Communications for Statistical Applications and Methods
2019, Vol. 26, No. 1, 47–55

https://doi.org/10.29220/CSAM.2019.26.1.047
Print ISSN 2287-7843 / Online ISSN 2383-4757

A two-step approach for variable selection in linear regression with measurement error

Jiyeon Song[a], Seung Jun Shin[1,b]

[a]Department of Statistics, University of Connecticut, USA;
[b]Department of Statistics, Korea University, Korea

Abstract

It is important to identify informative variables in high dimensional data analysis; however, it becomes a challenging task when covariates are contaminated by measurement error due to the bias induced by measurement error. In this article, we present a two-step approach for variable selection in the presence of measurement error. In the first step, we directly select important variables from the contaminated covariates as if there is no measurement error. We then apply, in the following step, orthogonal regression to obtain the unbiased estimates of regression coefficients identified in the previous step. In addition, we propose a modification of the two-step approach to further enhance the variable selection performance. Various simulation studies demonstrate the promising performance of the proposed method.

Keywords: measurement error, penalized orthogonal regression, SIMEX

1. Introduction

With the growth of high dimensional data, variable selection becomes a primal task in statistical learning. Since the prediction accuracy of final models relies heavily on selected variables. Regularization has become one of the canonical approaches for variable selection due to its fast and promising performance under high-dimensional setups since the proposal of the least absolute shrinkage and selection operator (LASSO) (Tibshirani, 1996). In addition to the L_1 penalty for LASSO, nonconvex penalties such as the smoothly clipped absolute deviation (SCAD) (Fan and Li, 2001) and the minimax concave plus penalty (MCP) (Zhang, 2010) have been widely employed as alternatives.

Measurement error in variables is commonly observed in practice. Let the dataset be $(y_i, \mathbf{x}_i) \in \mathbb{R} \times \mathbb{R}^p$ for $i = 1, \dots, n$ generated from $y_i = \mathbf{x}_i\boldsymbol{\beta} + \epsilon_i$ where $\boldsymbol{\beta} = (\beta_1, \dots, \beta_p)$ and ϵ_i is the random error with mean 0 and variance σ^2, and independent of $\mathbf{x}_i$, $i = 1, \dots, n$. Without measurement error, the least squares estimate would be an obvious choice. However, suppose that instead of $\mathbf{x}_i$, we observe the contaminated predictor $\mathbf{w}_i$ by the measurement error $\mathbf{u}_i$. That is,

$$\mathbf{w}_i = \mathbf{x}_i + \mathbf{u}_i, \quad i = 1, \dots, n, \tag{1.1}$$

where $\mathbf{u}_i$ denotes the measurement error with mean zero and covariance $\boldsymbol{\Sigma}_\mathbf{u}$, and independent of both $\mathbf{x}_i$ and ϵ_i. Assuming $E(\mathbf{x}) = \mathbf{0}$, $\mathrm{Cov}(\mathbf{x}) = \sigma_\mathbf{x}^2\mathbf{I}_p$, and $\mathrm{Cov}(\mathbf{u}) = \sigma_\mathbf{u}^2\mathbf{I}_p$ where $\sigma_\mathbf{x}$ and $\sigma_\mathbf{u}$ are scalars, we

[1] Corresponding author: Department of Statistics, Korea University, 145 Anam-Ro, Sungbuk-Gu, Seoul 02841, Korea. E-mail: sjshin@korea.ac.kr

Published 31 January 2019 / journal homepage: http://csam.or.kr

compute the population regression coefficient of y on $\mathbf{w}$ denoted by $\boldsymbol{\beta}^*$ in order to illustrate the effect of the measurement error:

$$\boldsymbol{\beta}^* = \{\mathrm{Cov}(\mathbf{w})\}^{-1}\mathrm{Cov}(\mathbf{w}, Y) = \{\mathrm{Cov}(\mathbf{x} + \mathbf{u})\}^{-1}\mathrm{Cov}(\mathbf{x} + \mathbf{u}, Y) = \frac{\sigma_{\mathbf{x}}^2}{\sigma_{\mathbf{u}}^2 + \sigma_{\mathbf{x}}^2}\boldsymbol{\beta},$$

where $\boldsymbol{\beta} = \{\mathrm{Cov}(\mathbf{x})\}^{-1}\mathrm{Cov}(\mathbf{x}, y)$ denotes the true target of interest. This shows that $\boldsymbol{\beta}^*$ does not consider a measurement error biased toward zero and this bias is called attenuation (Fuller, 1987). Therefore the implication of ignoring a measurement error could be considerable in inferential procedure.

Numerous methods have been developed for measurement error models (Carroll *et al.*, 2006, and references); however, is limited research on variable selection considering the measurement error effect. Liang and Li (2009) proposed SCAD-penalized orthogonal regression for a partially linear model and Ma and Li (2010) elegantly developed the penalized version of an estimating equation that can be applied to more broad families of semi-parametric models. Both ideas attempt to achieve variable selection and estimation simultaneously.

