Adaptive LASSO for Varying-Coefficient Partially Linear Measurement Error Models

HaiYing Wang[1], Guohua Zou[2], and Alan T.K. Wan[3]

[1]Department of Statistics, University of Missouri, Columbia, Missouri 65211, U.S.A. (*Email: hwzq7@mail.missouri.edu*)
[2]MADIS, Academy of Mathematics and Systems Science, Chinese Academy of Sciences, Beijing 100190, P. R. China (*Email: ghzou@amss.ac.cn*)
[3]Department of Management Sciences, City University of Hong Kong, Kowloon, Hong Kong (*Email: Alan.Wan@cityu.edu.hk*)

Abstract

This paper extends the adaptive LASSO (ALASSO) for simultaneous parameter estimation and variable selection to a varying-coefficient partially linear model where some of the covariates are subject to measurement errors of an additive form. We draw comparisons with the SCAD, and prove that both the ALASSO and SCAD attain the oracle property under this setup. We further develop an algorithm in the spirit of LARS for finding the solution path of the ALASSO in practical applications. Finite sample properties of the proposed methods are examined in a simulation study, and a real data example based on the U.S. Department of Agriculture's Continuing Survey of Food Intakes by Individuals (CSFII) is considered.

Keywords: Adaptive LASSO, LARS, Measurement Errors, Model Selection, Oracle Property, SCAD, Semi-parametric Model

1 Introduction

Consider the following semi-parametric varying-coefficient partially linear model with additive measurement errors on some of the covariates:

$$\begin{cases} Y &= \mathbf{X}^\top \boldsymbol{\beta} + \mathbf{Z}^\top \boldsymbol{\alpha}(T) + \varepsilon, \\ \mathbf{W} &= \mathbf{X} + \mathbf{U}, \\ \boldsymbol{\zeta} &= \mathbf{Z} + \mathbf{V}, \end{cases} \tag{1}$$

where Y is a univariate response variable; $\mathbf{X}$ and $\mathbf{Z}$ are $d\times 1$ and $r\times 1$ covariate vectors respectively; $\boldsymbol{\beta}$ is a d-dimensional unknown parameter vector associated with $\mathbf{X}$; $\boldsymbol{\alpha}(\cdot) = (\alpha_1(\cdot), ..., \alpha_r(\cdot))^\top$ is a r-dimensional unknown function vector associated with $\mathbf{Z}$; ε is a disturbance term with mean 0 and variance σ^2, and $\mathbf{U}$ and $\mathbf{V}$, which have mean 0 and variance-covariance matrices Σ_u and Σ_v respectively, represent the measurement errors associated with $\mathbf{W}$ and $\boldsymbol{\zeta}$, the proxies for the unobserved $\mathbf{X}$ and $\mathbf{Z}$. It is assumed for simplicity that T is univariate; $(\mathbf{X}^\top, \mathbf{Z}^\top, T)$, $\mathbf{U}$, $\mathbf{V}$ and ε are mutually independent, and Σ_u and Σ_v are known. The (more realistic) case where Σ_u and Σ_v are unknown will be taken up later in the paper. We refer to model (1) as the varying-coefficient

partially linear measurement error (VCPLE) model. Clearly, when $\mathbf{U} \equiv 0$ and $\mathbf{V} \equiv 0$, the VCPLE model reduces to the well-known varying-coefficient partially linear (VCPL) model. The main attraction of the VCPL model is that it allows T and $\mathbf{Z}$ to interact in a flexible way such that each different level of T is associated with a different linear model. Recent papers on the VCPL model emphasize the development of estimation procedures; e.g., Zhang *et al.* (2002), Fan and Huang (2005), You and Zhou (2006) and Huang and Zhang (2009).

There is a long standing literature on statistical modeling subject to measurement errors. More recently, attention has focused on refinements to various semi-parametric estimation methods in the face of measurement errors. Liang *et al.* (1999) applied the so-called "correction for attenuation" to the semi-parametric partially linear model in the context of measurement errors, and derived the asymptotic properties of the resultant estimator. Liang (2000) demonstrated the asymptotic normality of the estimator for the parametric component in a partially linear model when variables in the non-parametric component are measured with errors. You *et al.* (2006) proposed a corrected local polynomial estimator for the varying-coefficient model when the covariates are measured with errors. You and Chen (2006) modified the estimation method of Fan and Huang (2005) for the VCPL model to account for measurement errors in the covariates of the parametric part. Liang and Li (2009) considered the problem of variable selection in a partially linear model based on the SCAD penalty function (Fan and Li, 2001), and established the oracle property of the proposed estimator. Ma and Li (2009) studied the variable selection problem for the general non-linear and a class of semi-parametric models under measurement errors. Other studies on semi-parametric modeling involving errors-in-variables include Tsiatis and Ma (2004), Ma and Carroll (2006), Hall and Ma (2007), Liang *et al.* (2007), among others.

The current paper proposes a unified estimation and variable selection method for the VCPLE model in the spirit of the adaptive LASSO (ALASSO) developed by Zou (2006) and Zhang and Lu (2007). The original LASSO ("least absolute shrinkage and selection operator"), introduced by Tibshirani (1996), is a technique for simultaneous parameter estimation and variable selection based on the penalized least-squares method. It is a variant of the Bridge (Frank and Friedman, 1993), the Smoothly Clipped Absolute Deviation (SCAD) (Fan and Li, 2001), and the Least Angle Regression Selection (LARS) (Efron *et al.*, 2004) estimators. One characteristic of the LASSO is that it shrinks some of the coefficients to exactly zero, and in doing so, reduces the estimation variance while providing an interpretable final model. The LASSO technique has found widespread applications in many fields of science. While the LASSO is known to be near mini-max optimal as well as consistent under certain regularity conditions, Zou (2006) showed that it falls short of attaining the oracle property (Fan and Li, 2001, 2002, 2004; Fan and Peng, 2004; Cai *et al.*, 2005). By this latter property, an estimator estimates a zero coefficient exactly as zero with probability approaching one, while still being asymptotically normal for the non-zero coefficients in large samples. In this respect, the LASSO is inferior to other competing methods like the SCAD estimator which possesses the oracle property. To reconcile this shortcoming of the LASSO, Zou (2006) and Zhang and Lu (2007) independently developed the ALASSO, which uses adaptive weights for penalizing different coefficients in the ℓ_1 penalty. This is unlike the original LASSO that uses the same penalty for all the coefficients. Zou (2006) and Zhang and Lu (2007) demonstrated that the ALASSO possesses the aforementioned oracle property with optimal convergence rate in addition to enjoying all the good properties of the LASSO including near mini-max optimality. Compared to the SCAD estimator which has a drawback in that its penalty function is non-convex, the ALASSO has the advantage of having a convex penalty form which guarantees the existence of a unique solution. For the linear regression model, Zou (2006) showed that the ALASSO estimates can be readily calculated using the LARS algorithm (Efron *et al.*, 2004). Generalizations of the LASSO and ALASSO to cases of variable selection by groups rather than individually were made

by Yuan and Lin (2006) and Wang and Leng (2008).

In the context of the VCPL model, Zhao and Xue (2009) tackled the variable selection problem in the parametric component of the model by the SCAD method. Zhao and Xue (2010) applied the group version of the ALASSO developed by Wang and Leng (2008) for variable selection in the VCPLE model. More recently, Zhao and Xue (2011) considered a VCPLE model, a special case of model (1), in which the covariates of the non-parametric part are assumed to be free of measurement errors.

In this paper, a modified version of the LARS algorithm is proposed to obtain the solutions for the target function. This makes it possible to provide the entire solution paths of the coefficients corresponding to all tuning parameters, whereas with other algorithms such as the quadratic approximation, the solution pertains only to the tuning parameter specified at the outset. Another attraction of the LARS is that it gives the exact minimum of the target function, whereas the quadratic approximation only results in an approximate minimum. Yet in spite of these merits, LARS does not require enormous amounts of computing power to execute; for the linear model, LARS is no more intricate computationally than an ordinary least-squares fit to the full model (Efron *et al.*, 2004). In the context of the linear model, Zou (2006) used LARS to obtain the solution path of the ALASSO. To the best of our knowledge, the implementation of LARS coupled with a LASSO-type penalty function has not been explored when the covariates cannot be observed precisely. One purpose of the present paper is to take some steps in this direction by modifying the existing LARS algorithm to cater for the special features of the VCPLE model. We find that the modified algorithm performs well for both variable selection and parameter estimation. We also examine the asymptotic properties of the SCAD estimator in the context of the VCPLE model given by (1). Our results show that SCAD retains the oracle property under this set-up.

