Optimal variance estimation without estimating the mean function

Tiejun Tong[1,*], Yanyuan Ma[2,**] and Yuedong Wang[3,***]

[1]Department of Mathematics, Hong Kong Baptist University, Hong Kong
[2]Department of Statistics, Texas A&M University, College Station, Texas, USA
[3]Department of Statistics and Applied Probability, University of California, Santa Barbara, California, USA

* Corresponding author. Email: tong@hkbu.edu.hk
** Email: ma@stat.tamu.edu
*** Email: yuedong@pstat.ucsb.edu

February 1, 2012

Abstract

We study the least squares estimator in the residual variance estimation context. We show that the mean squared differences of paired observations are asymptotically normally distributed. We further establish that, by regressing the mean squared differences of these paired observations on the squared distances between paired covariates via a simple least squares procedure, the resulting variance estimator is not only asymptotically normal and root-n consistent, but also reaches the optimal bound in terms of estimation variance. We also demonstrate the advantage of the least squares estimator in comparison with existing methods in terms of the second order asymptotic properties.

Key Words: Asymptotic normality; Difference-based estimator; Generalized least squares; Nonparametric regression; Optimal bound; Residual variance.

1 Introduction

Consider the following nonparametric regression model

$$y_i = g(x_i) + \varepsilon_i, \quad 0 \le x_i \le 1, \ i = 1, \ldots, n, \tag{1}$$

where y_i is the observation of the mean function g evaluated at design point x_i plus random error ε_i. We assume that ε_i's are independent and identically distributed with mean zero and variance σ^2. Many nonparametric regression methods have been developed to estimate the mean function g in the literature. Often, for choosing the amount of smoothing, testing goodness of fit or estimating model complexity, one needs an estimate of σ^2 that does not require estimating the mean function g first (Eubank & Spiegelman 1990, Gasser, Kneip & Kohler 1991, Ye 1998). For example, an estimate of σ^2 is required in the unbiased risk criterion for selecting the smoothing parameter in spline smoothing (see Section 3.3 in Wang (2011)).

One popular class of estimators of σ^2 which bypasses the estimation of g is the so-called difference-based estimators. The basic idea of difference-based estimation is to use differences to remove trend in the mean function. Assume that $0 \le x_1 \le \cdots \le x_n \le 1$. Rice (1984) proposed the first-order difference-based estimator

$$\hat{\sigma}_R^2 = \frac{1}{2(n-1)} \sum_{i=2}^{n} (y_i - y_{i-1})^2. \tag{2}$$

Gasser, Sroka & Jennen-Steinmetz (1986) and Hall, Kay & Titterington (1990) extended the Rice estimator to the second- and higher-order difference-based estimators, respectively. More difference-based estimators can be found in Dette, Munk & Wagner (1998) and Müller, Schick & Wefelmeyer (2003).

Tong & Wang (2005) proposed a variation of the difference-based estimator. For simplicity, consider the equally-spaced design where $x_i = i/n$. Define the lag-k Rice

estimators as

$$s_k = \frac{1}{2(n-k)} \sum_{i=1}^{n-k} (y_{i+k} - y_i)^2, \quad k = 1, 2, \ldots. \tag{3}$$

For any $k = o(n)$, it can be shown that $\mathrm{E}(s_k) = \sigma^2 + Jd_k + o(d_k)$ where $J = \int_0^1 \{g'(x)\}^2 dx/2$ and $d_k = k^2/n^2$. That is, the lag-k Rice estimator overestimates σ^2 by Jd_k. To reduce bias, they proposed fitting a linear regression model

$$s_k = \beta_0 + \beta_1 d_k + \epsilon_k, \quad k = 1, \ldots, m, \tag{4}$$

where $m = o(n)$ and using the least squares type of estimate of the intercept as an estimate of σ^2.

For ease of notation, let $\boldsymbol{s} = (s_1, \ldots, s_m)^T$, $\boldsymbol{\beta} = (\beta_0, \beta_1)^T$, $\boldsymbol{\epsilon} = (\epsilon_1, \ldots, \epsilon_m)^T$, $\mathbf{1} = (1, \ldots, 1)^T$, $\boldsymbol{d} = (d_1, \ldots, d_m)^T$, and $X = (\mathbf{1}, \boldsymbol{d})$ be the design matrix. Then (4) leads to $\boldsymbol{s} = X\boldsymbol{\beta} + \boldsymbol{\epsilon}$. Note that s_k is the average of $(n-k)$ lag-k differences and there are a total of $N = (n-1) + (n-2) + \cdots + (n-m) = nm - m(m+1)/2$ pairs of differences involved in the regression. Tong & Wang (2005) assigned weight $w_k = (n-k)/N$ to the observation s_k and then fitted the linear regression using the weighted least squares with weight matrix $W = \mathrm{diag}(w_1, \ldots, w_m)$. This results in $\hat{\boldsymbol{\beta}}_{\mathrm{WLS}} = (X^T W^{-1} X)^{-1} X^T W^{-1} \boldsymbol{s}$. Consequently, the weighted least squares estimator of σ^2 is

$$\hat{\sigma}^2 = \hat{\beta}_{0,\mathrm{WLS}} = \sum_{k=1}^{m} w_k s_k - \hat{\beta}_{1,\mathrm{WLS}} \bar{d}_w, \tag{5}$$

where $\bar{d}_w = \sum_{k=1}^m w_k d_k$ and $\hat{\beta}_{1,\mathrm{WLS}} = \sum_{k=1}^m w_k s_k (d_k - \bar{d}_w) / \sum_{k=1}^m w_k (d_k - \bar{d}_w)^2$. For simplicity, the above weighted least squares estimator $\hat{\sigma}^2$ is referred to as the least squares estimator in this paper. In Section 3 we will show that the above weighted least squares estimator is asymptotically equivalent to the ordinary least squares estimator and the generalized least squares estimator which takes into account the correlations between s_k's.

In this paper, we investigate the asymptotic distribution and efficiency of the least squares estimator. We show that the least squares estimator is asymptotically normally distributed in Section 2. We further show that the least squares estimator is asymptotically equivalent to the generalized least squares estimator where correlations among s_k are accounted for in Section 3. In Section 4, we derive the optimal efficiency bound for any estimation procedure and show that the least squares estimator reaches this optimal efficiency bound. In Section 5, we derived the mean squared error (MSE) for Müller & Stadtmüller (1999)'s estimator and then compare it to the least squares estimator. A real example is also provided. Finally, we conclude the paper in Section 6 with some simulation studies.

2 Least Squares Estimator

Let $\boldsymbol{y} = (y_1, \ldots, y_n)^T$, $\boldsymbol{g} = (g(x_1), \ldots, g(x_n))^T$, and $\boldsymbol{\varepsilon} = (\varepsilon_1, \ldots, \varepsilon_n)^T$. Then $\boldsymbol{y} = \boldsymbol{g} + \boldsymbol{\varepsilon}$. Let $\gamma_i = \mathrm{E}(\varepsilon^i)/\sigma^i$ for $i = 3, 4$, and $\xrightarrow{\mathcal{D}}$ denote convergence in distribution. Assume that $\gamma_4 > 1$. We first establish asymptotic normality for the Rice estimator.

