Statistica Sinica **12**(2002), 1191-1205

ON REGRESSION ESTIMATORS WITH DE-NOISED VARIABLES

Hengjian Cui, Xuming He and Lixing Zhu

Beijing Normal University, University of Illinois and The University of Hong Kong

Abstract: We consider linear measurement error models where the variables in error are observed together with an auxiliary variable, say, time. Cai, Naik and Tsai (2000) studied this problem and proposed using a de-noising process prior to a least squares analysis. The present paper focuses on the asymptotic distributions of such de-noised estimators. We demonstrate that the use of de-noising contributes to an efficiency gain over other estimators of measurement error models that do not make use of any auxiliary information. We also extend the results to cases with dependent errors, and to a general class of M-estimators that have better robustness properties than least squares.

Key words and phrases: De-noising, errors-in-variables, kernel, linear model, measurement error, smoothing.

1. Introduction

Regression models with measurement errors arise frequently in practice and have attracted attention in the statistics literature. Since the least squares (LS) estimator is not consistent in the presence of measurement errors, a number of alternatives have been proposed. For example, the method of moments, leading to the adjusted least squares (ALS), can be used to correct for bias. A likelihood-based argument leads to least squares with orthogonal distances (OLS). A simulation-extrapolation method (SIMEX), which is equivalent to ALS in linear models, also works for nonlinear errors-in-variables models. We refer to Fuller (1987), Carroll, Ruppert and Stefanski (1995), and Cook and Stefanski (1994) for more details.

The present paper focuses on linear models where the variables in error are observed together with an auxiliary variable, say, time. Time will be used for the rest of the paper, but it is clear that any other auxiliary variable can take its place. Let $(\xi, \eta) \in R^p \times R^1$ be variables of interest that satisfy a linear relationship

$$\eta = \xi^\tau \beta_0 + z^\tau \alpha_0 \tag{1.1}$$

with some additional covariates $z \in R^q$. Measurements of (ξ, η) are collected over time to yield a data set of $\{(x_i, y_i, z_i), 1 \le i \le n\}$ with

$$x_i = \xi(t_i) + u_i \quad \text{and} \quad y_i = \eta(t_i) + v_i, \tag{1.2}$$

where t_i is time for the ith measurement, u_i and v_i are measurement errors. We assume that the z_i's are observed without error, and consider the problem of estimating the unknown parameters (β_0, α_0).

A key ingredient of our model is that both ξ and η are time-dependent. For a given time t, they can be viewed as the (unknown) population means of certain underlying variables. One example of using this model was given in Cai, Naik and Tsai (2000) for estimating the relationship between awareness and television rating points of TV commercials for certain products. The variable z_i may include the constant 1 to reflect the intercept in the model.

To be in line with usual regression models, we may rewrite (1.1) as

$$y_i = \xi_i^\tau \beta_0 + z_i^\tau \alpha_0 + v_i, \tag{1.3}$$

where $\xi_i = \xi(t_i)$ is subject to measurement error and the v_i's and u_i's are independent error variables. The cases where v_i are correlated over time will be discussed in Section 3.

The ordinary least squares estimate (LS) of (1.3) is biased and inconsistent. Cai, Naik and Tsai (2000) used wavelets to filter out noise in the observed variables. Let $\tilde{x}$ and $\tilde{y}$ denote the de-noised variables of x and y respectively. Under some smoothness conditions on $\xi(t)$ and $\eta(t)$, the least squares method applied to the de-noised variables yields a consistent estimator of β_0, called the DLS estimator.

The purpose of the present paper is to further study the effect of de-noising and the asymptotic properties of such de-noised estimators. In Section 2, we consider a specific procedure of DLS by using a kernel-type smoothing for x_i followed by least squares regression of y_i on $(\tilde{x}_i, z_i)$. We note that the de-noising of x is essential but there is no need to de-noise y. The consistency and asymptotic normality for the DLS are established. We also confirm that DLS enjoys an efficiency gain over ALS and OLS by making use of the time information in the data. In Section 3, we obtain asymptotic results for the cases where $\{v_i\}$ is a linear stationary process, thus extending the DLS methodology to a wider range of problems. In Section 4 we generalize our results to a class of de-noised M-estimators. The M-estimators are more robust than least squares against outliers in y_i, and are often more efficient for non-normal measurement errors. For robustness of M-estimators, we refer to Huber (1981). Section 5 reports a Monte Carlo comparison between DLS and ALS and illustrates the DLS methodology through two examples.

2. DLS Estimators and Their Asymptotics

The DLS estimators we consider here are in the spirit of Cai, Naik and Tsai (2000) but differ on some details. First, we de-noise only the x variable but not the y variable. This brings (1.3) closer to the traditional errors-in-variables regression framework. More importantly, the de-noising of y_i does not enhance the performance of the estimator. Second, we use a (convolution) kernel-type smoothing instead of wavelets de-noising. This difference is merely technical. The kernel-type smoothing is better known in statistics and easier to analyze, but under appropriate conditions our asymptotic normality results for DLS also hold for wavelets de-noising such as the one used by Antoniadis, Gregoire and McKeague (1994, p.1340).

For the model specification (1.1) and (1.2), we further assume that $u_i \in R^p$ and $v_i \in R$ are two independent random samples with mean 0 and variance-covariance Σ_u and σ_v^2 respectively. Without less of generality, we assume that the observations are taken at $0 = t_0 \le t_1 \le t_2 \le \cdots \le t_n \le 1$, where t_i can be time or any other input parameter for ξ and η. Note that any monotone and smooth transformation of t can be used in our set-up. For the sake of technical convenience, we assume in this section that

(B1) $d_n =: \max_{1\le i\le n}\{|t_i - i/n|\} = O(\log n/\sqrt{n})$, and $\max_{1\le i\le n}|t_i - t_{i-1}| = O(\log n/n)$.

If the unsorted $\{t_i\}$ is a random sample from any probability distribution F, then a monotone transformation $F(t_i)$ satisfies **(B1)** almost surely.