The remainder of the paper is organized as follows. In Section 2, we discuss the estimation method and the ALASSO penalty function. In Section 3, in addition to providing the main theoretical results, we also describe the modified LARS algorithm. Results of simulation experiments designed to investigate the small sample properties of the method along with an example based on real data are contained in Section 4. Section 5 concludes, and proofs of technical results are given in two appendices.

2 Estimation method and the ALASSO penalty

We will consider parameter estimation and variable selection within the framework of profile least-squares estimation (Fan and Huang, 2005; You and Chen, 2006). To motivate discussion, assume temporarily that there are no measurement errors and we observe i.i.d. samples of $\{\mathbf{Y}_i, \mathbf{X}_i, \mathbf{Z}_i, T_i\}$, $i = 1, ..., n$. For any t in the neighbourhood of t_0, let $\alpha_j(t)$ be approximated by the following linear function:

$$\alpha_j(t) \approx \alpha_j(t_0) + \alpha_j'(t_0)(t - t_0) \equiv a_j + b_j(t - t_0),\ j = 1, 2, ..., r.$$

If $\boldsymbol{\beta}$ is known, then we can obtain solutions to a_j and b_j by solving the following weighted local least-squares problem:

$$\min_{a_1,...,a_r,b_1,...b_r} \sum_{i=1}^{n} \left\{ Y_i - \mathbf{X}_i^\top \boldsymbol{\beta} - \sum_{j=1}^{r} Z_{ij}\left[a_j + b_j(T_i - t_0)\right] \right\}^2 K_h(T_i - t_0),$$

where $K_h(\cdot) = K(\cdot/h)/h$, $K(\cdot)$ is a kernel function and h is a bandwidth. The solution is given by

$$\left(\breve{a}_1(t), ..., \breve{a}_r(t), h\breve{b}_1(t), ..., h\breve{b}_r(t)\right)^\top = \left((D_t^Z)^\top \Omega_t D_t^Z\right)^{-1} (D_t^Z)^\top \Omega_t (\mathbf{Y} - \mathcal{X}\beta),$$

where $\mathbf{Y} = (Y_1, ..., Y_n)^\top$, $\mathcal{X} = (\mathbf{X}_1, ..., \mathbf{X}_n)^\top$,

$$\Omega_t = \text{diag}\{K_h(T_1 - t), ..., K_h(T_n - t)\} \text{ and } D_t^Z = \begin{pmatrix} \mathbf{Z}_1^\top & \frac{T_1 - t}{h}\mathbf{Z}_1^\top \\ \vdots & \vdots \\ \mathbf{Z}_n^\top & \frac{T_n - t}{h}\mathbf{Z}_n^\top \end{pmatrix}_{n \times 2r}$$

Substituting $(\breve{a}_1(t), ..., \breve{a}_r(t))^\top$ into model (1), we obtain

$$Y_i - \breve{Y}_i = (\mathbf{X}_i - \breve{\mathbf{X}}_i)^\top \boldsymbol{\beta} + \varepsilon_i,$$

where $\breve{Y}_i = (\mathbf{Z}_i^\top,\ 0)\left[(D_{t_i}^Z)^\top \Omega_{t_i} D_{t_i}^Z\right]^{-1}(D_{t_i}^Z)^\top \Omega_{t_i}\mathbf{Y}$, and $\breve{\mathbf{X}}_i = \left\{(\mathbf{Z}_i^\top,\ 0)\left[(D_{t_i}^Z)^\top \Omega_{t_i} D_{t_i}^Z\right]^{-1}(D_{t_i}^Z)^\top \Omega_{t_i}\mathcal{X}\right\}^\top$. Then $\boldsymbol{\beta}$ in the above regression can be estimated, as in Fan and Huang (2005), by the least-squares estimator $\breve{\boldsymbol{\beta}}_{\text{LS}} = \left\{\sum_{i=1}^n (\mathbf{X}_i - \breve{\mathbf{X}}_i)^{\otimes 2}\right\}^{-1}\left\{\sum_{i=1}^n (\mathbf{X}_i - \breve{\mathbf{X}}_i)(Y_i - \breve{Y}_i)\right\}$, where $M^{\otimes 2} = MM^\top$.

When the covariates are subject to measurement errors such that $\mathbf{X}_i$'s are unobserved and replaced by the surrogates $\mathbf{W}_i$'s defined above, You and Chen (2006) proposed the following modified least-squares estimator for estimating β:

$$\breve{\boldsymbol{\beta}}_{\text{MLS}} = \arg\min_{\boldsymbol{\beta}}\left\{\sum_{i=1}^n \left[Y_i - \breve{Y}_i - (\mathbf{W}_i - \breve{\mathbf{W}}_i)^\top \boldsymbol{\beta}\right]^2 - n\boldsymbol{\beta}^\top \Sigma_u \boldsymbol{\beta}\right\},$$

where $\breve{\mathbf{W}}_i = \breve{\mathbf{X}}_i + \breve{\mathbf{U}}_i$, $\breve{\mathbf{U}}_i = \left\{(\mathbf{Z}_i^\top,\ 0)\left[(D_{t_i}^Z)^\top \Omega_{t_i} D_{t_i}^Z\right]^{-1}(D_{t_i}^Z)^\top \Omega_{t_i}\mathcal{U}\right\}^\top$, and $\mathcal{U} = (\mathbf{U}_1, ..., \mathbf{U}_n)^\top$.

Now, when $\mathbf{Z}_i$'s are also subject to measurement errors and $\boldsymbol{\zeta}_i$'s are used instead, we propose to modify $\breve{\boldsymbol{\beta}}_{\text{MLS}}$ to

$$\hat{\boldsymbol{\beta}}_{\text{MLS}} = \arg\min_{\boldsymbol{\beta}}\left\{\sum_{i=1}^n \left[Y_i - \hat{Y}_i - (\mathbf{W}_i - \hat{\mathbf{W}}_i)^\top \boldsymbol{\beta}\right]^2 - n\boldsymbol{\beta}^\top \Sigma_u \boldsymbol{\beta}\right\}, \tag{2}$$

where $\hat{Y}_i = \psi_i \mathbf{Y}$, $\psi_i = (\boldsymbol{\zeta}_i^\top,\ 0)\left[(D_{t_i}^\zeta)^\top \Omega_{t_i} D_{t_i}^\zeta - \phi_{t_i}\right]^{-1}(D_{t_i}^\zeta)^\top \Omega_{t_i}$, $\hat{\mathbf{W}}_i = \hat{\mathbf{X}}_i + \hat{\mathbf{U}}_i$, $\hat{\mathbf{X}}_i = \{\psi_i \mathcal{X}\}^\top$, $\hat{\mathbf{U}}_i = \{\psi_i \mathcal{U}\}^\top$,

$$D_{t_i}^\zeta = \begin{pmatrix} \boldsymbol{\zeta}_1^\top & \frac{T_1 - t_i}{h}\boldsymbol{\zeta}_1^\top \\ \vdots & \vdots \\ \boldsymbol{\zeta}_n^\top & \frac{T_n - t_i}{h}\boldsymbol{\zeta}_n^\top \end{pmatrix} \text{ and } \phi_{t_i} = \sum_{j=1}^n \begin{pmatrix} 1 & \frac{T_j - t_i}{h} \\ \frac{T_j - t_i}{h} & \frac{(T_j - t_i)^2}{h^2} \end{pmatrix} \otimes \Sigma_v K_h(T_j - t_i).$$

The term ϕ_t is a correction term suggested by You *et al.* (2006) for the the varying-coefficient model under measurement errors. It has the purpose of correcting the bias introduced by measurement errors. You *et al.* (2006) showed that the estimator of the unknown function under their model setup is inconsistent if this term is dropped.