Theorem 1. *Assume that g has a bounded second derivative. For any $k = n^r$ with $0 < r < 3/4$, the lag-k Rice estimator satisfies $\sqrt{n}(s_k - \sigma^2) \xrightarrow{\mathcal{D}} N(0, \gamma_4\sigma^4)$ as $n \to \infty$.*

Proof of Theorem 1 can be found in Appendix 1. Next we establish asymptotic normality for the least squares estimator (5). Following the result in Tong & Wang (2005), the least squares estimator (5) has a quadratic form $\hat{\sigma}^2 = \boldsymbol{y}^T D\boldsymbol{y}/\mathrm{tr}(D)$, where $D = (d_{ij})_{n\times n}$ is a symmetric matrix with elements

$$d_{ij} = \begin{cases} \sum_{k=1}^{m} b_k + \sum_{k=0}^{\min(i-1,n-i,m)} b_k & 1 \le i = j \le n, \\ -b_{|i-j|} & 0 < |i-j| \le m, \\ 0 & \text{otherwise}, \end{cases}$$

where $b_0 = b_{m+1} = 0$ and $b_k = 1 - \bar{d}_w(d_k - \bar{d}_w)/\sum_{k=1}^{m} w_k(d_k - \bar{d}_w)^2$ for $k = 1, \ldots, m$.

Theorem 2. *Assume that g has a bounded second derivative and $E(\varepsilon^6)$ is finite. Then for any $m = n^r$ with $0 < r < 1/2$, the least squares estimator $\hat{\sigma}^2$ satisfies $\sqrt{n}(\hat{\sigma}^2 - \sigma^2) \xrightarrow{\mathcal{D}} N\{0, (\gamma_4 - 1)\sigma^4\}$ as $n \to \infty$.*

Proof of Theorem 2 can be found in Appendix 2. Given that $\mathrm{E}(\varepsilon^6)$ is finite, Theorems 1 and 2 show that the least squares estimator is more efficient than the Rice estimator. Theorem 2 also indicates that the least squares estimator is as efficient as the sample variance based on independent and identically distributed samples, regardless of whether the unknown mean function is a constant or not.

Theorem 2 can be used to construct confidence intervals for σ^2. Assume that $n > (\gamma_4 - 1)z^2_{\alpha/2}$ where z_α is the upper α-th percentile of the standard normal distribution. Then an approximate $1-\alpha$ confidence interval for σ^2 is $[\hat{\sigma}^2/\{1+z_{\alpha/2}\sqrt{(\gamma_4 - 1)/n}\}, \hat{\sigma}^2/\{1-z_{\alpha/2}\sqrt{(\gamma_4 - 1)/n}\}]$. For the special case when the ε_i's are distributed from $\mathrm{N}(0, \sigma^2)$, we have $\gamma_4 = 3$. In general, the parameter γ_4 can be replaced by an estimate. Finally, by Box (1954) and Rotar (1973), the finite sample distribution of $\hat{\sigma}^2$ can be approximated by the scaled chi-squared distribution, $(\sigma^2/\nu)\chi^2(\nu)$, where $\nu = \{\mathrm{tr}(D)\}^2/\mathrm{tr}(D^2)$.

3 Generalized Least Squares Estimator

In Appendix 3 we show that, for any $1 \le b < k = n^r$ with $0 < r < 2/3$, $\mathrm{Cov}(s_b, s_k) = n^{-1}(\gamma_4 - 1)\sigma^4 + o(n^{-1})$. Combined with the results in Theorem 1, we have $\mathrm{Corr}(s_b, s_k) \to (\gamma_4 - 1)/\gamma_4$ as $n \to \infty$. In the case when the ε_i's are normally distributed, $\gamma_4 = 3$ and the correlation coefficients between the lag-k Rice estimators are all asymptotically equal to $2/3$.

In the construction of the least squares estimator in Section 2 we have ignored the correlation between s_k's. Given that the correlation between lag-k Rice estimators are high, a natural question is whether the least squares estimator can be improved by the

following generalized least squares estimator

$$\hat{\boldsymbol{\beta}}_{\text{GLS}} = (X^T\Sigma^{-1}X)^{-1}X^T\Sigma^{-1}\boldsymbol{s}, \tag{6}$$

where $\Sigma = \gamma_4\sigma^4\{(1-\rho)I + \rho\mathbf{1}^T\mathbf{1}\}/n$ is the asymptotic variance-covariance matrix, $\rho = (\gamma_4 - 1)/\gamma_4$, and I is the identity matrix. It is known that $\hat{\boldsymbol{\beta}}_{\text{GLS}}$ is the best linear unbiased estimator of $\boldsymbol{\beta}$ (Kariya & Kurata 2004). Since Σ has the compound symmetry structure and the first column of X is $\mathbf{1}$, by McElroy (1967), the generalized least squares estimator $\hat{\boldsymbol{\beta}}_{\text{GLS}}$ is identical to the ordinary least squares estimator $\hat{\boldsymbol{\beta}}_{\text{OLS}} = (X^TX)^{-1}X^T\boldsymbol{s}$. Furthermore, for any $m = o(n)$, it is not difficult to show that $\hat{\boldsymbol{\beta}}_{\text{WLS}}$ is equivalent to $\hat{\boldsymbol{\beta}}_{\text{OLS}}$. Therefore, $\hat{\boldsymbol{\beta}}_{\text{OLS}}$, $\hat{\boldsymbol{\beta}}_{\text{GLS}}$ and $\hat{\boldsymbol{\beta}}_{\text{WLS}}$ are all asymptotically equivalent.

4 The Optimal Efficiency Bound for Estimating σ^2

In this section, we derive the optimal semiparametric efficiency bound for estimating σ^2 in model (1) for any estimation procedure and show that the least squares estimator reaches this bound.

Consider the estimation of σ^2 in model (1) regardless of how the estimation is carried out. For simplicity, we omit the subindex i. Under (1), the only assumption is that $\varepsilon = Y - g(X)$ are independent and identically distributed with mean zero, and are independent of X. Denote the model of the probability density function of ε as $\eta(\varepsilon)$.

The probability density function model of (x, y) can be written as $f_X(x)\eta\{y - g(x)\} = f_X(x)\eta(\varepsilon)$, where $f_X(\cdot)$ is a marginal probability density function model of X and η is a probability density function model that ensures zero mean, i.e. $\int \eta(\varepsilon)d\varepsilon = 1$ and $\int \varepsilon\eta(\varepsilon)d\varepsilon = 0$. Viewing f_X, η and g as the nuisance parameters and $\sigma^2 = \mathrm{E}(\varepsilon^2)$ as the parameter of interest, this becomes a semiparametric problem and one can derive the efficient influence function through projecting any influence function onto

the tangent space associated with f_X, η and g.