In this paper, we propose a two-step procedure for variable selection in linear regression with measurement errors. The proposed process conducts selection and estimation separately. We firstly identify important variables without considering measurement error by applying conventional regularized methods directly to $(y_i, \mathbf{w}_i)$. We then obtain unbiased coefficient estimates only for the selected variables via orthogonal regression. Our idea shares a conceptual similarity with the least angle regression - ordinary least squares (LARS-OLS) hybrid (Efron *et al.*, 2004) and the relaxed LASSO (Meinshausen, 2007) that separate variable selection and estimation steps in linear regression without measurement error. In addition, we propose a simple modification in the first step to enhance the variable selection performance by reducing false positives via random partitioning.

2. Orthogonal regression

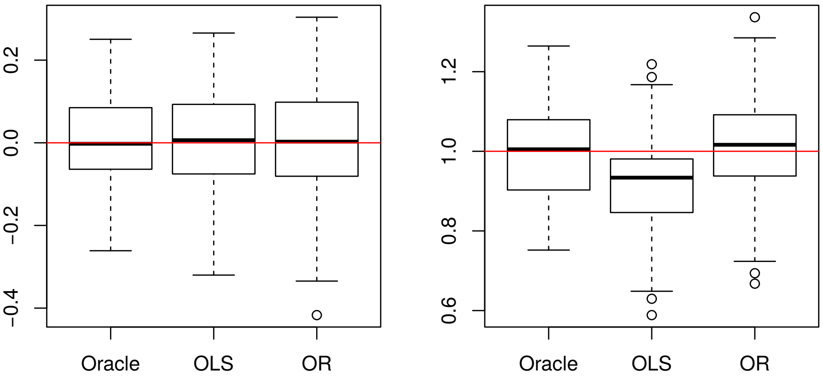
Orthogonal regression has been regarded as one of the popular choices for bias correction. The orthogonal regression assumes that $\mathbf{X} = (\mathbf{x}_1, \ldots, \mathbf{x}_n)^\top$ is an unknown fixed constant to be estimated. Consequently, the orthogonal regression minimizes the following objective function with respect to the unknown parameter $\boldsymbol{\beta}$ and the unobservable covariate $\mathbf{X}$.

$$L(\boldsymbol{\beta}, \mathbf{X}) = (\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^\top(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}) + \gamma \sum_{i=1}^{n}(\mathbf{w}_i - \mathbf{x}_i)^\top \boldsymbol{\Sigma}_{\mathbf{u}}^{-1}(\mathbf{w}_i - \mathbf{x}_i), \tag{2.1}$$

where $\mathbf{y} = (y_1, \ldots, y_n)^\top$, $\gamma = \sigma^2$ which is unknown. Here we used a new notation γ to denote σ^2 to emphasize that it will be treated as a tuning parameter instead of a model parameter. In this regards, we develop a data-adaptive method to choose a proper value of γ, which will be discussed in Section 4. The measurement error variance $\boldsymbol{\Sigma}_{\mathbf{u}}$ is often assumed to be known in practice since it can be estimated by repeatedly measuring the predictors (or is sometimes available from external sources). When $\gamma = 1$, the objective function becomes the orthogonal distance between $(y_i, \mathbf{w}_i)$ and $(\mathbf{x}_i^\top\boldsymbol{\beta}, \mathbf{x}_i)$, which explains its name. Here we let $\hat{\boldsymbol{\beta}}$ be the orthogonal regression estimator where $(\hat{\boldsymbol{\beta}}, \hat{\mathbf{X}}) = \mathrm{argmin}_{\boldsymbol{\beta},\mathbf{X}}\, L(\boldsymbol{\beta}, \mathbf{X})$.

To illustrate this, we consider a set of data $(y_i, \mathbf{w}_i)$, $i = 1, \ldots, n$ from the following linear regression model with the measurement error.

$$\begin{aligned} y_i &= \mathbf{x}_i^\top\boldsymbol{\beta} + \epsilon_i, \\ \mathbf{w}_i &= \mathbf{x}_i + \mathbf{u}_i, \end{aligned}$$



(a) Uninformative variable

(b) Informative variable

Figure 1: *Boxplots for coefficient estimates of (a) uninformative variable (β_1) and (b) informative variable (β_2).*

where $\boldsymbol{\beta} = (0, 1)$, $\mathbf{x}_i \overset{\text{iid}}{\sim} N_2(0, \mathbf{I})$, $\epsilon_i \overset{\text{iid}}{\sim} N(0, 1)$, and $\mathbf{u}_i \overset{\text{iid}}{\sim} N_2(0, \mathbf{I})$ with $n = 100$. We use the three estimation methods: least squares regression on $\mathbf{X}$ (denoted by Oracle) and least squares regression on $\mathbf{W}$ (OLS), and the orthogonal regression on $\mathbf{W}$ (OR). Figure 1 shows the boxplots of the coefficient estimates obtained by the three methods over 100 independent repetitions. Notice that β_1 is uninformative and there is no attenuation observed even for OLS (subpanel (a)). However, subpanel (b) for the informative variable β_2 depicts that attenuation is clearly observed in OLS. The orthogonal regression corrects the attenuation and its distribution is similar to the oracle estimates without the measurement error.