Now, by incorporating the ℓ_1 penalty in the objective function in (2), we obtain the LASSO estimator of $\boldsymbol{\beta}$, defined as:

$$\hat{\boldsymbol{\beta}}_{\text{LASSO}} = \arg\min_{\boldsymbol{\beta}}\left\{\sum_{i=1}^n \left[Y_i - \hat{Y}_i - (\mathbf{W}_i - \hat{\mathbf{W}}_i)^\top \boldsymbol{\beta}\right]^2 - n\boldsymbol{\beta}^\top \Sigma_u \boldsymbol{\beta} + \lambda_n \sum_{j=1}^d |\beta_j|\right\}, \tag{3}$$

where the last term in the above equation is the ℓ_1 penalty. The purpose of this penalty is to shrink some of the coefficients to exactly zero. This makes the LASSO a simultaneous estimation

and variable selection procedure. However, as noted by Zou (2006), because the ℓ_1 penalty forces the coefficients to be equally penalized, no estimator based on the LASSO can attain the oracle property. To reconcile this difficulty, Zou (2006) introduced the ALASSO, and proved under a linear model setup that it possesses the oracle property. Here, we adopt Zou's (2006) idea, and propose the following ALASSO estimator under the VCPLE model setup discussed above:

$$\hat{\boldsymbol{\beta}} = \arg\min_{\boldsymbol{\beta}} \left\{ \sum_{i=1}^{n} \left[Y_i - \hat{Y}_i - (\mathbf{W}_i - \hat{\mathbf{W}}_i)^\top \boldsymbol{\beta} \right]^2 - n\boldsymbol{\beta}^\top \Sigma_u \boldsymbol{\beta} + \lambda_n \sum_{j=1}^{d} \frac{|\beta_j|}{|\hat{\beta}_j^*|^\gamma} \right\}, \tag{4}$$

where $\hat{\boldsymbol{\beta}}^*$ is a consistent estimator of $\boldsymbol{\beta}$, and $\gamma > 0$ is a constant - large values of γ generally result in sparse models, and vice versa; Zhang and Lu (2007) fixed γ to 1, but in general, this parameter may be chosen by cross-validation. One possible choice for $\hat{\boldsymbol{\beta}}^*$ is the consistent estimator $\hat{\boldsymbol{\beta}}_{\text{MLS}}$ in equation (2). The implementation of $\hat{\boldsymbol{\beta}}$ also requires estimates of the unknown Σ_u and Σ_v. To estimate these matrices, it is useful to assume that there exist partially replicated observations such that $\mathbf{W}_{ij} = \mathbf{X}_i + \mathbf{U}_{ij}$ and $\boldsymbol{\zeta}_{il} = \mathbf{Z}_i + \mathbf{V}_{il}$ are observed, $j = 1, ..., J$, $l = 1, ..., L$, $i = 1, ..., n$ (Carroll *et al.*, 2006; Liang and Li, 2009). Then

$$\hat{\Sigma}_u = \frac{1}{n(J-1)} \sum_{i=1}^{n} \sum_{j=1}^{J} (\mathbf{W}_{ij} - \bar{\mathbf{W}}_i)^{\otimes 2}$$

and

$$\hat{\Sigma}_v = \frac{1}{n(L-1)} \sum_{i=1}^{n} \sum_{l=1}^{L} (\boldsymbol{\zeta}_{il} - \bar{\boldsymbol{\zeta}}_i)^{\otimes 2}$$

are consistent and unbiased estimators of Σ_u and Σ_v respectively, where $\bar{\mathbf{W}}_i = \sum_{j=1}^{J} \mathbf{W}_{ij}/J$ and $\bar{\boldsymbol{\zeta}}_i = \sum_{l=1}^{L} \mathbf{W}_{il}/L$. Now, denote $\bar{\mathbf{U}}_i = \sum_{j=1}^{J} \mathbf{U}_{ij}/J$ and $\bar{\mathbf{V}}_i = \sum_{l=1}^{L} \mathbf{V}_{il}/L$, then model (1) is modified to

$$\begin{cases} Y_i & = \mathbf{X}_i^\top \boldsymbol{\beta} + \mathbf{Z}_i^\top \boldsymbol{\alpha}(T_i) + \varepsilon_i, \\ \bar{\mathbf{W}}_i & = \mathbf{X}_i + \bar{\mathbf{U}}_i \\ \bar{\boldsymbol{\zeta}}_i & = \mathbf{Z}_i + \bar{\mathbf{V}}_i. \end{cases} \tag{5}$$

Correspondingly, the ALASSO estimator given in (4) is modified to

$$\hat{\boldsymbol{\beta}} = \arg\min_{\boldsymbol{\beta}} \left\{ \sum_{i=1}^{n} \left[\left(Y_i - \hat{\bar{Y}}_i - (\bar{\mathbf{W}}_i - \hat{\bar{\mathbf{W}}}_i)^\top \boldsymbol{\beta} \right)^2 - \frac{1}{J} \boldsymbol{\beta}^\top \hat{\Sigma}_u \boldsymbol{\beta} \right] + \lambda_n \sum_{j=1}^{d} \frac{|\beta_j|}{|\hat{\beta}_j^*|^\gamma} \right\}, \tag{6}$$

where $\hat{\bar{Y}}_i = \bar{\psi}_i \mathbf{Y}$, $\hat{\bar{\mathbf{W}}}_i = \{\bar{\psi}_i \bar{\mathcal{W}}\}^\top$, $\bar{\psi}_i = (\bar{\boldsymbol{\zeta}}_i^\top,\ 0) \left[(D_{t_i}^{\bar{\zeta}})^\top \Omega_{t_i} D_{t_i}^{\bar{\zeta}} - \bar{\phi}_{t_i} \right]^{-1} (D_{t_i}^{\bar{\zeta}})^\top \Omega_{t_i}$, $\bar{\mathcal{W}} = (\bar{\mathbf{W}}_1, ..., \bar{\mathbf{W}}_n)$,

$$D_{t_i}^{\bar{\zeta}} = \begin{pmatrix} \bar{\boldsymbol{\zeta}}_1^\top & \frac{T_1 - t_i}{h} \bar{\boldsymbol{\zeta}}_1^\top \\ \vdots & \vdots \\ \bar{\boldsymbol{\zeta}}_n^\top & \frac{T_n - t_i}{h} \bar{\boldsymbol{\zeta}}_n^\top \end{pmatrix} \text{ and } \bar{\phi}_{t_i} = \frac{1}{L} \sum_{j=1}^{n} \begin{pmatrix} 1 & \frac{T_j - t_i}{h} \\ \frac{T_j - t_i}{h} & \frac{(T_j - t_i)^2}{h^2} \end{pmatrix} \otimes \hat{\Sigma}_v K_h(T_j - t_i).$$

3 Main results and a modified LARS algorithm

The purpose of this section is three-fold. We will first present the key theoretical properties of the ALASSO estimators in (4) and (6). This is followed by the development of a method for constructing standard errors of the ALASSO estimates. Finally, we will discuss a LARS-type algorithm for computing the ALASSO estimates in practice.

3.1 Oracle property and standard errors construction

Without loss of generality, let the true value of $\boldsymbol{\beta}$ be $\boldsymbol{\beta}_0 = (\boldsymbol{\beta}_{10}^\top, \boldsymbol{\beta}_{20}^\top)^\top$, where $\boldsymbol{\beta}_{10}$ and $\boldsymbol{\beta}_{20}$ are non-zero and zero vectors of dimensions s and $d-s$ respectively. Let $\mathbf{W}^{(1)}$, $\mathbf{X}^{(1)}$ and $\mathbf{U}^{(1)}$ be the upper $s \times 1$ sub-vectors of $\mathbf{W}$, $\mathbf{X}$ and $\mathbf{U}$ respectively, and $\Sigma_u^{(11)}$ be the $s \times s$ upper-left sub-matrix of Σ_u that corresponds to $\mathbf{U}^{(1)}$.

Theorem 1. *Assume that Assumptions 1 - 6 in Appendix A hold. If $\lambda_n/\sqrt{n} \to 0$, and there exists a sequence $d_n \to \infty$ such that $d_n(\hat{\boldsymbol{\beta}}^* - \boldsymbol{\beta}_0) = O_P(1)$ and $d_n^\gamma \lambda_n/\sqrt{n} \to \infty$ as $n \to \infty$, then with probability approaching 1, the ALASSO estimator $\hat{\boldsymbol{\beta}}$ satisfies the following properties:*

(a) Sparsity, i.e., $\hat{\boldsymbol{\beta}}_2 = \mathbf{0}$, where $\hat{\boldsymbol{\beta}}_2$ is the estimator of $\boldsymbol{\beta}_{20}$.