Simple calculation yields the tangent space of model (1) to be

$$\begin{aligned}\Lambda_{\mathcal{T}} &= \{h(x) + f(\varepsilon) + \eta_0'(\varepsilon)/\eta_0(\varepsilon)a(x) : \\ &\quad \forall h, f \text{ such that } \mathrm{E}(h) = 0, \mathrm{E}(f) = \mathrm{E}(\varepsilon f) = 0, \text{ and } \forall a\}, \end{aligned} \tag{7}$$

where $\eta_0(\cdot)$ denotes the true probability density function of ε. Following the procedure in Chapter 4 of Tsiatis (2006), we consider an arbitrary parametric submodel, denoted as $\eta(\varepsilon, \boldsymbol{\mu})$. Here $\boldsymbol{\mu}$ is a finite dimensional vector of parameters and there exists $\boldsymbol{\mu}_0$, such that $\eta(\varepsilon, \boldsymbol{\mu}_0) = \eta_0(\varepsilon)$. In addition, $\eta(\varepsilon, \boldsymbol{\mu})$ is a valid probability density function and $\int \varepsilon\eta(\varepsilon; \boldsymbol{\mu})d\varepsilon = 0$ for all $\boldsymbol{\mu}$ in a local neighborhood of $\boldsymbol{\mu}_0$. We have $\partial \int \varepsilon^2\eta(\varepsilon, \boldsymbol{\mu})d\varepsilon/\partial\boldsymbol{\mu} = \mathrm{E}(\varepsilon^2 S_{\boldsymbol{\mu}})$, where $S_{\boldsymbol{\mu}} = \partial \log \eta(\varepsilon, \boldsymbol{\mu})/\partial\boldsymbol{\mu}$ is the score vector with respect to $\boldsymbol{\mu}$. Hence $\varepsilon^2 - \sigma^2$ is a valid influence function. We decompose $\varepsilon^2 - \sigma^2$ into

$$\varepsilon^2 - \sigma^2 = \{\varepsilon^2 - \sigma^2 + \gamma_3\sigma^3\eta_0'(\varepsilon)/\eta_0(\varepsilon)\} - \gamma_3\sigma^3\eta_0'(\varepsilon)/\eta_0(\varepsilon).$$

It is not difficult to verify that $\varepsilon^2 - \sigma^2 + \gamma_3\sigma^3\eta_0'(\varepsilon)/\eta_0(\varepsilon)$ satisfies the requirement on f in (7). Hence, it is a qualified $f(\varepsilon)$ function. Letting $a(x)$ in (7) be $-\gamma_3\sigma^3$ yields $-\gamma_3\sigma^3\eta_0'(\varepsilon)/\eta_0(\varepsilon)$. Thus, $\varepsilon^2 - \sigma^2 \in \Lambda_{\mathcal{T}}$, and consequently it is the efficient influence function. The corresponding efficient estimation variance is $n^{-1}E\{(\varepsilon^2 - \sigma^2)^2\} = n^{-1}(\gamma_4 - 1)\sigma^4$, which agrees with the result in Theorem 2. This shows that the least squares estimator is indeed optimal in terms of its estimation variability among the class of all root-n consistent estimators.

In the above derivation, we have not taken into account that X_i's are actually equally spaced instead of being random. However, assuming $f_X(x)$ to be uniform or more generally assuming $f_X(x)$ to have any particular form does not change the efficiency result. This is because the calculation relies on the property of ε only, which is independent of X.

5 Variance Estimator of Müller and Stadtmüller

Müller & Stadtmüller (1999) proposed a similar least squares type estimator for the equally-spaced design where $x_i = i/n$. Define

$$z_k = \frac{1}{2(n-L)} \sum_{i=1}^{n-L} (y_{i+k} - y_i)^2, \quad 1 \le k \le L,$$

where $L = L(n) \ge 1$. In the context of testing if the mean function contains jump discontinuities, Müller & Stadtmüller (1999) fitted a linear model that regresses z_k on two independent variables, one for the sum of the squared jump sizes and the other for the integrated squared first derivative, and then estimate the residual variance as the intercept. In the case when the function is smooth, that is, when the sum of the squared jump sizes equals to zero, the variance estimator in Müller & Stadtmüller (1999) reduces to

$$\hat{\sigma}^2_{\text{MS}} = \frac{3}{L(L-1)(L-2)} \sum_{k=1}^{L} \{3L^2 + 3L + 2 - 6(2L+1)k + 10k^2\} z_k. \tag{8}$$

The dependent variable z_k in Müller & Stadtmüller (1999) uses the first $n-L$ terms in the lag-k Rice estimator s_k while the last $L-k$ terms are ignored. This makes z_k a less efficient estimator of σ^2, especially when $L-k$ is large. In addition, noting that $\hat{\sigma}^2_{\text{MS}}$ is a weighted average of z_k with larger weights assigned to small k and more terms are ignored with small k, the efficiency loss of $\hat{\sigma}^2_{\text{MS}}$ over $\hat{\sigma}^2$ can be severe for small sample sizes.

Let $a_0 = 0$ and $a_k = 3\{3L^2 + 3L + 2 - 6(2L+1)k + 10k^2\}/\{L(L-1)(L-2)\}$ for $k = 1, \ldots, L$. By Lemma A5 in Müller & Stadtmüller (1999), we have $\sum_{k=1}^{L} a_k = 1$. Then $\hat{\sigma}^2_{\text{MS}}$ can be represented as the quadratic form, $\hat{\sigma}^2_{\text{MS}} = \boldsymbol{y}^T M \boldsymbol{y}$, where $M = (m_{ij})_{n \times n}$

is a symmetric matrix with elements

$$m_{ij} = \begin{cases} 1 + \sum_{k=0}^{i-1} a_k & i = j = 1, \ldots, L, \\ 2 & i = j = L+1, \ldots, n-L, \\ \sum_{k=i}^{n} a_{k+L-n} & i = j = n-L+1, \ldots, n, \\ -a_{j-i} & 0 < j - i \leq L \text{ and } i \leq n-L, \\ -a_{i-j} & 0 < i - j \leq L \text{ and } j \leq n-L, \\ 0 & \text{otherwise.} \end{cases}$$

Let $\text{diag}(M)$ denote the diagonal matrix of M. By Dette et al. (1998) we have

$$\begin{aligned} \text{MSE}(\hat{\sigma}^2_{\text{MS}}) &= \left[(\boldsymbol{g}^T M \boldsymbol{g})^2 + 4\sigma^2 \boldsymbol{g}^T M^2 \boldsymbol{g} + 4\boldsymbol{g}^T M \text{diag}(M) \mathbf{1} \sigma^3 \gamma_3 \right. \\ &\quad \left. + \sigma^4 \text{tr}[\{\text{diag}(M)\}^2](\gamma_4 - 3) + 2\sigma^4 \text{tr}(M^2)\right] / \text{tr}(M)^2, \end{aligned} \tag{9}$$

where the first term in (9) is the squared bias and the last four terms make up the variance.