We now specify a kernel-type smoothing procedure for the x_i. To this end, let $K(\cdot) \ge 0$ be a symmetric and Lipschitz kernel supported on [-1,1] with $\int_{-1}^{1} K(x)dx = 1$. Let $w_n(s,t)$ $(0 \le s,\ t \le 1)$ be a weight function depending only on $\{t_1,\ldots,t_n\}$ and satisfying $\int_0^1 w_n(s,t)dt = 1$ for any $0 \le s \le 1$. More specifically, we take

$$w_n(s,t) = \frac{1}{h}\Big[K\Big(\frac{s-t}{h}\Big) + K\Big(\frac{s+t}{h}\Big)I_{\{0\le s,t\le h\}} + K\Big(\frac{2-s-t}{h}\Big)I_{\{1-h\le s,t\le 1\}}\Big] \qquad (2.1)$$

for some smoothing parameter $h = h_n$ satisfying $h \in (0, 1/2)$, $h \to 0$, and $nh/\log n \to \infty$ as $n \to \infty$. Then, the de-noised variable $\tilde{x}$ is given by

$$\tilde{x}_i = \sum_{j=1}^{n} x_j \int_{A_j} w_n(s,t_i)ds, \qquad (2.2)$$

where $A_1 = [0, (t_1 + t_2)/2)$, $A_j = [(t_{j-1} + t_j)/2, (t_j + t_{j+1})/2))$ $(2 \le j \le n-1)$, and $A_n = [(t_{n-1} + t_n)/2, 1]$. This corresponds to a smoothing method used by Gasser and Müller (1979). The additional terms in (2.1) for (s,t) near the ends (0 or 1) make bias corrections for kernel smoothing at the boundaries.

For notational convenience, let $X = (x_1,\ldots,x_n)^\tau \in R^{n\times p}$, $\tilde{X} = (\tilde{x}_1,\ldots,\tilde{x}_n)^\tau$

$\in R^{n\times p}$, $Z = (z_1, \ldots, z_n)^\tau \in R^{n\times q}$, $Y = (y_1, \ldots, y_n)^\tau \in R^n$, $\Xi = (\xi_1, \ldots, \xi_n)^\tau \in R^{n\times p}$, $U = (u_1, \ldots, u_n)^\tau \in R^{n\times p}$, and $V = (v_1, \ldots, v_n)^\tau \in R^n$. Also let

$$\Omega_n = \frac{1}{n}\begin{pmatrix} \Xi^\tau\Xi & \Xi^\tau Z \\ Z^\tau\Xi & Z^\tau Z \end{pmatrix}, \quad \tilde{\Omega}_n = \frac{1}{n}\begin{pmatrix} \tilde{X}^\tau\tilde{X} & \tilde{X}^\tau Z \\ Z^\tau\tilde{X} & Z^\tau Z \end{pmatrix}. \tag{2.3}$$

Note by the consistency of the smoother (2.2), $\sup_i |\tilde{x}_i - \xi(t_i)| \to 0$ and $\Omega_n - \tilde{\Omega}_n \to 0$ in probability as $n \to \infty$. The DLS method proceeds by regressing y_i on $(\tilde{x}_i, z_i)$ to get

$$\begin{pmatrix} \hat{\beta} \\ \hat{\alpha} \end{pmatrix} = (n\tilde{\Omega}_n)^{-1}\begin{pmatrix} \tilde{X}^\tau \\ Z^\tau \end{pmatrix} Y. \tag{2.4}$$

The following conditions are also assumed for the asymptotic results in this section.

(B2) There exist $0 < C_1 < C_2 < \infty$ such that $C_1 < \lambda_{min}(\Omega_n) \le \lambda_{max}(\Omega_n) \le C_2$ for all n, where λ_{min} and λ_{max} stand for the minimum and maximum eigenvalues of a matrix.

(B3) $\xi(t)$ is continuous in t, and so is $z(t)$ if $z_i = z(t_i)$.

(B4) $\xi'(t)$ is Lipschitz with order γ $(0 < \gamma \le 1)$, and so is $z'(t)$ if $z_i = z(t_i)$.

In typical cases where $\Omega_n \to \Omega$ for some matrix Ω as $n \to \infty$, the condition **(B2)** follows from positive definiteness of Ω. The smoothness condition on $z(t)$ as given in **(B3)** and in **(B4)** can be weakened so that it holds everywhere except at finitely many points. Also, **(B4)** is a stronger condition than **(B3)**. If $\xi(t)$ and $z(t)$ have bounded second order derivatives, then **(B4)** holds with $\gamma = 1$. Our main results follow.

Theorem 1. *Assume* **(B1), (B2)** *and* **(B3)**. *The estimator* $(\hat{\beta}, \hat{\alpha})$ *given by* (2.4) *is consistent for* (β_0, α_0).

Theorem 2. *Assume* **(B1), (B2)** *and* **(B4)**, *and* $n(h/\log n)^2 \to \infty$ *as* $n \to \infty$. *Then*

$$\begin{pmatrix} \hat{\beta} \\ \hat{\alpha} \end{pmatrix} - \begin{pmatrix} \beta_0 \\ \alpha_0 \end{pmatrix} = \frac{1}{n}\Omega_n^{-1}\begin{pmatrix} \Xi^\tau(V - U\beta_0) \\ Z^\tau(V - U\beta_0) \end{pmatrix} + O_p\Big(\log n/(nh) + h^{1+\gamma}\Big), \tag{2.5}$$

$$\frac{\sqrt{n}\Omega_n^{1/2}}{\sqrt{\sigma_v^2 + \beta_0^\tau\Sigma_u\beta_0}}\left[\begin{pmatrix} \hat{\beta} \\ \hat{\alpha} \end{pmatrix} - \begin{pmatrix} \beta_0 \\ \alpha_0 \end{pmatrix}\right] \xrightarrow{d.} N(0, I_{p+q}), \tag{2.6}$$

provided that $nh^{2+2\gamma} \to 0$ *as* $n \to \infty$.

Remark 1. Under **(B4)** with $\gamma = 1$, the smoothing parameter that leads to the optimal rate of convergence in estimating $\xi(t)$ is of order $n^{-1/5}$. Our asymptotic normality result (2.6) requires a smaller h strictly between $n^{-1/2}$ and $n^{-1/4}$ in this case. This corresponds to undersmoothing ξ so that the bias is kept small. Based on the asymptotic analysis and empirical experience, we suggest a rule of thumb as follows: the smoothing parameter h is so chosen that intervals of size $2h$ would contain around 5 points for n up to 100 and between $\frac{1}{8}n^{2/3}$ and $\frac{1}{4}n^{2/3}$ points for larger n.

It is easy to show that $s^2 = n^{-1}(Y - X\hat{\beta} - Z\hat{\alpha})^\tau(Y - X\hat{\beta} - Z\hat{\alpha})$ is a consistent estimate of $\sigma_v^2 + \beta_0^\tau \Sigma_u \beta_0$. We can estimate the large sample variance-covariance of $(\hat{\beta}, \hat{\alpha})$ by $s^2(n\tilde{\Omega}_n)^{-1}$. We can also estimate the measurement error variances of Σ_u and σ_v^2 by

$$\hat{\Sigma}_u = n^{-1}(X^\tau D_n X - \tilde{X}^\tau D_n \tilde{X}), \quad \hat{\sigma}_v^2 = n^{-1}(Y - \tilde{X}\hat{\beta} - Z\hat{\alpha})^\tau(Y - \tilde{X}\hat{\beta} - Z\hat{\alpha}),$$

where $D_n = I_n - n^{-1}ll^\tau$ is a centering matrix with $l = (1, ..., 1)^\tau \in R^n$ and I_n is the n by n identity matrix.