(b) Asymptotic normality, i.e.,

$$\sqrt{n}(\hat{\boldsymbol{\beta}}_1 - \boldsymbol{\beta}_{10}) \xrightarrow{d} N\Big(0,\ (B^{(11)})^{-1} F^{(11)} (B^{(11)})^{-1}\Big), \tag{7}$$

where $\hat{\boldsymbol{\beta}}_1$ is the estimator of $\boldsymbol{\beta}_{10}$,

$$\begin{aligned} B^{(11)} =& \mathbf{E}\big[\mathbf{X}^{(1)}(\mathbf{X}^{(1)})^\top\big] - \mathbf{E}\left[\mathbf{E}(\mathbf{X}^{(1)}\mathbf{Z}^\top|T)[\mathbf{E}(\mathbf{Z}\mathbf{Z}^\top|T)]^{-1}\mathbf{E}(\mathbf{X}^{(1)}\mathbf{Z}^\top|T)^\top\right] \\ &+ \mathbf{E}\left[\mathbf{E}(\mathbf{X}^{(1)}\mathbf{Z}^\top|T)[\mathbf{E}(\mathbf{Z}\mathbf{Z}^\top|T)]^{-1}\Sigma_v[\mathbf{E}(\mathbf{Z}\mathbf{Z}^\top|T)]^{-1}\mathbf{E}(\mathbf{X}^{(1)}\mathbf{Z}^\top|T)^\top\right], \text{and} \\ F^{(11)} =& \mathbf{E}\left(\left[\mathbf{W}^{(1)} - \mathbf{E}(\mathbf{X}^{(1)}\mathbf{Z}^\top|T)[\mathbf{E}(\mathbf{Z}\mathbf{Z}^\top|T)]^{-1}\boldsymbol{\zeta}\right](\varepsilon - \mathbf{U}^\top\boldsymbol{\beta}_0) + \Sigma_u^{(11)}\boldsymbol{\beta}_{10}\right)^{\otimes 2}, \end{aligned}$$

if Σ_u and Σ_v are known, or

$$\begin{aligned} B^{(11)} =& \mathbf{E}\big[\mathbf{X}^{(1)}(\mathbf{X}^{(1)})^\top\big] - \mathbf{E}\left[\mathbf{E}(\mathbf{X}^{(1)}\mathbf{Z}^\top|T)[\mathbf{E}(\mathbf{Z}\mathbf{Z}^\top|T)]^{-1}\mathbf{E}(\mathbf{X}^{(1)}\mathbf{Z}^\top|T)^\top\right] \\ &+ \frac{1}{L}\mathbf{E}\left[\mathbf{E}(\mathbf{X}^{(1)}\mathbf{Z}^\top|T)[\mathbf{E}(\mathbf{Z}\mathbf{Z}^\top|T)]^{-1}\Sigma_v[\mathbf{E}(\mathbf{Z}\mathbf{Z}^\top|T)]^{-1}\mathbf{E}(\mathbf{X}^{(1)}\mathbf{Z}^\top|T)^\top\right], \text{and} \\ F^{(11)} =& \mathbf{E}\left(\left[\bar{\mathbf{W}}_i^{(1)} - \mathbf{E}(\mathbf{X}_i^{(1)}\mathbf{Z}_i^\top|T)[\mathbf{E}(\mathbf{Z}_i\mathbf{Z}_i^\top|T)]^{-1}\boldsymbol{\zeta}\right](\varepsilon_i - \bar{\mathbf{U}}_i^\top\boldsymbol{\beta}_0) + \frac{\sum_{j=1}^J (\mathbf{W}_{ij}^{(1)} - \bar{\mathbf{W}}_i^{(1)})^{\otimes 2}\boldsymbol{\beta}_{10}}{J(J-1)}\right)^{\otimes 2}, \end{aligned}$$

if Σ_u and Σ_v are unknown.

In the special case where the measurement errors are symmetrically distributed, expression $F^{(11)}$ may be simplified to

$$\left(\sigma^2 + \boldsymbol{\beta}_0^\top\Sigma_u\boldsymbol{\beta}_0\right)B^{(11)} + \sigma^2\Sigma_u^{(11)} + \mathbf{E}\Big(\big[\big(\mathbf{U}^{(1)}\big)^{\otimes 2} - \Sigma_u^{(11)}\big]\boldsymbol{\beta}_{10}\Big)^{\otimes 2},$$

if Σ_u and Σ_v are known.

By Theorem 1, the ALASSO estimator estimates a zero coefficient exactly as zero with probability that tends to 1, as well as being $\sqrt{n}-$consistent for the non-zero coefficients in large samples. Also, the estimators of the non-zero coefficients have the same asymptotic variance-covariance matrix as when the true model is known. The proof of Theorem 1 is contained in Appendix A. In Appendix B, we will show that the SCAD estimator when applied to the VCPLE model is also an oracle procedure.

Next, we develop methods of constructing standard errors of the ALASSO estimates. Let $\hat{\boldsymbol{\beta}}^{nz}$ and $\hat{\boldsymbol{\beta}}^{z}$ be the non-zero and zero components of $\hat{\boldsymbol{\beta}}$ (note that $\hat{\boldsymbol{\beta}}^{nz}$ and $\hat{\boldsymbol{\beta}}^{z}$ are *not* necessarily the

same as $\hat{\boldsymbol{\beta}}_1$ and $\hat{\boldsymbol{\beta}}_2$ defined above). Now, if a coefficient is estimated as 0, the variance of the estimate is also is 0 (Tibshirani, 1996; Fan and Li, 2001; Zou, 2006). So, the standard errors of all elements in $\hat{\boldsymbol{\beta}}^z$ are zero. To construct standard errors for $\hat{\boldsymbol{\beta}}^{nz}$, we follow Fan and Li's (2001) and Zou's (2006) approaches of approximating the ALASSO penalty for a nonzero β_j by the quadratic function

$$\frac{|\beta_j|}{|\hat{\beta}_j^*|^\gamma} \approx \frac{|\beta_{j0}|}{|\hat{\beta}_j^*|^\gamma} + \frac{\beta_j^2 - \beta_{j0}^2}{2|\beta_{j0}(\hat{\beta}_j^*)^\gamma|},\ j = 1, ..., s.$$

Then by using arguments similar to Fan and Li (2001), the ALASSO estimates can be approximated by computing

$$\hat{\boldsymbol{\beta}}^{(k)} = \left\{ \sum_{i=1}^{n} \left[(\mathbf{W}_i - \hat{\mathbf{W}}_i)^{\otimes 2} - \Sigma_u \right] + \lambda_n \Sigma(\hat{\boldsymbol{\beta}}^{(k-1)}) \right\}^{-1} \sum_{i=1}^{n} (\mathbf{W}_i - \hat{\mathbf{W}}_i)(Y_i - \hat{Y}_i)$$

iteratively, where $\hat{\boldsymbol{\beta}}^{(k)}$ is the estimate at the k-th iteration and $\hat{\boldsymbol{\beta}}^{(k-1)}$ is the estimate at the $(k-1)$-th iteration, $\Sigma(\boldsymbol{\beta}) = \text{diag}\ (\frac{I_{(\beta_1 \neq 0)}}{|\beta_1(\hat{\beta}_1^*)^\gamma|}\ , ..., \frac{I_{(\beta_d \neq 0)}}{|\beta_d(\hat{\beta}_d^*)^\gamma|})$, and $I_{(\cdot)}$ is an indicator function. This leads to the following estimated variance-covariance matrix of $\hat{\boldsymbol{\beta}}^{nz}$:

$$\widehat{\mathbf{COV}}(\hat{\boldsymbol{\beta}}^{nz}) = \frac{1}{n} \left[\hat{B}_n^{nz} + \frac{\lambda_n}{n} \Sigma(\hat{\boldsymbol{\beta}}^{nz}) \right]^{-1} \hat{F}_n^{nz} \left[\hat{B}_n^{nz} + \frac{\lambda_n}{n} \Sigma(\hat{\boldsymbol{\beta}}^{nz}) \right]^{-1}, \tag{8}$$

where $\hat{B}_n^{nz} = \frac{1}{n} \sum_{i=1}^{n} \left[(\mathbf{W}_i^{nz} - \hat{\mathbf{W}}_i^{nz})^{\otimes 2} - \Sigma_u^{nz} \right]$, $\hat{F}_n^{nz} = \frac{1}{n} \sum_{i=1}^{n} \Big((\mathbf{W}_i^{nz} - \hat{\mathbf{W}}_i^{nz}) \big[Y_i - \hat{Y}_i - (\mathbf{W}_i^{nz} - \hat{\mathbf{W}}_i^{nz})^\top \hat{\boldsymbol{\beta}}^{nz} \big] + \Sigma_u^{nz} \hat{\boldsymbol{\beta}}^{nz} \Big)^{\otimes 2}$, and $\mathbf{W}_i^{nz}$, $\hat{\mathbf{W}}_i^{nz}$ and Σ_u^{nz} are sub-matrices defined analogously to $\mathbf{W}_i$, $\hat{\mathbf{W}}_i$ and Σ_u respectively, and having dimensions conformable to $\hat{\boldsymbol{\beta}}^{nz}$. If Σ_u and Σ_v are unknown, Σ_u can be replaced by $\frac{1}{J}\hat{\Sigma}_u$ developed previously and $\hat{\mathbf{W}}_i$ and $\hat{Y}_i$ can be replaced by $\hat{\tilde{\mathbf{W}}}_i$ and $\hat{\tilde{Y}}_i$ respectively.