Theorem 3. *Assume that g has a bounded second derivative. Then for the equally spaced design with $n \to \infty$, $L \to \infty$ and $L/n \to 0$, we have the following bias, variance, and the mean squared error for the estimator (8),*

$$\begin{aligned} \text{Bias}(\hat{\sigma}^2_{MS}) &= o\left(\frac{L^2}{n^2}\right), \\ \text{var}(\hat{\sigma}^2_{MS}) &= \frac{1}{n}\text{var}(\varepsilon^2) + \frac{73L}{70n^2}\text{var}(\varepsilon^2) + \frac{9}{Ln}\sigma^4 + o\left(\frac{L}{n^2}\right) + o\left(\frac{1}{Ln}\right), \\ \text{MSE}(\hat{\sigma}^2_{MS}) &= \frac{1}{n}\text{var}(\varepsilon^2) + \frac{73L}{70n^2}\text{var}(\varepsilon^2) + \frac{9}{Ln}\sigma^4 + o\left(\frac{L}{n^2}\right) + o\left(\frac{1}{Ln}\right) + o\left(\frac{L^4}{n^4}\right). \end{aligned} \tag{10}$$

Proof of Theorem 3 can be found in Appendix 4. The asymptotical optimal bandwidth is $L_{opt} = \sqrt{630n\sigma^4/73\text{var}(\varepsilon^2)}$. Substituting L_{opt} into (10) leads to

$$\text{MSE}(\hat{\sigma}^2_{\text{MS}}(L_{opt})) = \frac{1}{n}\text{var}(\epsilon^2) + \frac{\sqrt{45990}}{35}\left\{\sigma^4 \text{var}(\varepsilon^2)\right\}^{1/2} n^{-3/2} + o\left(n^{-3/2}\right). \tag{11}$$

The optimal MSE of $\hat{\sigma}^2$ is (Tong & Wang 2005)

$$\text{MSE}(\hat{\sigma}^2(m_{opt})) = \frac{1}{n}\text{var}(\epsilon^2) + \frac{\sqrt{567}}{28}\left\{\sigma^4 \text{var}(\varepsilon^2)\right\}^{1/2} n^{-3/2} + o\left(n^{-3/2}\right).$$

It is clear that both $\hat{\sigma}^2$ and $\hat{\sigma}^2_{\rm MS}$ reach the optimal efficiency bound with the same first order term. However, the coefficient of the higher order term for $\hat{\sigma}^2_{\rm MS}$ is about seven times of that for $\hat{\sigma}^2$. Since the higher order term is not negligible for small to moderate sample sizes, $\hat{\sigma}^2$ often provides a much smaller MSE than $\hat{\sigma}^2_{\rm MS}$ in such situations. See simulation results in Section 6.

Even though the two estimators $\hat{\sigma}^2$ and $\hat{\sigma}^2_{\rm MS}$ look similar for one-dimensional equally spaced case, there is a fundamental difference behind the motivations for these estimators: the regression estimator in Tong & Wang (2005) was developed to estimate variances in nonparametric regression on general domains while the regression estimator in Müller & Stadtmüller (1999) was developed for assessing whether a one-dimensional mean function is smooth. Specifically, consider model (1) with $x_i \in \mathcal{T}$ where $\mathcal{T}$ is an arbitary subset in a normed space. Let $d_{ij} = \|x_i - x_j\|^2$ and $s_{ij} = \frac{1}{2}(y_i - y_j)^2$ for all pairs i and j, where $1 \le i < j \le n$. We fit the following simple linear model

$$s_{ij} = \beta_0 + \beta_1 d_{ij} + \epsilon_{ij}, \qquad d_{ij} \le m, \tag{12}$$

using the least squares where $m > 0$ is the bandwidth. The estimate of σ^2 is $\hat{\sigma}^2 = \hat{\beta}_0$. The variance estimator in Müller & Stadtmüller (1999) requires an ordering of the design points which may not be available for a general domain.

For the purpose of illustration, consider the Lake Acidity Data which contains measurements of 112 lakes in the southern Blue Ridge mountains area (Gu & Wahba 1993). Of interest is the dependence of the water pH level (ph) on the calcium concentration in $\log_{10}$ milligrams per liter (t_1) and the geographical location ($\mathbf{t}_2 = (t_{21}, t_{22})$ with t_{21}=latitude and t_{22}=longitude). For illustration, we consider the nonparametric regression model (1) with three different cases of x: $x = t_1$, $x = \mathbf{t}_2$ and $x = (t_1, \mathbf{t}_2)$. These three cases correspond to three different domains of one, two and three dimensions, respectively. For the first two cases, we use simple Euclidean norms. For the third case,

we rescale t_1 and $\|\mathbf{t}_2\|$ to the same scale before estimating the variance. Estimates of σ^2 for the above three cases with $m = n^{1/2}$ are 0.0821, 0.0884 and 0.0544, respectively using our method. The method in Müller & Stadtmüller (1999) does not apply to any one of these three cases.

6 Simulation Studies

In this section, we conduct simulations to compare the performance of the estimators $\hat{\sigma}^2$ and $\hat{\sigma}^2_{\text{MS}}$. The design points are $x_i = i/n$ and ε_i are independent and identically distributed from $N(0, \sigma^2)$. We consider three mean functions, $g_1(x) = 5x$, $g_2(x) = 5x(1-x)$, and $g_3(x) = 5\sin(2\pi x)$. Note that the first two functions were used in Müller & Stadtmüller (1999) and the last one was used in Tong & Wang (2005). We set coefficients of all three functions to be 5. For each mean function, we consider $n = 30$, 100 and 1000, corresponding to small, moderate and large sample sizes respectively, and $\sigma^2 = 0.25$ and 4, corresponding to small and large variances respectively. In total, we have 18 combinations of simulation settings.

For each simulation setting, we generate observations and compute the estimators $\hat{\sigma}^2(m)$ and $\hat{\sigma}^2_{\text{MS}}(L)$. For the bandwidth m, we choose $m_s = n^{1/2}$ and $m_t = n^{1/3}$ as suggested in Tong & Wang (2005). For the bandwidth L, Müller & Stadtmüller (1999) observed that the estimator $\hat{\sigma}^2_{\text{MS}}$ is quite stable and does not vary much with L. Therefore, we also choose $L_s = n^{1/2}$ and $L_t = n^{1/3}$ for ease of comparison. The cross-validation method may also be used to select the bandwidth m in $\hat{\sigma}^2(m)$ (Tong & Wang 2005). Nevertheless, we did not include this option in our simulations since the cross-validation method is not readily available for the estimator $\hat{\sigma}^2_{\text{MS}}$.

We repeat the simulation 1000 times and compute the relative mean squared errors $n\text{MSE}/(2\sigma^4)$. Table 1 lists relative mean squared errors for all simulation settings. Note that neither D nor M is guaranteed to be positive definite. Therefore, $\hat{\sigma}^2$ and

$\hat{\sigma}^2_{\text{MS}}$ may take negative values. Simulations indicate that a negative estimate occurs very rarely for $\hat{\sigma}^2$ (Tong & Wang 2005), while $\hat{\sigma}^2_{\text{MS}}$ tends to be negative when L is large (Müller & Stadtmüller 1999). We replace negative estimates by zero in the calculation of the relative mean squared errors.

n	σ^2	g	$\hat{\sigma}^2(m_s)$	$\hat{\sigma}^2(m_t)$	$\hat{\sigma}^2_{\text{MS}}(L_s)$	$\hat{\sigma}^2_{\text{MS}}(L_t)$
30	0.25	g_1	1.33	1.58	3.97	10.80
		g_2	1.34	1.57	3.97	10.79
		g_3	8.64	2.19	6.91	11.60
	4	g_1	1.32	1.57	3.91	10.75
		g_2	1.32	1.57	3.91	10.75
		g_3	1.38	1.59	4.02	10.83
100	0.25	g_1	1.25	1.43	2.09	5.53
		g_2	1.25	1.43	2.08	5.55
		g_3	2.06	1.45	2.30	5.50
	4	g_1	1.25	1.43	2.09	5.54
		g_2	1.25	1.43	2.08	5.54
		g_3	1.27	1.43	2.09	5.52
1000	0.25	g_1	1.18	1.30	1.35	1.83
		g_2	1.18	1.30	1.35	1.83
		g_3	1.19	1.30	1.35	1.83
	4	g_1	1.18	1.30	1.35	1.83
		g_2	1.18	1.30	1.35	1.83
		g_3	1.18	1.30	1.35	1.83

Table 1: Relative mean squared errors for the two estimators with bandwidths $m_s = L_s = n^{1/2}$ and $m_t = L_t = n^{1/3}$, respectively.