For any symmetric matrix A, let $vec(A)$ be the vectorization of A, that is, a vector that consists of all the elements in the upper triangular part of A.

Theorem 3. *Assume the conditions of Theorem* 2. *If* $E\|u_1\|^4 < +\infty$, *we have* $\sqrt{n}\Big(vec(\hat{\Sigma}_u) - vec(\Sigma_u)\Big) \xrightarrow{d.} N\Big(0, Var[vec(u_1 u_1^\tau)]\Big)$. *If* $Ev_1^4 < \infty$, *then* $\sqrt{n}(\hat{\sigma}_v^2 - \sigma_v^2) \xrightarrow{d.} N(0, Var(v^2))$.

The asymptotic normality (2.6) enables us to compare the asymptotic efficiency of the DLS estimator with other estimators developed for errors-in-variables models. For simplicity, we consider the case of $z_i = 1$ with $q = 1$.

Let $S_{\xi\xi} = n^{-1}\sum_{i=1}^n(\xi(t_i) - \overline{\xi})(\xi(t_i) - \overline{\xi})^\tau$ with $\overline{\xi} = n^{-1}\sum_{i=1}^n \xi(t_i)$, and $V_{0n} = (\sigma_v^2 + \beta_0^\tau \Sigma_u \beta_0)S_{\xi\xi}^{-1}$. It follows from Theorem 2 that $\sqrt{n}V_{0n}^{-1/2}(\hat{\beta}_1 - \beta_0) \xrightarrow{d.} N(0, I_p)$. If $\Sigma_u > 0$ is known, the ALS estimator of β_0 is $\hat{\beta}_{ALS} = \Big(S_{xx} - \Sigma_u\Big)^{-1} S_{xy}$, where $S_{xx} = n^{-1}\sum_{i=1}^n(x_i - \overline{x})(x_i - \overline{x})^\tau$, $S_{xy} = n^{-1}\sum_{i=1}^n(x_i - \overline{x})(y_i - \overline{y})$. Under appropriate conditions, we have $\sqrt{n}V_{1n}^{-1/2}(\hat{\beta}_{ALS} - \beta_0) \xrightarrow{d.} N(0, I_p)$ with $V_{1n} = V_{0n} + S_{\xi\xi}^{-1}Cov[u(v - u^\tau\beta_0)]S_{\xi\xi}^{-1}$, which means that the asymptotic variance-covariance of DLS is strictly smaller than that of ALS. The amount of difference increases with the covariance of $u(v - u^\tau\beta_0)$ relative to $S_{\xi\xi}$.

In the special case of $\Sigma_u = \sigma_v^2 I_p$ and $V_{0n} = (1 + \|\beta_0\|^2)\sigma_v^2 S_{\xi\xi}^{-1}$, we can use the orthogonal least squares estimate $\hat{\beta}_{OLS}$ of β_0, see Madansky (1959). In this case, we have $\sqrt{n}V_{2n}^{-1/2}(\hat{\beta}_{OLS} - \beta_0) \xrightarrow{d.} N(0, I_p)$ with

$$V_{2n} = V_{0n} + S_{\xi\xi}^{-1}Cov\Big\{(v - u^\tau\beta_0)[u + \frac{(v - u^\tau\beta_0)\beta_0}{1 + \|\beta_0\|^2}]\Big\}S_{\xi\xi}^{-1},$$

so DLS is strictly more efficient than the orthogonal least squares. More details about the likelihood-based orthogonal least squares estimates can be found in Cui and Li (1998), Liang, Hädle and Carroll (1999), and He and Liang (2000).

The efficiency gains of DLS over ALS and OLS can be substantial when the size of the measurement error is large. The method of de-noising makes use of the auxiliary information to consistently estimate the true values ξ_i. Unlike the other two methods, DLS needs no additional knowledge about the size of Σ_u relative to σ_v^2, and the asymptotic variance is not affected by the fourth moments of u_i and v_i.

3. DLS with Dependent Errors

Since the measurements of η are taken over time, it is often the case that v_i are not independent but should be modeled as a time series. The asymptotic consistency and normality results in Section 2 can be extended to such cases.

Suppose that $\{v_i\}$ is a linear stationary process of the form

$$v_i = \sum_{j=-\infty}^{\infty} b_j e_{i-j}, \tag{3.1}$$

where $\{e_j\}$ are i.i.d. with mean 0 and variance σ_e^2. Let $\rho(k) = cov(v_i, v_{i+k})$. We state additional conditions as

(B5) $0 < \sum_{j=-\infty}^{+\infty} b_j^2 < +\infty$.

(B6) $0 < \sum_{j=-\infty}^{+\infty} |b_j| < +\infty$.

Theorem 4. *Under the conditions of Theorem* 1 *and* (**B5**), *the DLS estimate* (2.4) *is consistent. Furthermore, if* (**B6**) *holds,*

$$\sqrt{n}\Big(\frac{1}{n}\Omega_n^{-1}(\Xi, Z)^\tau R_n(\Xi, Z)\Omega_n^{-1} + \beta_0^\tau \Sigma_u \beta_0 \Omega_n^{-1}\Big)^{-1/2}\left[\begin{pmatrix}\hat{\beta}_1\\ \hat{\alpha}_1\end{pmatrix} - \begin{pmatrix}\beta_0\\ \alpha_0\end{pmatrix}\right] \xrightarrow{d.} N(0, I_{p+q}),$$

where R_n is the n by n matrix whose ij-th element is $\rho(i-j)$.

Regular ARMA models are linear in the form of (3.1) with **(B6)** holding automatically, see Brokwell and Richard (1991). In particular, if $\{v_i\}$ follows an $AR(1)$ model such that $v_i = \rho v_{i-1} + e_i$ where $|\rho| < 1$, e_i are i.i.d. with mean 0 and variance σ_e^2, then $b_j = \rho^{|j|}$ and $R_n = (\sigma_e^2 \rho^{|i-j|})$. It is easy to get a consistent estimate of σ_e^2 and ρ in this case from the DLS residuals, and therefore Theorem 4 can be used to estimate the standard errors of the DLS estimates, see Example 2 in Section 5.

