(x_i)$ if (4.5) is assumed.

References

[1] E.A.Nadaraya (1964). On estimating regression. *Theory of probability and its applications*, 9, p141-142.

[2] E.A.Nadaraya (1965). On Non-parametric estimates of density functions and regression curves. *Theory of probability and its applications*, 10, p186-190.

[3] E.A.Nadaraya (1970). (Translated by B.Seckler from Russian) Remarks on non-parametric estimates for density functions and regression curves. *Theory of probability and its applications*, 15, p134-137.

[4] G.S. Watson (1963). Smooth regression analysis. *Sankhya SeriesA*, 26, p359-372.

[5] G.S. Watson and M.R. Leadbetter (1964). On the estimation of probability density I. *Annals of Mathematical statistics*, 34, p480-491.

[6] Mack, Y.P. (1981). Local properties of k-NN regression estimates. *SIAM journal of algebraic and discrete methods*, 2, p311-323.

[7] A.Pagan and A.Ullah (1999). Nonparametric Econometrics. *Cambridge University Press.*

[8] C.-K.Chu and J.S.Marron (1991). Choosing a Kernel Regression Estimator. *Statistical science* Vol.6, No.4, p404-436.

[9] Marron and Härdle (1986). Random approximations to some measures of accuracy in nonparametric curve estimation. *Journal of multivariate Analysis*, 20, p91-113.

[10] Härdle, Müller, Sperlich and Werwatz (2004). Nonparametric and Semiparametric Models. *Springer Series in Statistics.*

[11] Rudemo, M. (1982). Empirical choice of histgrams and kernel density estimators. *Scandinavian Journal of Statistics*, 9, p65-78.

[12] Bowman, A, W. (1984). An alternative method of cross-validation for the smoothing of density estimates. *Biometroka*, 71, p353-360.

[13] Hall, P. (1983). Large sample optimality of least square cross-validation in density estimation. *Annals of statistics*, 11, p1156-1174.

[14] J.Fan and Q.Yao (1998). Efficient estimation of conditional variance functions in stochastic regression. *Biometrika*, 85, 3, p645-660.

[15] H.G.Müller and U.Stadtmüller (1987). Estimation of heteroscedasticity in regression analysis. *The annals of statistics* Vol.15, No.2, p610-625.

[16] L.D.Brown and M.Levine (2007). Variance estimation in nonparametric regression via the difference sequence method. *The annals of statistics*, Vol.35, No.5, p2219-2232.