We observe that $\hat{\sigma}^2$ has smaller relative mean squared errors than $\hat{\sigma}^2_{\text{MS}}$ for all settings except for the case $(n, \sigma^2, g) = (30, 0.25, g_3)$. For this exceptional case, we plot in Figure 1 the histograms of the non-truncated estimates (including negative estimates) $\hat{\sigma}^2(m_s)$ and $\hat{\sigma}^2_{\text{MS}}(L_s)$. A relatively large portion of $\hat{\sigma}^2_{\text{MS}}(L_s)$ takes negative values. The choice of the bandwidth m_s is too large for $\hat{\sigma}^2$ when n is small (Tong & Wang 2005). Overall, the estimator $\hat{\sigma}^2$ performs better than $\hat{\sigma}^2_{\text{MS}}$, confirming the theoretical results in Section 5. Comparisons between $\hat{\sigma}^2(m_s)$ and $\hat{\sigma}^2(m_t)$ are similar to those in Tong & Wang (2005).

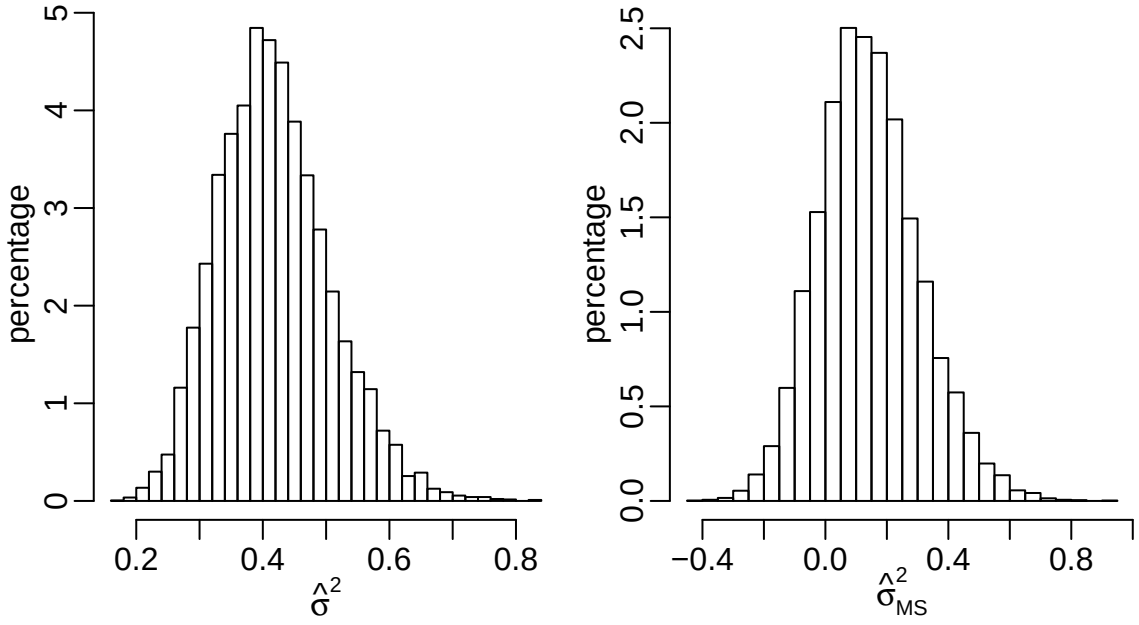

Figure 1: Histograms of the variance estimates $\hat{\sigma}^2(m_s)$ (left) and $\hat{\sigma}^2_{\text{MS}}(L_s)$ (right) for the case $(n, \sigma^2, g) = (30, 0.25, g_3)$.

Acknowledgement

Tiejun Tong's research was supported by Hong Kong RGC grant HKBU202711, and Hong Kong Baptist University grants FRG1/10-11/031 and FRG2/10-11/020. Yanyuan Ma's research was supported by NSF grant DMS0906341 and NINDS grant R01-NS073671. Yuedong Wang's research was supported by NSF grant DMS0706886. The authors thank the editor, the associate editor, and a referee for their constructive comments that substantially improved an earlier draft.

References

Box, G. E. P. (1954). Some theorems on quadratic forms applied in the study of analysis of variance problems, i. effect of inequality of variance in the one-way

classification, *Annals of Mathematical Statistics* **25**: 290–302.

Brockwell, P. J. and Davis, R. A. (1991). *Time Series: Theory and Methods*, Springer.

Dette, H., Munk, A. and Wagner, T. (1998). Estimating the variance in nonparametric regression - what is a reasonable choice?, *Journal of the Royal Statistical Society, Series B* **60**: 751–764.

Eubank, R. L. and Spiegelman, C. H. (1990). Testing the goodness of fit of a linear model via nonparametric regression techniques, *Journal of the American Statistical Association* **85**: 387–392.

Gasser, T., Kneip, A. and Kohler, W. (1991). A flexible and fast method for automatic smoothing, *Journal of the American Statistical Association* **86**: 643–52.

Gasser, T., Sroka, L. and Jennen-Steinmetz, C. (1986). Residual variance and residual pattern in nonlinear regression, *Biometrika* **73**: 625–633.

Gu, C. and Wahba, G. (1993). Semiparametric ANOVA with tensor product thin plate spline, *Journal of the Royal Statistical Society, Series B* **55**: 353–368.

Hall, P., Kay, J. W. and Titterington, D. M. (1990). Asymptotically optimal difference-based estimation of variance in nonparametric regression, *Biometrika* **77**: 521–528.

Kariya, T. and Kurata, H. (2004). *Generalized Least Squares*, Wiley.

McElroy, F. W. (1967). A necessary and sufficient condition that ordinary least-squares estimators be best linear unbiased, *Journal of the American Statistical Association* **62**: 1302–1304.

Müller, H. and Stadtmüller, U. (1999). Discontinuous versus smooth regression, *Annals of Statistics* **27**: 299–337.

Müller, U., Schick, A. and Wefelmeyer, W. (2003). Estimating the error variance in nonparametric regression by a covariate-matched U-statistic, *Statistics* **37**: 179–188.

Rice, J. A. (1984). Bandwidth choice for nonparametric regression, *Annals of Statistics* **12**: 1215–1230.

Rotar, V. I. (1973). Some limit theorems for polynomials of second degree, *Theory of Probability and its Applications* **18**: 527–534.

Tong, T. and Wang, Y. (2005). Estimating residual variance in nonparametric regression using least squares, *Biometrika* **92**: 821–830.

Tsiatis, A. A. (2006). *Semiparametric Theory and Missing Data*, Springer.

Wang, Y. (2011). *Smoothing Splines: Methods and Applications*, Chapman and Hall, New York.