4. De-noised M-Estimators

The least squares estimators of regression are known to be sensitive to outliers in the data. Robust estimators can be more efficient when the error distributions are non-Gaussian and can protect us from gross errors in the data.

M-estimators are arguably the most popular robust methods. To be more specific, we consider an M-estimator (β_0, α_0) as

$$(\hat{\beta}_{1n}, \hat{\alpha}_{1n}) =: arg \min_{\beta \in R^p, \alpha \in R^q} \frac{1}{n} \sum_{i=1}^{n} \rho(y_i - \tilde{x}_i^\tau \beta - z_i^\tau \alpha), \tag{4.1}$$

where $\rho(\cdot)$ is an even and convex loss function satisfying $\rho(x) = \int_0^x \psi(s) ds$ on $x > 0$, $\psi(s)$ a non-decreasing score function with $E\psi(v_1) = 0$, $E[\psi(v_1 + s)] = a_0 s + O(s^2)$, and $E[\psi(v_1 + s) - \psi(v_1)]^2 = o(s)$ as $s \to 0$ for some constant $a_0 > 0$.

Least squares regression is a special M-estimator with $\rho(x) = x^2$, but to be robust against outlying y_i values, we need to choose ρ such that $\psi(x)$ is bounded. The least absolute deviation regression with $\rho(x) = |x|$ and $\psi(x) = sgn(x)$ is a well-known example in this class. In this section, we assume $E\psi^2(v_1) < \infty$ instead of $Ev_1^2 < \infty$. This condition is automatically satisfied if ψ is bounded whether v_i has finite variance or not. The asymptotic properties of such M-estimators are given below. As in Section 2, we assume the v_i are i.i.d. in this section.

Theorem 5. *Under* **(B1), (B2)** *and* **(B3)***, the M-estimator* (4.1) *is consistent for* (β_0, α_0)*. Furthermore, under* **(B4)** *with* $nh^{2+2\gamma} \to 0$ *and* $nh^2/(\log n)^4 \to \infty$ *as* $n \to \infty$*, we have*

$$\frac{\sqrt{n}\Omega_n^{1/2}}{\sqrt{E[\psi^2(v_1)]/a_0^2 + \beta_0^\tau \Sigma_u \beta_0}} \begin{pmatrix} \hat{\beta}_{1n} - \beta_0 \\ \hat{\alpha}_{1n} - \alpha_0 \end{pmatrix} \xrightarrow{d.} N(0, I_{p+q}).$$

If ψ is differentiable, it is easy to show that $a_0 = E\psi'(v_1)$. If $\psi(x) = sgn(x)$, then $a_0 = 2f_v(0)$ where f_v is the p.d.f of v_1. Theorem 2 can be viewed as a special case of Theorem 5 except that the condition on h is slightly weaker in the former.

5. Empirical Investigations

We report a simulation study when data are generated from (1.3) with $p = q = 1$. More specifically, samples of size $n = 1000$ are generated from $y_i = 1 + \xi(t_i) + v_i$ and $x_i = \xi(t_i) + u_i$. The variances of the normal variables u_i and v_i are chosen for different levels of signal-to-noise ratios $snr_x = std(\xi)/std(u)$ and $snr_y = std(\eta)/std(v)$, where std stands for standard deviation. The function ξ is taken to be the Doppler function (see, e.g., Donoho and Johnstone, (1994)), and $t_i = i/n$ for $i = 1, \ldots, n$. A similar setting was used in the simulation study of Cai, Naik and Tsai (2000) where they compared the performance of their DLS with LS. Here, our comparison is with a consistent estimator ALS of $(\beta_0, \alpha_0) = (1, 1)$. In this case, we use the kernel-type smoothing with $h = 0.007$

and $K(t) = \frac{3}{4}(1-t^2)I_{\{|t|\le 1\}}$ on $t \in (-1,1)$. Note that the ALS estimators used here assume the measurement error variance Σ_u to be known. This additional information, usually unavailable in practice, is not needed for DLS.

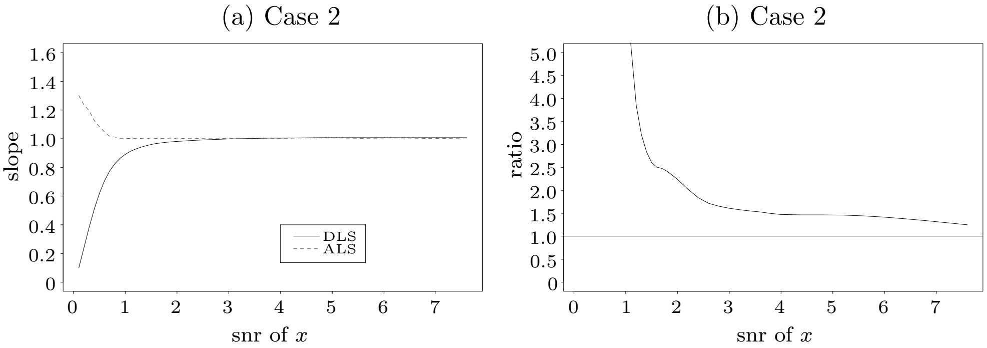


Figure 1. (a) Average slope estimates and (b) Ratio of standard errors of ALS relative to DLS when $snr_y = 10 - snr_x$.

Figure 1(a) gives the average parameter estimates of DLS and ALS (based on 500 Monte Carlo samples) for snr_x between 0.1 and 8 (in increments of 0.1) and $snr_y = 10 - snr_x$. It is clear that both estimators are nearly unbiased when $snr_x > 3$. When the signal-to-noise ratio is small, DLS has a negative bias but ALS has a positive bias. Figure 1(b) shows the estimated standard error of ALS relative to DLS under the same setup. As expected, the ALS is less efficient. At $snr_x = 2$, the relative efficiency of DLS relative to ALS is as large as 2, and increases rapidly as snr_x becomes smaller. When bias and variance are taken together, it is clear that DLS is a better performer, especially when snr_x is small. The comparisons are similar when we chose $snr_y = snr_x$.

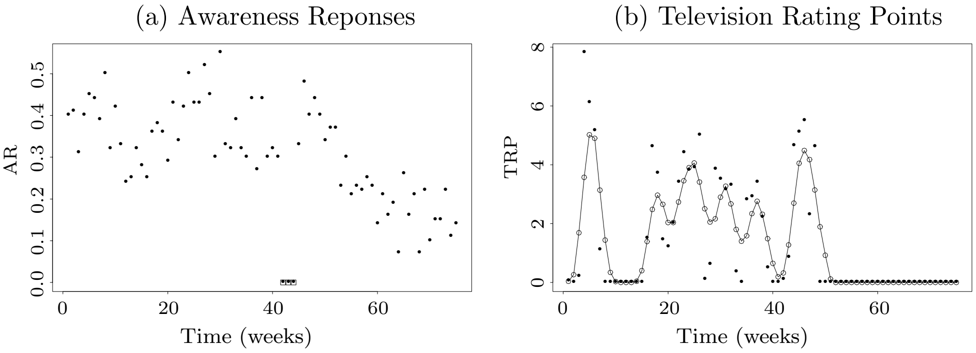
We now consider two examples to illustrate the applications of DLS.