3.2 A modified LARS algorithm

The LARS developed by Efron *et al.* (2004) is a variable selection algorithm. Under a linear model setup, Efron *et al.* (2004) and Zou (2006) showed that with slight modifications, this algorithm can be used to find the solution paths of the LASSO and ALASSO. Here, we modify the algorithm of Zou (2006) to account for the special features of the VCPLE model.

The basic idea underlying our method is as follows. Write $\widetilde{\mathbf{Y}} = (Y_1 - \hat{Y}_1, ..., Y_n - \hat{Y}_n)^\top$, $\widetilde{\mathcal{W}} = (\mathbf{W}_1 - \hat{\mathbf{W}}_1, ..., \mathbf{W}_n - \hat{\mathbf{W}}_n)^\top$ and $A = \widetilde{\mathcal{W}}^\top \widetilde{\mathcal{W}} - n\Sigma_u$. Then by some matrix manipulations, the objective function within (4) may be written as

$$\begin{aligned}
\mathcal{L}(\boldsymbol{\beta}) \equiv & \left\| \widetilde{\mathbf{Y}} - \widetilde{\mathcal{W}} \boldsymbol{\beta} \right\|^2 - n\boldsymbol{\beta}^\top \Sigma_u \boldsymbol{\beta} + \lambda_n \sum_{j=1}^{d} \frac{|\beta_j|}{|\hat{\beta}_j^*|^\gamma} \\
= & \widetilde{\mathbf{Y}}^\top (I - \widetilde{\mathcal{W}} A^{-1} \widetilde{\mathcal{W}}^\top) \widetilde{\mathbf{Y}} + \left\| A^{-\frac{1}{2}} \widetilde{\mathcal{W}}^\top \widetilde{\mathbf{Y}} - A^{\frac{1}{2}} \boldsymbol{\beta} \right\|^2 + \lambda_n \sum_{j=1}^{d} \frac{|\beta_j|}{|\hat{\beta}_j^*|^\gamma}.
\end{aligned} \tag{9}$$

Let $\mathbf{y} = A^{-\frac{1}{2}} \widetilde{\mathcal{W}}^\top \widetilde{\mathbf{Y}}$ and $(\mathbf{x}_1/|\hat{\beta}_1^*|^\gamma, ..., \mathbf{x}_d/|\hat{\beta}_d^*|^\gamma) = A^{\frac{1}{2}}$. Note that the first term on the r.h.s. of (9) does not involve $\boldsymbol{\beta}$. Thus, minimizing $\mathcal{L}(\boldsymbol{\beta})$ is equivalent to minimizing

$$\left\| \mathbf{y} - \sum_{j=1}^{d} \mathbf{x}_j \frac{\beta_j}{|\hat{\beta}_j^*|^\gamma} \right\|^2 + \lambda_n \sum_{j=1}^{d} \frac{|\beta_j|}{|\hat{\beta}_j^*|^\gamma}. \tag{10}$$

Denote $\tilde{\beta}_j = \frac{\beta_j}{|\hat{\beta}_j^*|^\gamma}$ $j = 1, ..., d$. Equation (10) then becomes

$$\left\| \mathbf{y} - \sum_{j=1}^{d} \mathbf{x}_j \tilde{\beta}_j \right\|^2 + \lambda_n \sum_{j=1}^{d} |\tilde{\beta}_j|, \tag{11}$$

which has the same form as the original LASSO penalty. This transformation allows the application of the LARS algorithm to find the solution path with respect to $\tilde{\beta}_j$, $j = 1, ..., d$. The steps of our algorithm are summarized as follows.

Steps of the modified LARS algorithm:

Step 1. Compute $A = \widetilde{\mathcal{W}}^\top \widetilde{\mathcal{W}} - n\Sigma_u$, and obtain $\mathbf{y} = A^{-\frac{1}{2}} \widetilde{\mathcal{W}}^\top \widetilde{\mathbf{Y}}$ and $\mathbf{x}_j = \mathbf{e}_j |\hat{\beta}_j^*|^\gamma$ for $j = 1, ..., d$, where $\mathbf{e}_j$ is the jth column of $A^{\frac{1}{2}}$; Σ_u in A may be replaced by $\frac{1}{J}\hat{\Sigma}_u$ if it is unknown, and $\hat{\mathbf{W}}_i$ and $\hat{Y}_i$ may be replaced by $\hat{\widetilde{\mathbf{W}}}_i$ and $\hat{\widetilde{Y}}_i$ if Σ_v is unknown. .

Step 2. Apply the steps of LARS as per Efron *et al.* (2004) to obtain the solution path of

$$\hat{\tilde{\boldsymbol{\beta}}} = \underset{\boldsymbol{\beta}}{\operatorname{argmin}} \left\{ \left\| \mathbf{y} - \sum_{j=1}^{d} \mathbf{x}_j \beta_j \right\|^2 + \lambda_n \sum_{j=1}^{d} |\beta_j| \right\}.$$

Step 3. Obtain the final solution $\hat{\beta}_j = \hat{\tilde{\beta}}_j |\hat{\beta}_j^*|^\gamma$, $j = 1, ..., d$.

It is worth pointing out that despite the transformation of the covariate matrix, the piecewise linearity property (Osborne *et al.*, 2000; Efron *et al.*, 2004) of the solution with respect to λ_n, the tuning parameter, still holds for our algorithm. Thus, after applying the transformation of Step 3 the entire solution path of the ALASSO corresponding to different λ_n values can be obtained. Indeed, one characteristic of the ALASSO estimates is their dependence on the choice of the tuning parameter. The fact that our algorithm yields the whole solution path, and not just the solution based on a single λ_n value, is a particularly strong feature of our method. The optimal λ_n can then be obtained by comparing the properties of the ALASSO estimates based on different tuning parameters. This particular merit of the ALASSO is not shared by other methods like the SCAD method which results in just one solution corresponding to the value of λ_n chosen in advance.

4 Simulation experiments and a real data example

4.1 Simulation experiments

In this subsection, we examine via simulations the small sample properties of the ALASSO estimator computed by the modified LARS algorithm. Note that the estimation of model (1) also involves the estimation of the non-parametric component even though our main interest centers on the selection of variables in the parametric component. As discussed in Section 1, our method uses local linear approximation as the basis for estimating the non-parametric component. In our numerical analysis, the selection of the associated bandwidth parameter is based on cross-validation, as in Fan and Huang (2005) and You and Chen (2006). We use the modified least-squares estimator $\hat{\boldsymbol{\beta}}_{\text{MLS}}$ defined in (2) as the consistent estimator $\hat{\boldsymbol{\beta}}^*$ included in the formula of the ALASSO estimator $\hat{\boldsymbol{\beta}}$, as seen from (4) and (6).