Whittle, P. (1964). On the convergence to normality of quadratic forms in independent variables, *Theory of Probability and Its Applications* **9**: 103–108.

Ye, J. (1998). On measuring and correcting the effects of data mining and model selection, *Journal of the American Statistical Association* **93**: 120–131.

Appendix 1: Proof of Theorem 1

For ease of notation, let $g_i = g(x_i)$, $i = 1, \ldots, n$. Write s_k as a sum of three parts, $s_k = L_1 + L_2 + L_3$, where

$$L_1 = \frac{1}{2(n-k)} \sum_{i=k+1}^{n} (g_i - g_{i-k})^2,$$

$$L_2 = \frac{1}{n-k} \sum_{i=k+1}^{n} (g_i - g_{i-k})(\varepsilon_i - \varepsilon_{i-k}),$$

$$L_3 = \frac{1}{2(n-k)} \sum_{i=k+1}^{n} (\varepsilon_i - \varepsilon_{i-k})^2.$$

Applying the Taylor expansion, it can be shown that $L_1 = (k^2/n^2)J + o(k^2/n^2) = o_p(n^{-1/2})$ when $k = n^r$ with $0 < r < 3/4$. For L_2, we have

$$\mathrm{E}(L_2^2) = \frac{2\sigma^2}{(n-k)^2} \left\{ \sum_{i=k+1}^{n} (g_i - g_{i-k})^2 - \sum_{i=k+1}^{n-k} (g_i - g_{i-k})(g_{i+1} - g_i) \right\} = O\left(\frac{k^2}{n^3}\right).$$

This implies that $L_2 = o_p(n^{-1/2})$ for any $k = o(n)$.

Rewrite L_3 as $L_3 = \sigma^2 + \sum_{i=k+1}^{n} \xi_i(k)/(n-k)$, where $\xi_i(k) = (\varepsilon_i - \varepsilon_{i-k})^2/2 - \sigma^2$. For any given k, $\{\xi_i(k),\ i = k+1, \ldots, n\}$ is a strictly stationary sequence of random variables with mean zero and autocovariance function

$$\gamma(\tau) = \gamma(s, s+\tau) = \begin{cases} (\gamma_4 + 1)\sigma^4/2 & \tau = 0, \\ (\gamma_4 - 1)\sigma^4/4 & \tau = k, \\ 0 & \text{otherwise.} \end{cases}$$

Note also that the sequence $\{\xi_i(k),\ i = k+1, \ldots, n\}$ is m-dependent with $m = k$. Thus by the central limit theorem for strictly stationary m-dependent sequences (Brockwell and Davis, 1991), $\sqrt{n}(L_3 - \sigma^2) \xrightarrow{\mathcal{D}} N(0, \nu_k^2)$ as $n \to \infty$, where $\nu_k^2 = \gamma(0) + 2\sum_{\tau=1}^{k} \gamma(\tau) = \gamma_4\sigma^4$. Finally, noting that $s_k = L_1 + L_2 + L_3 = L_3 + o_p(n^{-1/2})$, we have $\sqrt{n}(s_k - \sigma^2) \xrightarrow{\mathcal{D}} N(0, \gamma_4\sigma^4)$ as $n \to \infty$.

Appendix 2: Proof of Theorem 2

We first state two lemmas. Lemma 1 is an immediate result from Whittle (1964). Lemma 2 was derived, in essence, in Tong and Wang (2005).

Lemma 1. *Assume that the matrix* $A = (a_{ij})_{n\times n}$ *satisfies* $a_{ij} = a_{i-j}$ *and* $\sum_{-\infty}^{\infty} a_k^2 < \infty$. *Furthermore, assume that* $E(\varepsilon^6)$ *is finite. Then*

$$\frac{1}{n}\boldsymbol{\varepsilon}^T A\boldsymbol{\varepsilon} = \frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{n} a_{i-j}\varepsilon_i\varepsilon_j \xrightarrow{\mathcal{D}} N(a_0\sigma^2, \sigma_A^2), \quad \text{as } n \to \infty,$$

where $\sigma_A^2 = (\gamma_4 - 3)a_0^2\sigma^4/n + 2\sigma^4\sum_{i=1}^{n}\sum_{j=1}^{n} a_{i-j}^2/n^2$.

Lemma 2. *Assume that* $m \to \infty$ *and* $m/n \to 0$. *Then*

(i) $\sum_{k=1}^{m} b_k = m - \frac{5m^2}{16n} + o(m)$;

(ii) $\sum_{k=j}^{m} b_k = m - \frac{9}{4}j + \frac{5j^3}{4m^2} + o(m),\ 1 \le j \le m$;

(iii) $\sum_{k=1}^{m} b_k^2 = \frac{9}{4}m + o(m)$;

(iv) $\boldsymbol{g}^T D\boldsymbol{g} = O(m^4/n^2)$;

(v) $\boldsymbol{g}^T D^2\boldsymbol{g} = O(m^5/n^2)$.

Proof of Theorem 2: Noting that $\boldsymbol{y} = \boldsymbol{g} + \boldsymbol{\varepsilon}$ and $\mathrm{tr}(D) = 2N$, we have

$$\hat{\sigma}^2 = \frac{1}{2N}\boldsymbol{g}^T D\boldsymbol{g} + \frac{1}{N}\boldsymbol{g}^T D\boldsymbol{\varepsilon} + \frac{1}{2N}\boldsymbol{\varepsilon}^T D\boldsymbol{\varepsilon}. \tag{13}$$

The first term in (13) corresponds to the bias term of the least squares estimator. By Lemma 2, we have $\boldsymbol{g}^T D\boldsymbol{g}/(2N) = O(m^3/n^3)$. Thus, for any $m = n^r$ with $0 < r < 5/6$,

$$\frac{1}{2N}\boldsymbol{g}^T D\boldsymbol{g} = o(n^{-1/2}). \tag{14}$$

For the second term in (13), by Lemma 2 we have $E\left(\boldsymbol{g}^T D\boldsymbol{\varepsilon}/N\right)^2 = \boldsymbol{g}^T D^2\boldsymbol{g}/N^2 = O(m^3/n^4)$. This implies that, for any $m = o(n)$,

$$\frac{1}{N}\boldsymbol{g}^T D\boldsymbol{\varepsilon} = o_p(n^{-1/2}). \tag{15}$$

Now we derive the limiting distribution of the third term in (13). Let $nD/(2N) = C - H$, where $C = (c_{ij})_{n\times n}$ with elements

$$c_{ij} = \begin{cases} n\sum_{k=1}^{m} b_k/N & 1 \le i = j \le n, \\ -nb_{|i-j|}/(2N) & 0 < |i-j| \le m, \\ 0 & \text{otherwise}, \end{cases}$$

and $H = \mathrm{diag}(h_1, h_2, \ldots, h_n)$ with elements $h_i = n\sum_{\min(i,n+1-i,m+1)}^{m+1} b_k/(2N)$. Then

$$\frac{1}{2N}\boldsymbol{\varepsilon}^T D\boldsymbol{\varepsilon} = \frac{1}{n}\boldsymbol{\varepsilon}^T C\boldsymbol{\varepsilon} - \frac{1}{n}\boldsymbol{\varepsilon}^T H\boldsymbol{\varepsilon}. \tag{16}$$