Example 1. (Advertising) In measuring the effectiveness of television advertising, people have developed Television Rating Points (TRP) as a rough estimate of the extent of TV advertising. TRP is calculated based on several factors, including the length of the TV commercial. This is related to an Awareness Response (AR) reflecting the percentage of people who have seen that advertisement in a small survey of consumers. We take $\eta(t)$ as the true Awareness Response (from the whole targeted population) at time t, and $\xi(t)$ as the "true" extent of advertising approximated by TRP at time t. Here, we observe TRP and AR as data with measurement errors. Figure 2 gives the scatter-plots of the weekly AR and TRP data for a TV advertisement in its first 75 weeks. The data are taken from West and Harrison (1989, p.581). We model the relationship between $AR(t)$ and

$TRP(t)$ as

$$AR_i = \alpha + \beta_1 \xi_i + \beta_2 \xi_{i-3} + v_i, \quad i = 1, \ldots, 75, \tag{5.1}$$

where $TRP_i = \xi_i + u_i$ and $t_i = i/75$.



(a) Awareness Reponses

(b) Television Rating Points

Figure 2. (a) Scatterplot of awareness response; (b) Scatter-plot of Television Rating Points and the de-noised curve.

Following Remark 1, we use $h = 0.031$ with the same Epanechnikov kernel as used for our simulation above. Since three observations for AR are missing, only 72 observations are used in estimating the model. The DLS estimate of $(\alpha, \beta_1, \beta_2)$ is (0.216, 0.034, 0.034) and the standard errors of both slope estimates are 0.007. The estimated measurement error $\hat{\Sigma}_u = 1.553$ and $snr_x = 1.24$. By contrast, the LS estimate of the parameters based on the raw data is (0.241, 0.027, 0.024). That is, the LS estimate underestimates the slopes by more than one standard error. In other words, the estimate based on DLS indicates that an increase of one unit in the extent of TV advertising every week would make additional 6.8% of the targeted population aware of the campaign, as compared to 5.1% estimated by least squares without de-noising.

If we choose to use the de-noised least absolute deviation estimator in this example, we get the parameter estimates (0.220, 0.037, 0.028) with the standard errors of both slope estimates at 0.008. Without de-noising, the estimates are (0.230, 0.028, 0.026), so attenuation mainly occurs for the first slope parameter for this estimator. The residual plots of this regression example shows that the errors are close to normal so least squares is expected to yield a more efficient estimator.

Example 2. (Volatility) In finance and security analysis, we often measure the risk of an individual stock as its (standardized) regression slope against a market index. If this slope, usually called beta, is greater than 1, the change in the stock price is expected to be more than that in the index and thus the stock is considered to be more risky. An index is usually chosen to represent a broad market. If we ask how risky a stock is relative to a relevant market, then the use of an index results in measurement errors if the index covers only a portion of the stocks in the market of interest. As a result, the value of beta may be under-estimated. A correction of this bias can be made with DLS.

We consider the common stock price of Microsoft (MSFT) during the first ten months of year 2000, using with daily closing prices. To measure its risk relative to the market of U.S blue chip stocks, we take the Standard & Poor's 100 Index as a proxy to this market. (The Standard & Poor's 500 Index is more commonly used in finance, but the same methodology applies.) Figure 3 gives the time series plots for the MSFT price and the index over the 10 month period with 206 observations.

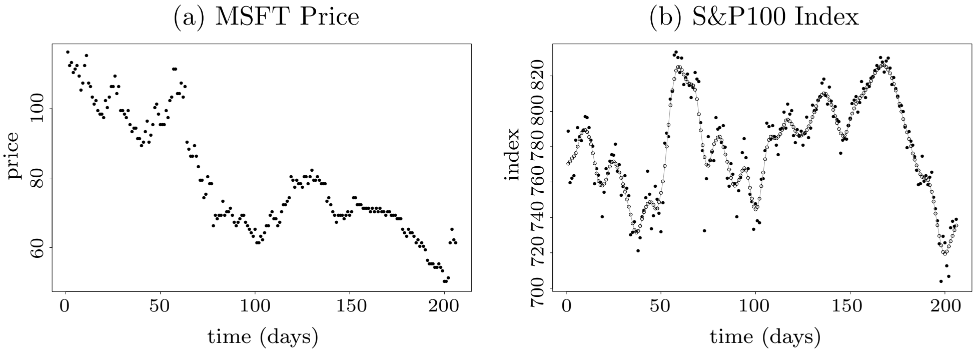
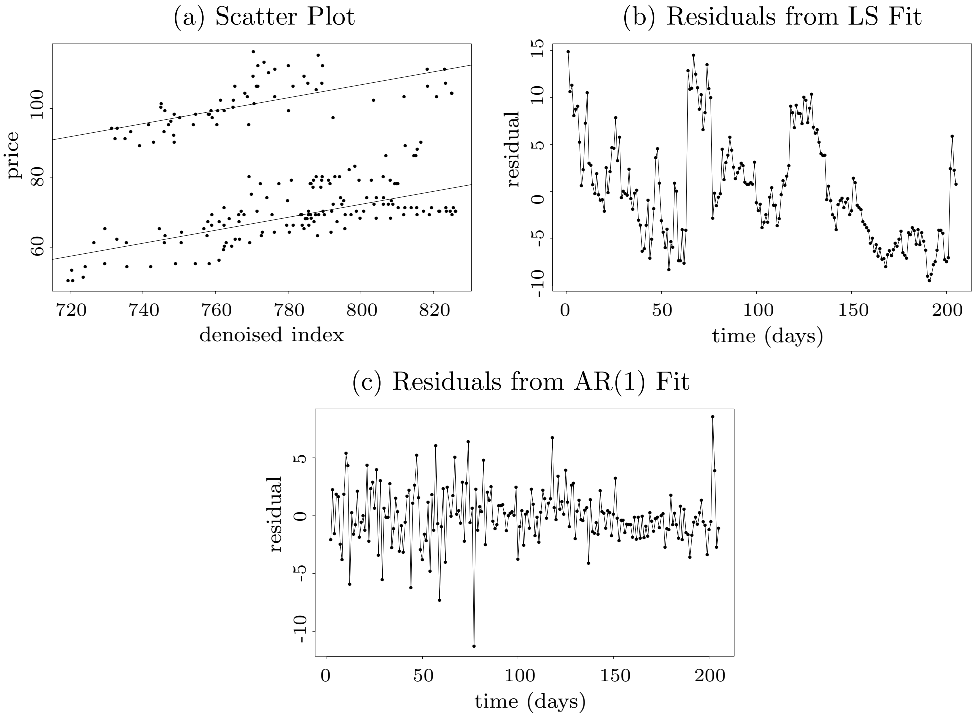


Figure 3. Time series plots of (a) Microsoft stock price and (b) S & P 100 index over the first ten months of year 2000.