The least squares estimate for $(y_i, \mathbf{w}_i)$ is biased; however, it still contains a significant amount of information for identifying informative variables $\mathcal{S} = \{j : \beta_j \neq 0\}$ since the goal of the variable selection is not to estimate exact value of β_j, $j = 1, \ldots, p$ but to check whether $\beta_j = 0$ or not. This motivates us to develop a two-step approach that separates selection step and estimation step.

3. Two-step procedure

We develop a two-step procedure for variable selection in linear regression with measurement error. The key idea of the two-step procedure is to separate selection and estimation.

In the first step we consider the following regularized linear regression for $(y_i, \mathbf{w}_i)$

$$\hat{\boldsymbol{\beta}}_{(1)} := \left(\beta_1^{(1)}, \ldots, \beta_p^{(1)}\right)^\top = \underset{\boldsymbol{\beta}}{\operatorname{argmin}} (\mathbf{y} - \mathbf{W}\boldsymbol{\beta})^\top (\mathbf{y} - \mathbf{W}\boldsymbol{\beta}) + n \sum_{j=1}^{p} p_\lambda(|\beta_j|),$$

where $\mathbf{W} = (\mathbf{w}_1, \ldots, \mathbf{w}_n)^\top$, p_λ denotes the sparsity-inducing penalty function such as L_1 norm penalty for LASSO and non-convex penalties including SCAD and MCP and $\lambda > 0$ is a tuning parameter. The role of λ is to control the degree of sparsity of the solution $\hat{\boldsymbol{\beta}}_{(1)}$ that can be data-adaptively chosen via cross-validation.

Notice that $\hat{\boldsymbol{\beta}}_{(1)}$ is obviously biased. However, the goal of the first step is not to estimate $\boldsymbol{\beta}$ but to estimate $\mathcal{S}$. Therefore, we have

$$\hat{\mathcal{S}} = \left\{ j \mid \hat{\beta}_j^{(1)} \neq 0, \ j = 1, \ldots, p \right\}$$

in the first step.

Given $\hat{\mathcal{S}}$, we can estimate $\boldsymbol{\beta}_{\hat{\mathcal{S}}} = \{\beta_j : j \in \hat{\mathcal{S}}\}$ by applying the orthogonal regression of y_i on the selected predictors only. Finally we have the final estimate denoted by $\hat{\boldsymbol{\beta}}_{\hat{\mathcal{S}}}$ that is defined by

$$\left(\hat{\boldsymbol{\beta}}_{\hat{\mathcal{S}}}, \hat{\mathbf{X}}_{\hat{\mathcal{S}}}\right) = \underset{\boldsymbol{\beta}_{\hat{\mathcal{S}}}, \mathbf{X}_{\hat{\mathcal{S}}}}{\operatorname{argmin}} \left(\mathbf{y} - \mathbf{X}_{\hat{\mathcal{S}}}\boldsymbol{\beta}_{\hat{\mathcal{S}}}\right)^\top \left(\mathbf{y} - \mathbf{X}_{\hat{\mathcal{S}}}\boldsymbol{\beta}_{\hat{\mathcal{S}}}\right) + \gamma \sum_{i=1}^{n} \left(\mathbf{w}_{i,\hat{\mathcal{S}}} - \mathbf{x}_{i,\hat{\mathcal{S}}}\right)^\top \boldsymbol{\Sigma}^{-1}_{\mathbf{u},\hat{\mathcal{S}}} \left(\mathbf{w}_{i,\hat{\mathcal{S}}} - \mathbf{x}_{i,\hat{\mathcal{S}}}\right),$$

where $\mathbf{w}_{i,\hat{\mathcal{S}}} = \{w_{ij} | j \in \hat{\mathcal{S}}\}$, $\mathbf{x}_{i,\hat{\mathcal{S}}} = \{x_{ij} | j \in \hat{\mathcal{S}}\}$, $\mathbf{X}_{\hat{\mathcal{S}}} = (\mathbf{x}_{1,\hat{\mathcal{S}}}, \ldots, \mathbf{x}_{n,\hat{\mathcal{S}}})^\top$, and $\boldsymbol{\Sigma}_{\mathbf{u},\hat{\mathcal{S}}}$ is the covariance matrix of $\mathbf{u}_{\hat{\mathcal{S}}} = \{u_j | j \in \hat{\mathcal{S}}\}$.