For the matrix C, let $c_{ij} = c_{i-j}$ with $c_0 = n\sum_{k=1}^{m} b_k/N$, $c_{i-j} = c_{j-i} = -nb_{|i-j|}/(2N)$ for $0 < |i-j| \le m$, and $c_{i-j} = c_{j-i} = 0$ for $|i-j| > m$. By Lemma 2, for any $m = o(n)$, $\sum_{-\infty}^{\infty} c_k^2 = c_0^2 + 2\sum_{k=1}^{m} c_k^2 = 1 + o(1) < \infty$. Then under the assumption that $\mathrm{E}(\varepsilon^6)$ is finite, by Lemma 1 we have

$$\sqrt{n}\left(\frac{1}{n}\boldsymbol{\varepsilon}^T C\boldsymbol{\varepsilon} - c_0\sigma^2\right) \xrightarrow{\mathcal{D}} N(0, \sigma_c^2), \quad \text{as } n \to \infty, \tag{17}$$

where

$$\sigma_c^2 = \frac{n^2(\gamma_4 - 1)\sigma^4}{N^2}\left(\sum_{k=1}^{m} b_k\right)^2 + \frac{n\sigma^4}{N^2}\sum_{k=1}^{m}(n-k)b_k^2.$$

For the second term in (16), note that $\boldsymbol{\varepsilon}^T H\boldsymbol{\varepsilon} = \sum_1^m h_i\varepsilon_i^2 + \sum_{n-m+1}^n h_i\varepsilon_i^2$. By Lemma 2, it is easy to see that

$$\begin{aligned} E\left(\sum_{i=1}^{m} h_i\varepsilon_i^2\right)^2 &= (\gamma_4 - 1)\sigma^4\frac{n^2}{4N^2}\sum_{i=1}^{m}\left(\sum_{\min(i,n+1-i,m+1)}^{m+1} b_k\right)^2 \\ &\quad + \frac{n^2\sigma^4}{4N^2}\left(\sum_{i=1}^{m}\sum_{\min(i,n+1-i,m+1)}^{m+1} b_k\right)^2 \\ &= O(m^2). \end{aligned}$$

Similarly, we have $E\left(\sum_{n-m+1}^n h_i\varepsilon_i^2\right)^2 = O(m^2)$. This leads to $\mathrm{E}(\boldsymbol{\varepsilon}^T H\boldsymbol{\varepsilon}/n)^2 = O(m^2/n^2)$. Further, for any $m = n^r$ with $0 < r < 1/2$,

$$\frac{1}{n}\boldsymbol{\varepsilon}^T H\boldsymbol{\varepsilon} = o_p(n^{-1/2}). \tag{18}$$

Combining (14), (15), (17) and (18), and applying the Slutsky theorem, we have

$$\frac{\sqrt{n}(\hat{\sigma}^2 - c_0\sigma^2)}{\sigma_c} \xrightarrow{\mathcal{D}} N(0,1), \quad \text{as } n \to \infty. \tag{19}$$

Note also that, by Lemma 2,

$$c_0 = \frac{n}{nm - m(m+1)/2}\left\{m - \frac{5m^2}{16n} + o(m)\right\} = 1 + O\left(\frac{m}{n}\right),$$
$$\sigma_c^2 = \frac{n^2(\gamma_4 - 1)\sigma^4}{N^2}\left(\sum_{k=1}^{m} b_k\right)^2 + \frac{n\sigma^4}{N^2}\sum_{k=1}^{m}(n-k)b_k^2 = (\gamma_4 - 1)\sigma^4 + o(1).$$

Thus for any $m = n^r$ with $0 < r < 1/2$, we have $\sqrt{n}(c_0 - 1) = o(1)$. In addition, $(\gamma_4 - 1)\sigma^4/\sigma_c^2 \to 1$ as $n \to \infty$. Then by (19) and the Slutsky theorem,

$$\begin{aligned}\frac{\sqrt{n}(\hat{\sigma}^2 - \sigma^2)}{\sqrt{(\gamma_4 - 1)\sigma^4}} &= \frac{\sigma_c}{\sqrt{(\gamma_4 - 1)\sigma^4}}\left\{\frac{\sqrt{n}(\hat{\sigma}^2 - c_0\sigma^2)}{\sigma_c} + \frac{\sqrt{n}(c_0 - 1)\sigma^2}{\sigma_c}\right\} \\ &\xrightarrow{\mathcal{D}} N(0,1), \quad \text{as } n \to \infty.\end{aligned}$$

Appendix 3: Derivation of covariances between Rice estimators

For any $1 \le b < k = o(n)$, we have

$$\begin{aligned}\mathrm{E}(s_b s_k) = \frac{1}{4(n-b)(n-k)} \quad & \left\{\sum_{i=k+1}^{n} \mathrm{E}(y_i - y_{i-k})^2(y_{i-k+b} - y_{i-k})^2\right. \\ & + \sum_{i=k+1}^{n} \mathrm{E}(y_i - y_{i-k})^2(y_i - y_{i-b})^2 \\ & + \sum_{i=k+b+1}^{n} \mathrm{E}(y_i - y_{i-k})^2(y_{i-k} - y_{i-k-b})^2 \\ & + \sum_{i=k+1}^{n-b} \mathrm{E}(y_i - y_{i-k})^2(y_{i+b} - y_i)^2 \\ & \left. + \sum_{(i,j)\in\mathcal{E}} \mathrm{E}(y_i - y_{i-k})^2(y_j - y_{j-b})^2\right\} \\ = \frac{1}{4(n-b)(n-k)} \quad & (I_1 + I_2 + I_3 + I_4 + I_5),\end{aligned}$$

where $\mathcal{E} = \{(i,j) : i = k+1,\ldots,n;\ j = b+1,\ldots,n;\ i \neq j;\ i \neq j-b;\ i-k \neq j;\ i-k \neq j-b\}$. It is easy to verify that $I_1 + I_2 = 2(n-k)(\gamma_4+3)\sigma^4 + O(k^2/n)$, $I_3 + I_4 = 2(n-k-b)(\gamma_4+3)\sigma^4 + O(k^2/n)$, and $I_5 = 4\{(n-k)(n-b) - 2(2n-2k-b)\}\sigma^4 + 4\sigma^2(n-b)(n-k)(b^2+k^2)J/n^2 + O(k^3/n)$. Therefore,

$$\mathrm{E}(s_b s_k) = \frac{2n-2k-b}{2(n-b)(n-k)}(\gamma_4-1)\sigma^4 + \sigma^4 + \frac{b^2+k^2}{n^2}J\sigma^2 + O\left(\frac{k^3}{n^3}\right).$$

Note also that $\mathrm{E}(s_b) = \sigma^2 + Jd_b + O(b^3/n^3) + o(1/n^2)$ and $\mathrm{E}(s_k) = \sigma^2 + Jd_k + O(k^3/n^3) + o(1/n^2)$. Thus,

$$\mathrm{Cov}(s_b, s_k) = \frac{2n-2k-b}{2(n-b)(n-k)}(\gamma_4-1)\sigma^4 + O\left(\frac{k^3}{n^3}\right) + o\left(\frac{1}{n^2}\right).$$

Finally, for any $k = n^r$ with $0 < r < 2/3$, we have $k^3/n^3 = o(1/n)$ and therefore $\mathrm{Cov}(s_b, s_k) = (\gamma_4-1)\sigma^4/n + o(1/n)$.