To account for a sudden shift in the MSFT price at the beginning of April (see Figure 4(a)), we model the stock price gains y_i (the price at i-th day divided by the price on day one) as $y_i = \xi_i\beta+\alpha_0 I(i \le 64)+\alpha_1 I(i > 64)+v_i$, $i = 1, ..., 206$, where ξ_i denotes the change in market value from day one to the i-th day, which is measured (with some error) by the change in the S&P100 index. The intercept parameter α_0 is used for the first three months (with 64 trading days) and α_1 for the remaining days.

Ordinary least squares gives $\hat{\beta}_{LS} = 1.167$. If we test the hypothesis that $\beta = 1$ against $\beta > 1$, the p-value from the t-test is 0.054. But we can see from the residual plot in Figure 4(b) that there is clear dependence in the residuals.

We now assume that v_i follows an $AR(1)$ process and use the DLS to estimate the model parameters. We choose $h = n^{-1/3}/8$ for de-noising. The resulting estimate is $\hat{\beta}_{DLS} = 1.276$ with standard error of 0.135. The estimate of the autocorrelation for $\{v_i\}$ is $\hat{\rho} = 0.86$. Figure 6(c) shows the residual plot after fitting the $AR(1)$ process to $\{v_i\}$. It is not significantly different from white noise, validating our choice of $AR(1)$ in this case. For the test of $\beta = 1$ versus $\beta > 1$, we have a p-value of 0.023 based on DLS. At the usual 5% level of significance, we may conclude that MSFT is a stock that was more volatile than the U.S. blue chip market as a whole.


(a) Scatter Plot

(b) Residuals from LS Fit

(c) Residuals from AR(1) Fit

Figure 4. (a) Microsoft stock price versus de-noised index (b) Residual plot from the LS fit. (c) Residuals after fitting AR(1) to the errors.

6. Proofs of Main Results

In this section, we use a generic positive constant C which may vary from line to line. For notational convenience, let $I_n(i) = \tilde{\xi}_i - \xi_i = \sum_{j=1}^n \xi(t_j) \int_{A_j} w_n(s,t_i)ds - \xi(t_i)$, and $J_n(j) = \sum_{i=1}^n \xi(t_i) \int_{A_j} w_n(s,t_i)ds - \xi(t_j)$. Note that $J_n(j)$ relates to $\tilde{u}_j - u_j$ as we shall see later in the section. We first provide bounds on I_n and J_n.

Lemma 6.1. *For the de-noising method* (2.2) *we have the following results.*

(i) $\max_{0 \le t \le 1} \int_0^1 |w_n(s,t)|ds = O(1)$, *and* $\max_{1 \le j \le n, 0 \le t \le 1} |\int_{A_j} w_n(s,t)ds| = O(\log n/(nh))$ *if* **(B1)** *holds.*

(ii) *If* **(B1)** *and* **(B3)** *hold, then* $\max_{1 \le i \le n} \|I_n(i)\| = o(1)$.

(iii) *If* **(B1)** *and* **(B4)** *hold with* $n(h/\log n)^2 \to \infty$, *then*

$$\|I_n(i)\| = \begin{cases} O(h) & \text{for } t_i < h \text{ or } t_i > 1-h \\ O(h^{1+\gamma} + \log n/n) & \text{for } h \le t_i \le 1-h \end{cases},$$

$$\max_{1\le j\le n} \|J_n(j)\| = O(\log n(h + (nh)^{-1}) + d_n/h).$$

Proof. We only prove (iii) here, as the first two statements are easier to verify using similar arguments.

By the definition of $w_n(s,t)$, we have $\int_0^1 w_n(s,t)ds = 1$. Together with (i), we have

$$\begin{aligned}\|I_n(i)\| &\le \|\sum_{j=1}^n \int_{A_j} w_n(s,t_i)[\xi(t_j)-\xi(s)]ds\| + \|\sum_{j=1}^n \int_{A_j} w_n(s,t_i)\xi(s)ds - \xi(t_i)\| \\ &\le C\log n/n + \|\int_0^1 w_n(s,t_i)[\xi(s)-\xi(t_i)]ds\|.\end{aligned}$$

Using the Lipschitz property of $\xi'(t)$, we can expand $\xi(s)$ at $s = t_i$ to show that

$$\int_0^1 w_n(s,t_i)[\xi(s)-\xi(t_i)]ds = O(h^{1+\gamma}), \quad \text{if} \;\; h \le t_i \le 1-h.$$

For other t_i near the boundaries, the above term is $O(h)$. Thus we have proven the bound on $I_n(i)$ in (iii).

Since K has a finite support, there are only $O(nh)$ nonzero terms in $\{K((t_i - t)/h), 1 \le i \le n\}$. Together with **(B1)**, we have

$$\frac{1}{nh}\sum_{i=1}^n K\Big(\frac{t_i-t}{h}\Big)\xi(t_i) - \int \frac{1}{h}K\Big(\frac{s-t}{h}\Big)\xi(s)ds = O(\log n/(nh) + d_n/h),$$

$$\max_{0\le s\le 1} \|\frac{1}{n}\sum_{i=1}^n w_n(s,t_i)\xi(t_i) - \int_0^1 w_n(s,t)\xi(t)dt\| = O(\log n/(nh) + d_n/h), \quad (6.1)$$

and therefore $J_n(j) = n\int_{A_j}[\int_0^1 w_n(s,t)(\xi(t)-\xi(t_j))dt]ds + O(\log n/(nh) + d_n/h) = O(\log n(h + (nh)^{-1}) + d_n/h)$, which completes the proof of Lemma 6.1.