The proposed two-step method shows promising performance; however, we empirically observe that uninformative variables are often selected in the first step due to the additional variability of $\mathbf{w}_i$ compared to $\mathbf{x}_i$. We suggest a simple modification for the selection step as follows. In order to reduce such false positives in the first step. We randomly split data into two subsets with the same size and apply regularized methods on the two subsets. Then we set $\hat{\mathcal{S}} = \hat{\mathcal{S}}_1 \cap \hat{\mathcal{S}}_2$ where $\hat{\mathcal{S}}_1$ and $\hat{\mathcal{S}}_2$ denote the selected sets from the two subsets respectively. Simulation studies show that the simple modification helps reduce false positives in the selection step.

However, we also remark that this random partitioning approach may not work well when the sample size is low and/or the signal of informative variables are not strong because the signal is too weak to detect from half of the data. Therefore we recommend in practice to employ this modification only when the sample size is large enough.

4. Tuning γ: SIMEX estimator of σ^2

In the orthogonal regression (2.1), γ which is in fact σ^2 is an unknown but crucial quantity, and thus it should be estimated before applying the proposed method. Without the measurement error, we can estimate σ^2 by

$$\hat{\sigma}^2 = \frac{1}{n-p-1}\left(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}^*\right)^\top \left(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}^*\right),$$

where $\hat{\boldsymbol{\beta}}^*$ denotes the least square estimate of $\mathbf{y}$ on $\mathbf{X}$. In the presence of measurement error, we can consider $\tilde{\sigma}^2(\gamma)$, a naive estimate of error variance from the orthogonal regression as:

$$\tilde{\sigma}^2(\gamma) = \frac{1}{n-p-1}\left(\mathbf{y} - \mathbf{W}\hat{\boldsymbol{\beta}}(\gamma)\right)^\top \left(\mathbf{y} - \mathbf{W}\hat{\boldsymbol{\beta}}(\gamma)\right), \tag{4.1}$$

where $\hat{\boldsymbol{\beta}}(\gamma)$ denotes the orthogonal regression estimate and it is a function of γ. However, $\tilde{\sigma}^2(\gamma)$ in (4.1) clearly over-estimates σ^2 since $\mathbf{W}$ are contaminated. Let

$$\Delta(\gamma) = \tilde{\sigma}^2(\gamma) - \hat{\sigma}^2 \tag{4.2}$$

be the difference between the two estimates. We propose to exploit the simulation-extrapolation (SIMEX) (Cook and Stefanski, 1994) to estimate $\Delta(\gamma)$ in (4.2). For the SIMEX method, we consider independent copies of $\mathbf{u}_i$, $\mathbf{u}_i^* \overset{\text{iid}}{\sim} N(0, \boldsymbol{\Sigma}_{\mathbf{u}})$, $i = 1, \ldots, n$ and define $\mathbf{W}^* = (\mathbf{w}_1^*, \ldots, \mathbf{w}_n^*)^\top$ with $\mathbf{w}_i^* = \mathbf{x}_i + \mathbf{u}_i^*$. SIMEX is motivated because the relation between $\mathbf{X}$ and $\mathbf{W}$ is similar to that between $\mathbf{W}$ and $\mathbf{W}^*$.

In particular, we denote $\mathbf{W}_b^* = (\mathbf{w}_{1,b}^*, \ldots, \mathbf{w}_{n,b}^*)^\top$ with be $\mathbf{w}_{i,b}^* = \mathbf{x}_i + \mathbf{u}_{i,b}$ where $\mathbf{u}_{i,b}$ for $i = 1, \ldots, n$ is the independently simulated measurement error of the b^{th} iteration for $b = 1, \ldots, B$. Given γ, we obtain the solution $\hat{\boldsymbol{\beta}}_b^*(\gamma)$ as the orthogonal regression estimator of $(\mathbf{y}, \mathbf{W}_b^*)$

$$\left(\hat{\boldsymbol{\beta}}_b^*(\gamma), \hat{\mathbf{X}}\right) = \underset{\boldsymbol{\beta},\mathbf{X}}{\operatorname{argmin}} (\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^\top (\mathbf{y} - \mathbf{X}\boldsymbol{\beta}) + \gamma \sum_{i=1}^{n} \left(\mathbf{w}_{i,b}^* - \mathbf{x}_i\right)^\top \{2\boldsymbol{\Sigma}_\mathbf{u}\}^{-1} \left(\mathbf{w}_{i,b}^* - \mathbf{x}_i\right)$$

since $\mathrm{Cov}(\mathbf{w}_{i,b}^* - \mathbf{x}_i) = 2\boldsymbol{\Sigma}_\mathbf{u}$. We then compute