Appendix 4: Proof of Theorem 3

Lemma 3. *Assume that g has a bounded second derivative. Then for the equally spaced design with $n \to \infty$, $L \to \infty$ and $L/n \to 0$, we have*

(i) $\mathrm{tr}(M) = 2(n-L)$;

(ii) $\mathrm{tr}[\{\mathrm{diag}(M)\}^2] = 4n - 134L/35 + o(L)$;

(iii) $\mathrm{tr}(M^2) = 4n - 134L/35 + 18n/L + o(L) + o(n/L)$;

(iv) $\boldsymbol{g}^T M^2 \boldsymbol{g} = O(L^3/n^2)$;

(v) $\boldsymbol{g}^T M \mathrm{diag}(M)\mathbf{1} = O(L^2/n)$.

Proof of Lemma 3. It is easy to verify that $\sum_{k=1}^{L} a_k = 1$, $\sum_{k=1}^{i} a_k = 9i/L - 18i^2/L^2 + 10i^3/L^3 + o(i/L)$ for $1 \le i \le L$, $\sum_{k=1}^{L} a_k^2 = 9/L + o(1/L)$, $\sum_{k=1}^{L} k a_k = O(L)$, and $\sum_{k=1}^{L} k^2 a_k = O(L^2)$.

(i) $\mathrm{tr}(M) = 2L\sum_{k=1}^{L} a_k + 2(n-2L)\sum_{k=1}^{L} a_k = 2(n-L)$.

(ii) Note that $a_0 = 0$ and $\sum_{k=n-L+i}^{n} a_{k+L-n} = 1 - \sum_{k=0}^{i-1} a_k$. We have

$$\begin{aligned}
\mathrm{tr}[\{\mathrm{diag}(M)\}^2] &= 4(n-2L) + \sum_{i=1}^{L}\left(1+\sum_{k=0}^{i-1} a_k\right)^2 + \sum_{i=1}^{L}\left(1-\sum_{k=0}^{i-1} a_k\right)^2 \\
&= 4n - 6L + 2\sum_{i=1}^{L}\left\{\frac{9i}{L} - \frac{18i^2}{L^2} + \frac{10i^3}{L^3} + o\left(\frac{i}{L}\right)\right\}^2 \\
&= 4n - \frac{134}{35}L + o(L).
\end{aligned}$$

(iii) By (ii), we have

$$\begin{aligned}
\mathrm{tr}(M^2) &= \mathrm{tr}[\{\mathrm{diag}(M)\}^2] + \sum_{i=1}^{L}\left(\sum_{k=1}^{L} a_k^2 + \sum_{k=0}^{i-1} a_k^2\right) + 2\sum_{i=L+1}^{n-L}\sum_{k=1}^{L} a_k^2 + \sum_{i=1}^{L}\sum_{k=i}^{L} a_k^2 \\
&= \mathrm{tr}[\{\mathrm{diag}(M)\}^2] + 2(n-L)\sum_{k=1}^{L} a_k^2 \\
&= 4n - \frac{134}{35}L + \frac{18n}{L} + o(L) + o\left(\frac{n}{L}\right).
\end{aligned}$$

(iv) Noting that M is a symmetric matrix, we have $\boldsymbol{g}^T M^2 \boldsymbol{g} = (M\boldsymbol{g})^T M\boldsymbol{g} \triangleq \boldsymbol{h}^T\boldsymbol{h}$ where $\boldsymbol{h} = M\boldsymbol{g} = (h_1, \ldots, h_n)^T$. Under the condition that g has a bounded second derivative, it is easy to verify that for $i \in [L+1, n-L]$,

$$h_i = \sum_{k=1}^{L} a_k(g_i - g_{i-k}) - \sum_{k=1}^{L} a_k(g_{i+k} - g_i) = -\frac{1}{n^2} g_i'' \sum_{k=1}^{L} k^2 a_k + o\left(\frac{m^3}{n^2}\right) = O\left(\frac{L^2}{n^2}\right).$$

Similarly, we can show that for $i \in [1, L]$ or $i \in [n-L+1, n]$, $h_i = O(L/n)$. Finally,

$$\boldsymbol{g}^T M^2 \boldsymbol{g} = \boldsymbol{h}^T\boldsymbol{h} = \sum_{i=1}^{L} h_i^2 + \sum_{i=L+1}^{n-L} h_i^2 + \sum_{i=n-L+1}^{n} h_i^2 = O\left(\frac{L^3}{n^2}\right).$$

(v) Note that $\boldsymbol{g}^T[M\mathrm{diag}(M)\mathbf{1}] = (M\boldsymbol{g})^T\mathrm{diag}(M)\mathbf{1} = \boldsymbol{h}^T\mathrm{diag}(M)\mathbf{1}$. We have

$$\boldsymbol{g}^T[M\mathrm{diag}(M)\mathbf{1}] = \sum_{i=1}^{L} h_i \cdot O(1) + \sum_{i=L+1}^{n-L} h_i \cdot O(1) + \sum_{i=n-L+1}^{n} h_i \cdot O(1) = O\left(\frac{L^2}{n}\right).$$

Proof of Theorem 3. By Müller and Stadtmüller (1999), $\text{Bias}(\hat{\sigma}^2_{\text{MS}}) = \boldsymbol{g}^T M \boldsymbol{g}/\text{tr}(M) = o(L^2/n^2)$. Note that the last four terms in (9) make up the variance. By Lemma 3 and the facts that $L/n \to 0$ and $\sigma^4(\gamma_4 - 3) = \text{var}(\varepsilon^2) - 2\sigma^4$, we have

$$\begin{aligned}
\text{var}(\hat{\sigma}^2_{\text{MS}}) &= \frac{1}{4(n-L)^2}\left[\left\{\text{var}(\varepsilon^2) - 2\sigma^4\right\}\left\{4n - \frac{134}{35}L + o(L)\right\}\right. \\
&\quad \left. +2\sigma^4\left\{4n - \frac{134}{35}L + \frac{18n}{L} + o(L) + o\left(\frac{n}{L}\right)\right\}\right] \\
&= \frac{1}{4(n-L)^2}\left\{\left(4n - \frac{134}{35}L\right)\text{var}(\varepsilon^2) + \frac{36n}{L}\sigma^4 + o(L) + o\left(\frac{n}{L}\right)\right\} \\
&= \frac{1}{n}\text{var}(\varepsilon^2) + \frac{73L}{70n^2}\text{var}(\varepsilon^2) + \frac{9}{Ln}\sigma^4 + o\left(\frac{L}{n^2}\right) + o\left(\frac{1}{Ln}\right).
\end{aligned}$$

Finally, we have

$$\text{MSE}(\hat{\sigma}^2_{\text{MS}}) = \frac{1}{n}\text{var}(\varepsilon^2) + \frac{73L}{70n^2}\text{var}(\varepsilon^2) + \frac{9}{Ln}\sigma^4 + o\left(\frac{L}{n^2}\right) + o\left(\frac{1}{Ln}\right) + o\left(\frac{L^4}{n^4}\right).$$