Proof of Theorem 1. Consistent with the notation in Section 2, let $\tilde\Xi = (\tilde\xi(t_1),\ldots,\tilde\xi(t_n))^\tau$, $\Pi = (\eta(t_1),\ldots,\eta(t_n))^\tau$, $\tilde U = (\tilde u_1,\ldots,\tilde u_n)^\tau$. By Lemma 6.1, we have

$$n^{-1}\|\tilde\Xi - \Xi\|^2 \le \max_{1\le i\le n}\|I_n(i)\|^2 = o(1), \quad n^{-1}\|(\tilde\Xi-\Xi)^\tau\Xi + \Xi^\tau(\tilde\Xi-\Xi)\| = o(1),$$

$$n^{-1}E[\tilde U^\tau \tilde U] = \frac{1}{n}\sum_{i=1}^n\sum_{j=1}^n[\int_{A_j} w_n(s,t_i)ds]^2\Sigma_u = O(\log n/(nh)), \quad (6.2)$$

Note that

$$\begin{aligned}n^{-1}(\tilde X^\tau\tilde X - \Xi^\tau\Xi) = n^{-1}[&(\tilde\Xi-\Xi)^\tau(\tilde\Xi-\Xi) + (\tilde\Xi-\Xi)^\tau\Xi + \Xi^\tau(\tilde\Xi-\Xi) + \tilde U^\tau\Xi + \Xi^\tau\tilde U \\ &+ (\tilde\Xi-\Xi)^\tau\tilde U + \tilde U^\tau(\tilde\Xi-\Xi) + \tilde U^\tau\tilde U] = o_p(1).\end{aligned}$$

Similar arguments lead to

$$\begin{pmatrix} \tilde{X}^\tau\tilde{X} & \tilde{X}^\tau Z \\ Z^\tau\tilde{X} & Z^\tau Z \end{pmatrix}^{-1} \begin{pmatrix} \tilde{X}^\tau\Xi & \tilde{X}^\tau Z \\ Z^\tau\Xi & Z^\tau Z \end{pmatrix} \to I_{p+q}, \quad and \quad \begin{pmatrix} \tilde{X}^\tau\tilde{X} & \tilde{X}^\tau Z \\ Z^\tau\tilde{X} & Z^\tau Z \end{pmatrix}^{-1} \begin{pmatrix} \tilde{X}^\tau \\ Z^\tau \end{pmatrix} V \to 0$$

in probability, from which the consistency result of Theorem 1 follows.

Proof of Theorem 2. By Lemma 6.1 and (6.2), we have

$$n^{-1}E\Big[(\tilde{X}-\Xi)^\tau(\tilde{X}-\Xi)\Big] = n^{-1}(\tilde{\Xi}-\Xi)^\tau(\tilde{\Xi}-\Xi)+n^{-1}E[\tilde{U}^\tau\tilde{U}] = O(h^2+\log n/(nh)),$$

and therefore

$$n^{-1}(\tilde{X}-\Xi)^\tau(\tilde{X}-\Xi) = O_p(h^2+\log n/(nh)). \tag{6.3}$$

By the definition of $I_n(i)$, we have $\Xi^\tau(\tilde{\Xi}-\Xi) = \sum_{i=1}^n \xi(t_i)I_n^\tau(i)$. Splitting the sum into those over $\{i: \ t_i \in (h, 1-h)\}$ and its compliment set (of size $O(nh)$), and using Lemma 6.1 again, we obtain

$$\begin{aligned} n^{-1}\|\Xi^\tau(\tilde{\Xi}-\Xi)\| &\le \frac{C}{n}\left[nh \max_{t_i<h, t_i>1-h}\|I_n(i)\| + n\max_{h\le t_i\le 1-h}\|I_n(i)\|\right] \\ = O(h^2) + O(h^{1+\gamma}+\log n/n) &= O_p(h^{1+\gamma}+\log n/n). \end{aligned} \tag{6.4}$$

Similarly, we have

$$\begin{aligned} n^{-2}E\|\Xi^\tau(\tilde{U}-U)\|^2 &= n^{-2}E\|\sum_{j=1}^n[\sum_{i=1}^n \xi(t_i)\int_{A_j} w_n(s,t_i)ds - \xi(t_j)]u_s\|^2 \\ \le \frac{\|\Sigma_u\|}{n^2}\sum_{j=1}^n\|J_n(j)\|^2 &= O\Big(n^{-1}[h\log n + \log n/(nh) + d_n/h]^2\Big), \end{aligned}$$

and thus (using $nh^2 \to \infty$)

$$n^{-1}\Xi^\tau(\tilde{U}-U) = O_p(h\log n/\sqrt{n} + d_n/(\sqrt{n}h)). \tag{6.5}$$

It follows from (6.3)−(6.5) that

$$\begin{aligned} n^{-1}(\tilde{X}^\tau\tilde{X}-\tilde{X}^\tau\Xi) &= n^{-1}\{(\tilde{X}-\Xi)^\tau(\tilde{X}-\Xi)+\Xi^\tau(\tilde{\Xi}-\Xi)+\Xi^\tau(\tilde{U}-U)+\Xi^\tau U\} \\ &= n^{-1}\Xi^\tau U + O_p(h^{1+\gamma}+\log n/(nh)). \end{aligned} \tag{6.6}$$

Similarly, we have

$$\begin{aligned} n^{-1}Z^\tau(\tilde{X}-\Xi) &= n^{-1}\{Z^\tau(\tilde{\Xi}-\Xi)+Z^\tau(\tilde{U}-U)+Z^\tau U\} \\ &= n^{-1}Z^\tau U + O_p(h^{1+\gamma}+\log n/(nh)). \end{aligned} \tag{6.7}$$

Based on (6.6) and (6.7), we have

$$\begin{pmatrix} \tilde{X}^\tau \tilde{X} & \tilde{X}^\tau Z \\ Z^\tau \tilde{X} & Z^\tau Z \end{pmatrix}^{-1} \begin{pmatrix} \tilde{X}^\tau \Xi & \tilde{X}^\tau Z \\ Z^\tau \Xi & Z^\tau Z \end{pmatrix} = I_{p+q} - (n\Omega_n)^{-1} \begin{pmatrix} \Xi^\tau U & 0 \\ Z^\tau U & 0 \end{pmatrix} + O_p(h^{1+\gamma} + \log n/(nh)).$$

Similarly, we have $n^{-1}(\tilde{X} - \Xi)^\tau V = o_p(h^{1+\gamma} + \log n/(nh))$, and therefore

$$\begin{aligned} \begin{pmatrix} \hat{\beta} \\ \hat{\alpha} \end{pmatrix} - \begin{pmatrix} \beta_0 \\ \alpha_0 \end{pmatrix} &= \left[\begin{pmatrix} \tilde{X}^\tau \tilde{X} & \tilde{X}^\tau Z \\ Z^\tau \tilde{X} & Z^\tau Z \end{pmatrix}^{-1} \begin{pmatrix} \tilde{X}^\tau \Xi & \tilde{X}^\tau Z \\ Z^\tau \Xi & Z^\tau Z \end{pmatrix} - I_{p+q} \right] \begin{pmatrix} \beta_0 \\ \alpha_0 \end{pmatrix} \\ &\quad + \begin{pmatrix} \tilde{X}^\tau \tilde{X} & \tilde{X}^\tau Z \\ Z^\tau \tilde{X} & Z^\tau Z \end{pmatrix}^{-1} \begin{pmatrix} \tilde{X}^\tau \\ Z^\tau \end{pmatrix} V \\ &= \frac{1}{n} \Omega_n^{-1} \begin{pmatrix} \Xi^\tau (V - U\beta_0) \\ Z^\tau (V - U\beta_0) \end{pmatrix} + O_p(h^{1+\gamma} + \log n/(nh)). \end{aligned}$$

The rest of the proof follows easily.