$$\tilde{\sigma}_b^{*2}(\gamma) = \frac{1}{n-p-1}\left(\mathbf{y} - \mathbf{W}^*\hat{\boldsymbol{\beta}}^*(\gamma)\right)^\top \left(\mathbf{y} - \mathbf{W}^*\hat{\boldsymbol{\beta}}^*(\gamma)\right), \quad b = 1, \ldots, B,$$

and finally, the SIMEX estimator of $\Delta_{\text{SIMEX}}(\gamma)$ is given by

$$\Delta_{\text{SIMEX}}(\gamma) = \frac{1}{B}\sum_{b=1}^{B} \tilde{\sigma}_b^{*2}(\gamma) - \tilde{\sigma}^2(\gamma). \tag{4.3}$$

and the estimated variance $\hat{\sigma}^2_{\text{SIMEX}}(\gamma)$ is

$$\hat{\sigma}^2_{\text{SIMEX}}(\gamma) = \tilde{\sigma}^2(\gamma) - \Delta_{\text{SIMEX}}(\gamma).$$

Finally, we can find $\hat{\gamma}$ such that $\hat{\gamma} = \operatorname{argmin}_\gamma |\hat{\sigma}^2_{\text{SIMEX}}(\gamma) - \gamma|$ via grid search.

5. Simulation

We conduct a simulation to investigate the performance of the two-step variable selection procedure and σ^2 estimation. We consider a SCAD penalty function and perform ten-fold cross validation for λ selection in the first step of the variable selection. We remark that we have two versions, the original and the modified version, of the two-step procedure method, denoted by TS1 and TS2, respectively. Competing methods considered in the simulation include the least squares regression of $\mathbf{y}$ on $\mathbf{X}$ as an oracle estimator (denoted by Oracle), the least squares regression of $\mathbf{y}$ on $\mathbf{W}$ (denoted by OLS), the orthogonal regression of $\mathbf{y}$ on $\mathbf{W}$ (denoted by OR) and the penalized least squares regression of $\mathbf{y}$ on $\mathbf{W}$ (denoted by PLS). We also consider the penalized orthogonal regression (denoted by POR) by directly adding a penalty term on (2.1), which can be viewed as a one-step procedure. For all orthogonal regressions considered here, we assume $\gamma = \sigma^2$ is known for simplicity.

Setting $(n, p) \in \{200, 500\} \times \{20\}$, we consider the following three regression models:

(M1) $y_i = x_{i16} + x_{i17} + x_{i18} + x_{i19} + x_{i20} + \epsilon_i$

(M2) $y_i = x_{i2} + x_{i7} + x_{i12} + x_{i13} + x_{i20} + \epsilon_i$

(M3) $y_i = -3.6x_{i3} - 1.5x_{i6} + x_{i11} + 2.3x_{i14} + 4x_{i20} + \epsilon_i$

where $\mathbf{x}_i \overset{\text{iid}}{\sim} N_p(0, \mathbf{I}_p)$ and $\epsilon_i \overset{\text{iid}}{\sim} N(0, 1)$. (M1) and (M2) are regarded as the models with relatively week signals while (M3) is the complex model with strong signals. When generating the measurement error $\mathbf{u}_i \overset{\text{iid}}{\sim} N_p(0, \boldsymbol{\Sigma}_\mathbf{u})$, we consider the two different structures of $\boldsymbol{\Sigma}_\mathbf{u}$ as:

- Independent (IND): $\boldsymbol{\Sigma}_\mathbf{u} = \sigma_\mathbf{u}^2\mathbf{I}_p$
- Autoregressive (AR): $\boldsymbol{\Sigma}_\mathbf{u} = \sigma_\mathbf{u}^2\mathbf{A}_p$ where $\mathbf{A}_p$ denotes a p-dimensional symmetric matrix whose $(i, j)^{th}$ element is $\rho^{|i-j|}$, $i, j = 1, \ldots, p$.