Consider the following data generating process:

$$\begin{cases} Y & = \mathbf{X}^\top \boldsymbol{\beta} + Z_1 \sin(2\pi T) + Z_2 \sin(6\pi T) + \varepsilon, \\ \mathbf{W} & = \mathbf{X} + \mathbf{U}, \\ \boldsymbol{\zeta} & = \mathbf{Z} + \mathbf{V}, \end{cases}$$

where $\mathbf{X}= (X_1, \cdots, X_8)^\top$; Z_1, Z_2 and the covariates in $\mathbf{X}$ are each $N(0, 3)$; ε is $N(0, \sigma^2)$; T is Uniform$(0, 1)$; $\mathbf{U}$ is $N(0,\ \sigma_u^2 I_8)$; $\mathbf{V}$ is $N(0,\ \sigma_u^2 I_2)$; and the correlation matrix of $(\mathbf{X}^\top, Z_1, Z_2)^\top$ is given by $\{C_{ij}\}_{10\times 10}$ with $C_{ij} = 0.5^{|i-j|}$, $i,\ j = 1, ..., 10$. We consider $\sigma = 1$; $\sigma_u = 0$, 0.1, 0.3, 0.5 and 1; $n = 100$, 200; and the following scenarios of $\boldsymbol{\beta}$:

$$\begin{aligned} &\text{S1: } \boldsymbol{\beta} = (3, 1.5, 0, 0, 2, 0, 0, 0)^\top, \\ &\text{S2: } \boldsymbol{\beta} = (0.85, 0.85, 0, 0, 0.85, 0, 0, 0)^\top, \\ &\text{S3: } \boldsymbol{\beta} = (3, 1.5, 0, 1, 2, 0, 1, 1)^\top, \\ &\text{S4: } \boldsymbol{\beta} = (3, 2, 0, 0, 0, 0.85, 0.5, 0)^\top \\ &\text{S5: } \boldsymbol{\beta} = (3, -2, 0, 0, 0, 0.85, -0.5, 0)^\top, \\ &\text{S6: } \boldsymbol{\beta} = (3, -2, 0, 0, 0, 0.85, 0.5, 0)^\top. \end{aligned}$$

Scenarios S1 and S2 represent models with large and small non-zero coefficients respectively, with the number of zero coefficients being five in both cases. Scenario S3 has only two non-zero coefficients and is a non-sparse model. Scenario S4 contains both large and small non-zero coefficients. Scenarios S5 and S6 are similar to Scenario 4 except for the signs of some of the coefficients. We assess the performance of estimators on the basis of mean squared errors, defined as MSE $= \mathbf{E}\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0\|^2$. For comparison with the ALASSO, we also evaluate the MSEs of the SCAD, LASSO, ORACLE and full model estimators. The ORACLE estimator is a "hypothetical" estimator computed using (2) based on the true model that contains none of the covariates with zero coefficients. The ORACLE estimator is expected to perform best since it is based on the true model which is unknown in practice, and thus serves as a benchmark for comparisons. We use two-fold cross-validation to select γ for the ALASSO, and five-fold cross-validation to select the tuning parameter λ_n in both the LASSO and ALASSO penalty functions. For the SCAD estimator, a perturbed version of the local quadratic approximation algorithm (Hunter and Li, 2005; Liang and Li, 2009) is used to optimize the target function. The tuning parameter is selected by the BIC following Wang *et al.* (2007) and Liang and Li (2009). All of our simulations are based on 1000 replications.

Table 1 reports the MSEs of the various estimators. The following general patterns are apparent. First, as expected, the hypothetical estimator ORACLE nearly always results in the best estimates. Second, with few exceptions, the full model estimator is the least preferred estimator in MSE terms. The full model estimator performs especially poorly when σ_u is large. This suggests that in the context of the VCPLE model, including all the covariates indiscriminately is generally an inferior strategy to variable selection, especially when the covariates are measured with large errors. Also, when $\sigma_u \leq 0.3$, the ALASSO always yields smaller MSE than the LASSO; when $\sigma_u = 0.5$, the ALASSO is still the preferred estimator except under S4 with $n = 100$, but when $\sigma_u = 1.0$, the ALASSO is inferior to the LASSO in the majority of cases. Across all cases considered the ALASSO generally have an edge over the SCAD estimator. Exceptions occur, for example, under scenario S1, where the model contains some large non-zero coefficients, and the SCAD is found to have an advantage over the ALASSO in seven out of ten cases. Similar to the ALASSO, the SCAD estimator is dominated by the LASSO when $\sigma_u = 1.0$, but unlike the ALASSO which generally has

smaller MSE than the LASSO for other values of σ_u, the SCAD can deliver worse estimates than the LASSO frequently when $\sigma_u = 0.5$, and occasionally when $\sigma_u = 0.3$. Overall speaking, there is a tendency for the performance of the ALASSO and the SCAD relative to the LASSO to deteriorate as the measurement errors grow. We think this may be attributable to the large variations in the estimator $\hat{\boldsymbol{\beta}}^*$ (which is included in the ALASSO estimator) when $\mathbf{W}$ and $\boldsymbol{\zeta}$ are poor surrogates for $\mathbf{X}$ and $\mathbf{Z}$.

It is worth noting that under scenario S3 where the true model is close to the full model, the (relative) performance of the full model estimator improves as expected, and can be superior to that of the LASSO, although the ALASSO is still the more favored estimator, second only to the ORACLE most of the time. Interestingly, under scenario S3 and $\sigma_u = 1.0$, the LASSO yields smaller MSE than the ORACLE. Generally speaking, the inclusion of negative alongside positive coefficients (as in Scenarios S5 and S6) does not appear to have any significant bearing on the results.

Table 2 presents the average number of "correct" and "incorrect" zero estimates for the ALASSO, SCAD and LASSO based on 1000 replications under the six scenarios; a zero estimate is considered to be "correct" if the actual coefficient is zero, and "incorrect" otherwise. For the $\boldsymbol{\beta}$ specifications we have chosen, the target values of correct zeros are 5, 5, 2, 4, 4 and 4 for scenarios S1, S2, S3, S4, S5 and S6 respectively, and 0 is the target value of incorrect zeros for all the six scenarios. We observe from the table that in all cases the ALASSO and the SCAD provide more accurate number of correct zeros than does the LASSO. For the majority of cases, the SCAD is to be preferred to the ALASSO in terms of producing the number of correct zeros. Interestingly, for a fixed sample size, for all scenarios except S3, as σ_u increases, the ALASSO generally improves in terms of its ability to correctly produce zero estimates for zero coefficients, but it also incorrectly estimates non-zero coefficients as zeroes more frequently than desired. This behaviour is also observed for the LASSO under all six scenarios including S3, which corresponds to a non-sparse model with relatively large parameters. This may be taken as an indication that these strategies tend to select sparser models as measurement errors grow. The reason for this rather curious finding is probably to do with the fact that cross-validation was adopted for selecting the tuning parameter λ_n; due to the presence of measurement errors, cross-validation favors a tuning parameter that results in more zero coefficients in the model in order to have lower model prediction errors. It is also found that under scenarios S4, S5 and S6, the LASSO less frequently estimates non-zero coefficients incorrectly as zero than the ALASSO or SCAD do, but its ability of generating correct zero estimates is inferior to that of its other two competitors.

Overall speaking, neither the SCAD nor the ALASSO dominates each other - the ALASSO generally results in more accurate estimators, while the SCAD is generally a better strategy for choosing the right variables in the model. On the other hand, probably due to its lack of the oracle property, the LASSO is frequently the worst strategy in terms of both performance yardsticks. We have also found that the ALASSO is computationally more efficient than the SCAD - the computing time required for producing the simulation results for the ALASSO is only about a quarter of the time for producing the same results for SCAD. This is encouraging particularly in view of the fact that the ALASSO uses cross-validation which is computationally intensive to select the tuning parameter.

We have also evaluated the accuracy of formula (8) for calculating standard errors of the ALASSO estimates of the non-zero coefficients. Table 3 reports the results based on 1000 simulation replications for scenario S1. The results for the other scenarios are similar and they are omitted for brevity. In Table 3, $\widehat{SE}$ is the average of the standard errors calculated using formula (8), whereas SE is the standard error of the estimates from the replicated samples. Although

$\widehat{SE}$ always underestimates SE, the two values are nevertheless very close for small σ_u's, especially when n=200. However, as σ_u increases, the accuracy of $\widehat{SE}$ deteriorates, *ceteris paribus*. As expected, other things being equal, more accurate values of $\widehat{SE}$ are obtained when n is large than when it is small. It also appears that the variability of ε has little effect on the accuracy of $\widehat{SE}$.