Proof of Theorem 3. Following Theorem 2, it is straightforward to verify that $\hat{\Sigma}_u = n^{-1} U^\tau U + o_p(1)$ and $\hat{\sigma}_v^2 = n^{-1} \sum_{i=1}^n v_i^2 + o_p(1)$.

Proof of Theorem 4. First we note that $\rho(k) = \sigma_e^2 \sum_j b_j b_{j-k}$ and $\sum_k |\rho(k)| \le (\sum_j |b_j|)^2$. Following the proof of Theorem 1, it suffices for the consistency result to verify

$$n^{-1}(\tilde{X}, Z)^\tau V = o_p(1), \tag{6.8}$$

This follows from $E \| \frac{1}{n} (\Xi,\, Z)^\tau V \|^2 = \frac{1}{n^2} trace\{ (\Xi,\, Z)^\tau E(VV^\tau)(\Xi, Z) \} = O(n^{-1} \lambda_{\max}(R_n)) = O(n^{-1} \sum_{k=0}^n |\rho(k)|) = o(1)$, and $E\|(\tilde{X} - \Xi)V\|^2 \le E\|\tilde{X} - \Xi\|^2 \lambda_{\max}(R_n) \le 2\sum_{k=0}^n |\rho(k)| E\|\tilde{X} - \Xi\|^2 = o(n^2)$.

For the asymptotic normality to work as in the proof of Theorem 2, we need to verify that $\frac{1}{n}(\tilde{X} - \Xi)V = o_p(h^{1+\gamma} + \log n/(nh))$ and $\frac{1}{\sqrt{n}}(\Xi, Z)^\tau V$ is asymptotically normal. They both follow from similar arguments leading to (6.8) using the weak correlations of v_i as dictated by **(B5)**. We omit the details.

Proof of Theorem 5. Using similar arguments to those of He and Shao (1996), a Bahadur-type representation holds for the M-estimator

$$\begin{pmatrix} \hat{\beta}_{1n} \\ \hat{\alpha}_{1n} \end{pmatrix} - \begin{pmatrix} \beta_0 \\ \alpha_0 \end{pmatrix} = -(a_0 n \Omega_n)^{-1} \sum_{i=1}^n \psi(e_i) \begin{pmatrix} \tilde{x}_i \\ z_i \end{pmatrix} + o_p(n^{-1/2}),$$

where $e_i = y_i - \tilde{x}_i^\tau \beta_0 - z_i^\tau \alpha_0$. Since the first order term of the above representation takes the same form as the least squares with e_i replaced by $\psi(e_i)/a_0$, the arguments used for the proof of Theorem 2 carry over. We omit the details.

Acknowledgements

Research of H. Cui was partially supported by Natural Science Foundation of China and EYTF of the State Education Commission of China. Research of X. He was partially supported by the University of Illinois Campus Research Board, NSF Grant DMS-0102411 and NSA Grant MDA904-02-1-0092. The authors thank a referee for helpful comments.

References

Antoniadis, A., Gregoire, G. and McKeague, I. W. (1994). Wavelet methods for curve estimation. *J. Amer. Statist. Assoc.* **89**, 1340-1353.

Brokwell P. J. and Richard A. D. (1991). *Times Series, Theory and Methods.* Springer-Verlag, New York.

Cai, Z., Naik, P. A. and Tsai, C. L. (2000). De-noised least squares estimators: an application to estimating advertising effectiveness. *Statist. Sinica* **10**, 1231-1243.

Carroll, R. J., Ruppert, D. and Stefanski, L. A. (1995). *Measurement Error in Nonlinear Models.* Chapman and Hall, London.

Cook, J. R. and Stefanski, L. A. (1994). Simulation-extrapolation estimation in parametric measurement error models. *J. Amer. Statist. Assoc.* **89**, 1314-1328.

Cui H. J. and Li, R. C. (1998). On parameter estimating in semi-linear EV models. *J. Multivariate Anal.* **64** 1-24.

Donoho, D. L. and Johnstone, I. M. (1994). Ideal spatial adaptation by wavelet shrinkage. *Biometrika* **81**, 425-455.

Fuller, W. A. (1987). *Measurement Error Models.* Wiley, New York.

Gasser, T. and Müller, H. G. (1979). Kernel estimation of regression function. In *Smoothing Techniques for curve Estimation* (Edited by Gasser and Rosenblatt). Springer-Verlag, Heidelberg.

He, X. and Liang H. (2000). Quantile regression estimates for a class of linear and partially linear errors-in-variables models. *Statist. Sinica* **10**, 129-140.

He, X. and Shao, Q. M. (1996). A general Bahadur representation of M-estimators and its application to linear regression with nonstochastic designs. *Ann. Statist.* **24**, 2608-2630.

Huber, P. (1981). *Robust Statistics*, Wiley, New York.

Liang H., Hädle, W. and Carroll, R. J. (1999). Estimation in a semiparametric partially linear errors-in-variables model. *Ann. Statist.* **27** 1519-1535.

Madansky, A. (1959). The fitting of straight lines when both variables are subject to error. *J. Roy. Statist. Soc. Ser. B* **54**, 173-205.

West, M. and Harrison J. (1997). *Bayesian Forecasting and Dynamic Models, Springer Series in Statistics.* Springer-Verlag, New York.

Department of Mathematics, Beijing Normal University, Beijing, China.
E-mail: hjcui@bnu.edu.cn

Department of Statistics, University of Illinois, Champaign, IL 61820, USA.
E-mail: he@stat.uiuc.edu

Department of Statistics and Actuarial Science, The University of Hong Kong, Hong Kong.
E-mail: lzhu@hku.hk

(Received February 2001; accepted May 2002)