Table 1: Simulation result for independent predictors

Model	Method	$\sigma_{\mathbf{u}} = 0.5$						$\sigma_{\mathbf{u}} = 1$					
		TP		FP		MSE		TP		FP		MSE	
M1	Oracle	5.00	(0.00)	0.00	(0.00)	0.010	(0.013)	5.00	(0.00)	0.00	(0.00)	0.010	(0.013)
	OLS	5.00	(0.00)	15.00	(0.00)	0.254	(0.052)	5.00	(0.00)	15.00	(0.00)	1.301	(0.122)
	OR	5.00	(0.00)	15.00	(0.00)	0.115	(0.038)	5.00	(0.00)	15.00	(0.00)	0.507	(0.235)
	PLS	5.00	(0.00)	0.74	(1.32)	0.206	(0.051)	5.00	(0.00)	1.21	(1.72)	1.254	(0.120)
	POR	5.00	(0.00)	1.12	(2.55)	0.046	(0.043)	5.00	(0.00)	1.84	(1.74)	0.236	(0.178)
	TS1	5.00	(0.00)	0.74	(1.50)	0.037	(0.035)	5.00	(0.00)	1.21	(2.08)	0.182	(0.159)
	TS2	5.00	(0.00)	0.54	(1.04)	0.035	(0.027)	5.00	(0.00)	0.93	(1.23)	0.142	(0.130)
M2	Oracle	5.00	(0.00)	0.00	(0.00)	0.010	(0.013)	5.00	(0.00)	0.00	(0.00)	0.010	(0.013)
	OLS	5.00	(0.00)	15.00	(0.00)	0.270	(0.053)	5.00	(0.00)	15.00	(0.00)	1.351	(0.128)
	OR	5.00	(0.00)	15.00	(0.00)	0.124	(0.039)	5.00	(0.00)	15.00	(0.00)	0.560	(0.201)
	PLS	5.00	(0.00)	0.78	(1.88)	0.224	(0.051)	5.00	(0.00)	1.52	(2.00)	1.296	(0.126)
	POR	5.00	(0.00)	1.58	(2.64)	0.055	(0.048)	5.00	(0.00)	2.04	(2.38)	0.281	(0.203)
	TS1	5.00	(0.00)	0.78	(1.75)	0.037	(0.034)	5.00	(0.00)	1.52	(2.48)	0.192	(0.165)
	TS2	5.00	(0.00)	0.68	(1.36)	0.034	(0.031)	5.00	(0.00)	1.08	(2.00)	0.136	(0.147)
M3	Oracle	5.00	(0.00)	0.00	(0.00)	0.010	(0.013)	5.00	(0.00)	0.00	(0.00)	0.010	(0.013)
	OLS	5.00	(0.00)	15.00	(0.00)	1.817	(0.301)	5.00	(0.00)	15.00	(0.00)	9.922	(0.909)
	OR	5.00	(0.00)	15.00	(0.00)	0.534	(0.165)	5.00	(0.00)	15.00	(0.00)	2.695	(1.189)
	PLS	5.00	(0.00)	1.45	(1.62)	1.641	(0.314)	4.95	(0.20)	2.94	(2.55)	9.871	(0.951)
	POR	5.00	(0.00)	1.86	(1.81)	0.267	(0.173)	4.94	(0.20)	3.41	(2.68)	1.604	(1.086)
	TS1	5.00	(0.00)	1.45	(1.66)	0.236	(0.159)	4.95	(0.20)	2.94	(2.56)	1.361	(0.900)
	TS2	5.00	(0.00)	1.00	(1.44)	0.194	(0.135)	4.85	(0.34)	2.33	(2.30)	1.179	(0.988)

Averaged TP, FP, and MSE over 100 independent repetitions are reported under (M1)–(M3) where $\Sigma_{\mathbf{u}}$ takes IND structure with $n = 500$ and $p = 20$. The SCAD penalty is used for a shrinkage method. The standard errors of TP, FP, and MSE are given in parenthesis. TP = true positives; FP = false positive; MSE = median of squared error.

For $\sigma_{\mathbf{u}}$, we consider $\{0.5, 1\}$ for IND and $0.1, 0.25\}$ for AR respectively.

For performance evaluation, we report the three values with their standard errors:

- TP: averaged true positives: the # of important variables selected in the first step.
- FP: averaged false positive: the # of unimportant variables selected in the first step.
- MSE: the median of squared error $||\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}||^2$.

The first two measures, TP and FP quantify the performance of the variable selection. MSE measures the accuracy of coefficient estimates. In our simulation study, perfect methods have 5 TP, 0 FP and the lowest MSE.

Table 1 and Table 2 report the simulation results when the measurement errors have independent and autoregressive structures, respectively. It is observed that our two-step approach outperforms the POR which is a one-step approach. Comparing the two versions of the two-step approach, the modified version substantially reduces false positives. We also note that our two-step approach using LASSO and MCP performs better than the one-step approach under the all scenarios in considered.

We next conduct additional simulations to examine the performance of our method to estimate $\gamma = \sigma^2$, the variance of error. We investigate the case where the data is generated from (M1) having the independent form $\Sigma_{\mathbf{u}}$ with $\sigma_{\mathbf{u}} \times n \times p = \{0.5, 1\} \times 1000 \times 20$. While $\sigma_{\mathbf{u}}$ is fixed, we vary the true values of $\gamma = \sigma^2$ between 0.5 to 1.5 to show the performance of the proposed method with $B = 100$.