4.2 A real data example

This subsection considers an application of the proposed method to a subset of data from the Continuing Survey of Food Intakes by Individuals (CSFII) conducted by the U.S. Department of Agriculture. The same data set has been used in a number of health and nutritional studies, e.g., (Thompson *et al.*, 1992). This data set contains dietary intake and related information of $n = 1827$ individuals between the age of 25 and 50. Using the available data, we specify the following model for calories intake, denoted by y:

$$y = \sum_{i=1}^{18} \beta_i x_i + f_0(t) + z f_1(t) + \varepsilon,$$

where x_1 is the body mass index, x_2, x_3, x_4, x_5 and x_6 are intake levels of fat, protein, carbohydrates, Vitamin A and Vitamin C respectively, x_7 - x_{13} and x_{14} - x_{18} are two groups of indicator variables representing various Hispanic and other race categories respectively, z is income, and t is age. We use a nonparametric function to address the age effect as the scatter plot of y and t reveals that the relationship between these two variables is nonlinear; we also postulate that the effect of age changes with income. In addition, x_5 and x_6 are measured with errors, and they are replaced by the mean values of the observed surrogates.

We adopt the same methods of choosing the bandwidth, tuning parameters and γ as per the simulation exercises of Section 4.1. Again, $\hat{\boldsymbol{\beta}}_{\text{MLS}}$ is used as the consistent estimator included in the ALASSO. The solution paths of the ALASSO based on the modified LARS algorithm are presented in Figure 1, where the various paths are labeled by the letters corresponding to the different variables in Table 4, and the vertical dotted line corresponds to the tuning parameter $\lambda_n = 0.1604235$ selected by five-fold cross validation. The coefficient estimates based on this tuning parameter are given in Table 4. For comparison purposes, we also provide the estimates obtained from the full model. The results show that the intake levels of fat, protein and carbohydrates are the only covariates selected by ALASSO, and as expected, calories intake has a positive relationship with each of these three covariates. Figure 2 gives the plot for the residuals $r_i = y_i - \mathbf{W}_i^\top \hat{\boldsymbol{\beta}} - \hat{f}_0(t_i) - z_i \hat{f}_1(t_i)$; with the two curves being the estimated curves of $f_0(\cdot)$ and $f_1(\cdot)$.

5 Discussion

The findings in this paper clearly demonstrate that the ALASSO has considerable appeal as a unified estimation and variable selection method for the VCPLE model. This is also the first time that the LARS algorithm is applied to models subject to measurement errors in the covariates. One potential difficulty, however, with the modified least-squares method which we rely upon for correcting the measurement errors is that while this method has desirable large sample properties, it may not possess similar properties in finite samples; for example, it is unsure if $\frac{1}{n}\sum_{i=1}^{n}(\mathbf{W}_i - \hat{\mathbf{W}}_i)^{\otimes 2} - \Sigma_u$ is a positive definite matrix when the sample size is small. For future research it would be worthwhile to apply other methods of measurement errors correction (e.g., the orthogonal regression method developed in Liang and Li (2009)) to the context of investigation considered here.

While we have used cross-validation which is computationally intensive to select the tuning parameter in the ALASSO, Wang *et al.* (2007) showed that in the case of the SCAD, if the BIC is used to select the tuning parameter, the resulting estimator has superior asymptotic properties to the estimator based on the tuning parameter chosen by cross-validation. It remains to explore the use of the BIC for selecting the tuning parameter in the context of the ALASSO. It is envisaged that some modifications to the BIC may be required to account for the features of the ALASSO.

Acknowledgements Zou's work was partially supported by a grant from the Hundred Talents Program of the Chinese Academy of Sciences and two grants from the National Natural Science Foundation of China (Grant nos. 70625004 and 11021161). Wan's work was partially supported by a General Research Fund from the Hong Kong Research Grants Council (Grant no. CityU-102709). We are very grateful to Professor Hua Liang and two referees for constructive comments and suggestions which have led to substantial improvement in the paper. The usual disclaimer applies.

A Proofs of theorems

Our theoretical results depend on the following technical assumptions which are common in the semi-parametric and measurement errors literatures. Fan and Huang (2005) and You and Chen (2006) also made the same assumptions in their studies.

1. The random variable T has bounded support $\mathbf{\Omega}$; in addition, the density f of T is Lipschitz continuous and bounded away from 0 on its support.

2. For each $T \in \mathbf{\Omega}$, $\mathbf{E}(\mathbf{Z}\mathbf{Z}^\top|T)$, $\mathbf{E}(\mathbf{X}\mathbf{X}^\top|T)$ and $\mathbf{E}(\mathbf{Z}\mathbf{X}^\top|T)$ are Lipschitz continuous.

3. There exists some $t > 2$ s.t. $\mathbf{E}\|\mathbf{X}\|^{2t} < \infty$, $\mathbf{E}\|\mathbf{Z}\|^{2t} < \infty$, $\mathbf{E}\|\mathbf{U}\|^{2t} < \infty$, $\mathbf{E}\|\mathbf{V}\|^{2t} < \infty$, and $\mathbf{E}\|\varepsilon\|^{2t} < \infty$; as well, there exists some $\rho < 2 - t^{-1}$ s.t. $nh^{2\rho-1} \to \infty$.

4. $\alpha_j(\cdot), j = 1, ..., r$, are twice continuously differentiable in $T \in \mathbf{\Omega}$.

5. $K(\cdot)$ is a symmetric density with compact support.

6. The bandwidth h satisfies the conditions $nh^8 \to 0$ and $nh^2/[\log(n)]^2 \to \infty$.

Assumptions 1, 2 and 4 are required for the smoothness of the models of interest, while Assumption 5 is required for obtaining a closed form of the estimator of the unknown function vector. Assumption 3 places conditions on the moments of covariates and the rate of convergence of the bandwidth to guarantee uniform consistency of the kernel estimators. It was first used by Mack and Silverman (1982), and subsequently adopted in a number of other semi-parametric studies then frequently used in semi-parametric modeling (e.g. Fan and Huang, 2005; You and Chen, 2006; Li and Liang, 2008). Assumption 3 is generally satisfied in practice. For instance, suppose $t = 3$, meaning that all the random variables have finite sixth moment, and let $\rho = \frac{4}{3}$ and $h = O(n^{-\frac{1}{5}})$. Then since $nh^{2\rho-1} = O(n^{\frac{2}{3}}) \to \infty$, all conditions of Assumption 3 are thus satisfied. The same bandwidth also satisfies Assumption 6 for guaranteeing the optimal convergence rate in the estimation of the linear component of the model.

We shall first give the proof of normality followed by that of sparsity, as the proof of the latter requires results of the former. The proofs need the following lemma:

Lemma 1. *Provided that Assumptions 1-6 hold, we have following result:*

$$\frac{1}{n}\sum_{i=1}^{n}(\mathbf{W}_i-\hat{\mathbf{W}}_i)^{\otimes 2}\xrightarrow{P}B+\Sigma_u,$$

where $B=\mathbf{E}(\mathbf{X}\mathbf{X}^\top)-\mathbf{E}\left(\Phi_T\Gamma_T^{-1}\Phi_T^\top\right)+\mathbf{E}\left(\Phi_T\Gamma_T^{-1}\Sigma_v\Gamma_T^{-1}\Phi_T^\top\right)$, $\Phi_T=\mathbf{E}(\mathbf{X}\mathbf{Z}^\top|T)$, *and* $\Gamma_T=\mathbf{E}(\mathbf{Z}\mathbf{Z}^\top|T)$.