Table 3 describes the average estimated γ and its standard error over 100 repetitions. As the true γ is increased, the proposed method performs well on average even though the noise strength $\sigma_{\mathbf{u}}$ is strong. High averages with low standard error shows the consistency of the estimator and the accuracy

Table 2: Simulation result for autoregressive predictors

Model	Method	$\sigma_{\mathbf{u}} = 0.5$						$\sigma_{\mathbf{u}} = 1$					
		TP		FP		MSE		TP		FP		MSE	
M1	Oracle	5.00	(0.00)	0.00	(0.00)	0.011	(0.013)	5.00	(0.00)	0.00	(0.00)	0.011	(0.013)
	OLS	5.00	(0.00)	15.00	(0.00)	0.048	(0.014)	5.00	(0.00)	15.00	(0.00)	0.134	(0.034)
	OR	5.00	(0.00)	15.00	(0.00)	0.047	(0.014)	5.00	(0.00)	15.00	(0.00)	0.078	(0.023)
	PLS	5.00	(0.00)	0.85	(1.42)	0.013	(0.010)	5.00	(0.00)	0.86	(2.11)	0.085	(0.035)
	POR	5.00	(0.00)	0.50	(2.12)	0.012	(0.013)	5.00	(0.00)	1.21	(2.16)	0.025	(0.023)
	TS1	5.00	(0.00)	0.85	(1.50)	0.015	(0.015)	5.00	(0.00)	0.86	(1.72)	0.025	(0.024)
	TS2	5.00	(0.00)	0.65	(1.44)	0.014	(0.011)	5.00	(0.00)	0.65	(2.04)	0.023	(0.021)
M2	Oracle	5.00	(0.00)	0.00	(0.00)	0.010	(0.013)	5.00	(0.00)	0.00	(0.00)	0.010	(0.013)
	OLS	5.00	(0.00)	15.00	(0.00)	0.045	(0.014)	5.00	(0.00)	15.00	(0.00)	0.086	(0.026)
	OR	5.00	(0.00)	15.00	(0.00)	0.046	(0.014)	5.00	(0.00)	15.00	(0.00)	0.065	(0.020)
	PLS	5.00	(0.00)	1.08	(1.57)	0.014	(0.010)	5.00	(0.00)	1.73	(2.92)	0.045	(0.026)
	POR	5.00	(0.00)	0.47	(1.61)	0.012	(0.009)	5.00	(0.00)	1.17	(2.68)	0.018	(0.018)
	TS1	5.00	(0.00)	1.08	(1.66)	0.017	(0.016)	5.00	(0.00)	1.73	(2.78)	0.023	(0.020)
	TS2	5.00	(0.00)	0.74	(1.46)	0.015	(0.013)	5.00	(0.00)	1.44	(2.11)	0.022	(0.019)
M3	Oracle	5.00	(0.00)	0.00	(0.00)	0.010	(0.013)	5.00	(0.00)	0.00	(0.00)	0.010	(0.013)
	OLS	5.00	(0.00)	15.00	(0.00)	0.064	(0.019)	5.00	(0.00)	15.00	(0.00)	0.343	(0.074)
	OR	5.00	(0.00)	15.00	(0.00)	0.059	(0.017)	5.00	(0.00)	15.00	(0.00)	0.158	(0.044)
	PLS	5.00	(0.00)	1.00	(1.74)	0.022	(0.015)	5.00	(0.00)	4.27	(4.01)	0.223	(0.089)
	POR	5.00	(0.00)	1.22	(2.13)	0.016	(0.015)	5.00	(0.00)	3.46	(3.52)	0.094	(0.055)
	TS1	5.00	(0.00)	1.00	(1.71)	0.021	(0.020)	5.00	(0.00)	4.27	(4.24)	0.083	(0.050)
	TS2	5.00	(0.00)	0.77	(1.62)	0.019	(0.016)	5.00	(0.00)	3.43	(3.64)	0.078	(0.047)

Averaged TP, FP, and MSE over 100 independent repetitions are reported under (M1)–(M3) where $\Sigma_{\mathbf{u}}$ takes AR structure with $n = 500$, $p = 20$, and $\rho = 0.5$. The SCAD penalty is used for a shrinkage method. The standard errors of TP, FP, and MSE are given in parenthesis. TP = true positives; FP = false positive; MSE = median of squared error.

Table 3: Simulation result for γ estimation

γ	$\sigma_{\mathbf{u}} = 0.5$		$\sigma_{\mathbf{u}} = 1$	
	AVER	SE	AVER	SE
0.50	0.512	(0.090)	0.638	(0.181)
0.70	0.717	(0.093)	0.834	(0.222)
1.00	1.016	(0.110)	1.118	(0.255)
1.30	1.303	(0.119)	1.409	(0.269)
1.50	1.526	(0.129)	1.600	(0.266)

The averaged estimated γ and its standard error over 100 repetitions are reported under (M1) with IND structure, $n = 1000$ and $p = 20$.

of our algorithm. Thus, the proposed estimation method based on the SIMEX idea shows promising performance to select a proper γ.

6. Real data illustration

For the real data illustration, we use the Boston housing data (Harrison and Rubinfeld, 1978) available in R. The response is the logarithm of the median value of owner-occupied homes in the Boston areas. The data originally contains thirteen predictors which are not contaminated. In order to check the performance of the proposed method under the presence of measurement error, we first exclude two discrete predictors and marginally standardize the eleven continuous predictors. We then generate $(p - 11)$ noise variables from the standard normal distribution so that we have p predictors in total where $p \in \{20, 30\}$. Finally, the measurement errors from the normal distribution with mean 0 and variance 0.5 which we assume to be known are generated and added to all predictors. We apply the

Table 4: Real-data-based comparison results

p	Penalty	PLS		TS		p-value
	LASSO	0.742	(0.065)	0.637	(0.124)	0.000
20	SCAD	0.698	(0.081)	0.658	(0.176)	0.013
	MCP	0.696	(0.081)	0.686	(0.220)	0.333
	LASSO	0.644	(0.049)	0.557	(0.076)	0.000
30	SCAD	0.607	(0.078)	0.584	(0.104)	0.000
	MCP	0.606	(0.077)	0.591	(0.105)	0.004

Averaged root mean square error over 100 independent repetitions along with the corresponding standard deviations. The last column contains the p-values for the (one-sided) pairwise t-test between the RMSEs of PLS and TS.

proposed two-step procedure (TS) to this artificially contaminated data and compare its performance to the penalized least regression of the response on the contaminated predictors. We repeat these steps 100 times independently; consequently, Table 4 reports the averaged root mean squared error (RMSE) of regression coefficients over the 100 independent repetitions. The last column reports the p-values for the pairwise mean difference between the RMSEs of TS and PLS. When computing the RMSE we treat the OLS estimators based on the eleven predictors before the contamination as true parameter values for informative predictors and 0 for all noise variables. It is observed that the proposed TS outperforms PLS that ignores measurement error which is concordant to what we have seen in Section 4.

7. Conclusion

In this paper, we develop a two-step variable selection method for measurement error models. The proposed method is based on the idea of separating selection and estimation; it first selects significant variables from contaminated covariates and then obtains the orthogonal regression estimates of the selected variables. Furthermore, we suggested a SIMEX-based method to estimate γ which is unknown in practice. Our simulation results illustrate that the proposed method works well under various scenarios with different types of variance-covariance matrices. A tactic assumption in both the variable section methods and the SIMEX method is that $\mathbf{\Sigma}_\mathbf{u}$ is known. The estimation of the error covariance matrix is still challenging in error-in-variable models; in addition, most research assumes that $\mathbf{\Sigma}_\mathbf{u}$ is diagonal such as $\mathbf{\Sigma}_\mathbf{u} = \sigma_\mathbf{u}^2\mathbf{I}_p$ Amemiya and Fuller (1984) and $\sigma_\mathbf{u}$ can be estimated from repeated measurements. Despite the drawback, the proposed methods are valuable for improving estimation in the measurement errors model; consequently, the estimation of $\mathbf{\Sigma}_\mathbf{u}$ remains a topic for future work.

Acknowledgement

This work is supported by National Research Foundation of Korea Grant (NRF-2018R1D1A1B07043 034).

References

Amemiya Y and Fuller WA (1984). Estimation for the multivariate errors-in-variables model with estimated error covariance matrix, *The Annals of Statistics*, **12**, 497–509.

Carroll RJ, Ruppert D, Stefanski LA, and Crainiceanu C (2006). *Measurement Error in Nonlinear Models: A Modern Perspective*, CRC Press.

Cook JR and Stefanski LA (1994). Simulation-extrapolation estimation in parametric measurement error models, *Journal of the American Statistical Association*, **89**, 1314–1328.

Efron B, Hastie T, Johnstone I, and Tibshirani R (2004). Least angle regression, *The Annals of Statistics*, **32**, 407–499.

Fan J and Li R (2001). Variable selection via nonconcave penalized likelihood and its oracle properties, *Journal of the American statistical Association*, **96**, 1348–1360.

Fuller WA (1987). *Measurement Error Models*, John Willey, New York.

Harrison D and Rubinfeld DL (1978). Hedonic housing prices and the demand for clean air, *Journal of Environmental Economics and Management*, **5**, 81–102.

Liang H and Li R (2009). Variable selection for partially linear models with measurement errors, *Journal of the American Statistical Association*, **104**, 234–248.

Ma Y and Li R (2010). Variable selection in measurement error models, *Bernoulli: official journal of the Bernoulli Society for Mathematical Statistics and Probability*, **16**, 274.

Meinshausen N (2007). Relaxed lasso, *Computational Statistics & Data Analysis*, **52**, 374–393.

Tibshirani R (1996). Regression shrinkage and selection via the lasso, *Journal of the Royal Statistical Society. Series B (Methodological)*, **58**, 267–288.

Zhang CH (2010). Nearly unbiased variable selection under minimax concave penalty, *The Annals of Statistics*, **38**, 894–942.

Received October 16, 2018; Revised December 27, 2018; Accepted January 7, 2019