Proof of Lemma 1. From Lemma 7.1 of Fan and Huang (2005), we can show that

$$(D_t^\zeta)^\top\Omega_t D_t^\zeta-\phi_t=\begin{pmatrix}1&0\\0&\mu_2\end{pmatrix}\otimes nf(t)\Gamma_t[1+O_P(c_n)],\text{and}$$
$$(D_t^\zeta)^\top\Omega_t\mathcal{W}=(1,0)^\top\otimes nf(t)\Phi_t[1+O_P(c_n)]$$

hold uniformly in $t\in\mathbf{\Omega}$, where $\mu_2=\int t^2K(t)\mathrm{d}t$ and $c_n=h^2+\log(1/h)/(nh)$. Combining these two equations, we have, uniformly in T_i,

$$\hat{\mathbf{W}}_i=\mathbf{E}(\mathbf{X}_i\mathbf{Z}_i^\top|T_i)[\mathbf{E}(\mathbf{Z}_i\mathbf{Z}_i^\top|T_i)]^{-1}\boldsymbol{\zeta}_i[1+O_P(c_n)]. \tag{12}$$

So,

$$\frac{1}{n}\sum_{i=1}^{n}(\mathbf{W}_i-\hat{\mathbf{W}}_i)^{\otimes 2}=\frac{1}{n}\sum_{i=1}^{n}\left[\mathbf{W}_i-\mathbf{E}(\mathbf{X}_i\mathbf{Z}_i^\top|T_i)[\mathbf{E}(\mathbf{Z}_i\mathbf{Z}_i^\top|T_i)]^{-1}\boldsymbol{\zeta}_i\right]^{\otimes 2}[1+O_P(c_n)].$$

The required result then follows from the law of large numbers. □

Proof of Theorem 1: asymptotic normality. For brevity, we only provide the proof for the Σ_u known case. The corresponding proof when Σ_u is unknown can be similarly obtained. Let $\boldsymbol{\beta}=\boldsymbol{\beta}_0+\frac{1}{\sqrt{n}}\mathbf{u}$, and write $\mathcal{Q}(\mathbf{u})=\mathcal{L}(\boldsymbol{\beta}_0+\frac{1}{\sqrt{n}}\mathbf{u})$. Clearly, minimizing $\mathcal{L}(\boldsymbol{\beta})$ is equivalent to minimizing $\mathcal{Q}(\mathbf{u})$. This also implies an equivalence between the ALASSO estimator $\hat{\boldsymbol{\beta}}$ and the estimator $\hat{\mathbf{u}}$ that minimizes $\mathcal{Q}(\mathbf{u})$. Hence, for our purpose it suffices to consider the minimization of $\mathcal{Q}(\mathbf{u})$ with respect to $\mathbf{u}$. By direct calculations, we obtain

$$\begin{aligned}
&\mathcal{Q}(\mathbf{u})-\mathcal{Q}(\mathbf{0})\\
=&\mathbf{u}^\top\Big[\frac{1}{n}\sum_{i=1}^{n}(\mathbf{W}_i-\hat{\mathbf{W}}_i)^{\otimes 2}-\Sigma_u\Big]\mathbf{u}\\
&-2\frac{1}{\sqrt{n}}\mathbf{u}^\top\sum_{i=1}^{n}\Big\{\big[Y_i-\hat{Y}_i-(\mathbf{W}_i-\hat{\mathbf{W}}_i)^\top\boldsymbol{\beta}_0\big](\mathbf{W}_i-\hat{\mathbf{W}}_i)+\Sigma_u\boldsymbol{\beta}_0\Big\}\\
&+\lambda_n\sum_{j=1}^{d}\frac{|\beta_{j0}+u_j/\sqrt{n}|-|\beta_{j0}|}{|\hat{\beta}_j^*|^\gamma}\\
\equiv&J_1-J_2+J_3.
\end{aligned}$$

To simplify J_1, note from Lemma 1 that the quantity between $\mathbf{u}^\top$ and $\mathbf{u}$ in J_1 goes to B in probability. To simplify J_2, following the idea of You and Chen (2006), we can write

$$\begin{aligned}
&Y_i-\hat{Y}_i-(\mathbf{W}_i-\hat{\mathbf{W}}_i)^\top\boldsymbol{\beta}_0\\
=&\mathbf{X}_i^\top\boldsymbol{\beta}_0+\mathbf{Z}_i^\top\boldsymbol{\alpha}(T_i)+\varepsilon_i-(\boldsymbol{\zeta}_i^\top,\,0)\left[(D_{t_i}^\zeta)^\top\Omega_{t_i}D_{t_i}^\zeta-\phi_{t_i}\right]^{-1}(D_{t_i}^\zeta)^\top\Omega_{t_i}\mathbf{Y}\\
&-\mathbf{X}_i^\top\boldsymbol{\beta}_0-\mathbf{U}_i^\top\boldsymbol{\beta}_0+(\boldsymbol{\zeta}_i^\top,\,0)\left[(D_{t_i}^\zeta)^\top\Omega_{t_i}D_{t_i}^\zeta-\phi_{t_i}\right]^{-1}(D_{t_i}^\zeta)^\top\Omega_{t_i}\mathcal{X}\boldsymbol{\beta}_0+\hat{\mathbf{U}}_i^\top\boldsymbol{\beta}_0\\
=&\varepsilon_i-\mathbf{U}_i^\top\boldsymbol{\beta}_0-\hat{\varepsilon}_i+\hat{\mathbf{U}}_i^\top\boldsymbol{\beta}_0+\mathbf{Z}_i^\top\boldsymbol{\alpha}(T_i)-(\boldsymbol{\zeta}_i^\top,\,0)\left[(D_{t_i}^\zeta)^\top\Omega_{t_i}D_{t_i}^\zeta-\phi_{t_i}\right]^{-1}(D_{t_i}^\zeta)^\top\Omega_{t_i}\mathbf{M},
\end{aligned}$$

where $\mathbf{M} = (\mathbf{Z}_1^\top \boldsymbol{\alpha}(T_1), ..., \mathbf{Z}_n^\top \boldsymbol{\alpha}(T_n))^\top$, $\hat{\varepsilon}_i = (\boldsymbol{\zeta}_i^\top,\ 0)\left[(D_{t_i}^\zeta)^\top \Omega_{t_i} D_{t_i}^\zeta - \phi_{t_i}\right]^{-1}(D_{t_i}^\zeta)^\top \Omega_{t_i}\boldsymbol{\varepsilon}$ and $\boldsymbol{\varepsilon} = (\varepsilon_1, ..., \varepsilon_n)^\top$. Hence,

$$
\begin{aligned}
\Delta_n \equiv & \sum_{i=1}^n \Big\{ \big[Y_i - \hat{Y}_i - (\mathbf{W}_i - \hat{\mathbf{W}}_i)^\top \boldsymbol{\beta}_0\big](\mathbf{W}_i - \hat{\mathbf{W}}_i) + \Sigma_u \boldsymbol{\beta}_0 \Big\} \\
= & \sum_{i=1}^n \big[(\mathbf{W}_i - \hat{\mathbf{W}}_i)(\varepsilon_i - \mathbf{U}_i^\top \boldsymbol{\beta}_0) + \Sigma_u \boldsymbol{\beta}_0\big] + \sum_{i=1}^n (\mathbf{W}_i - \hat{\mathbf{W}}_i)(\hat{\mathbf{U}}_i^\top \boldsymbol{\beta}_0 - \hat{\varepsilon}_i) \\
& + \sum_{i=1}^n (\mathbf{W}_i - \hat{\mathbf{W}}_i) \left\{ \mathbf{Z}_i^\top \boldsymbol{\alpha}(T_i) - (\boldsymbol{\zeta}_i^\top,\ 0)\left[(D_{t_i}^\zeta)^\top \Omega_{t_i} D_{t_i}^\zeta - \phi_{t_i}\right]^{-1}(D_{t_i}^\zeta)^\top \Omega_{t_i} \mathbf{M} \right\} \\
= & \sum_{i=1}^n \Big\{ \big[\mathbf{W}_i - \mathbf{E}(\mathbf{X}_i \mathbf{Z}_i^\top | T_i)[\mathbf{E}(\mathbf{Z}_i \mathbf{Z}_i^\top | T_i)]^{-1} \boldsymbol{\zeta}_i\big](\varepsilon_i - \mathbf{U}_i^\top \boldsymbol{\beta}_0) + \Sigma_u \boldsymbol{\beta}_0 \Big\} \qquad (13) \\
& + \sum_{i=1}^n \big[\mathbf{E}(\mathbf{X}_i \mathbf{Z}_i^\top | T_i)[\mathbf{E}(\mathbf{Z}_i \mathbf{Z}_i^\top | T_i)]^{-1} \boldsymbol{\zeta}_i - \hat{\mathbf{W}}_i\big](\varepsilon_i - \mathbf{U}_i^\top \boldsymbol{\beta}_0) \qquad (14) \\
& + \sum_{i=1}^n (\mathbf{W}_i - \hat{\mathbf{W}}_i)(\hat{\mathbf{U}}_i^\top \boldsymbol{\beta}_0 - \hat{\varepsilon}_i) \qquad (15) \\
& + \sum_{i=1}^n (\mathbf{W}_i - \hat{\mathbf{W}}_i) \left\{ \mathbf{Z}_i^\top \boldsymbol{\alpha}(T_i) - (\boldsymbol{\zeta}_i^\top,\ 0)\left[(D_{t_i}^\zeta)^\top \Omega_{t_i} D_{t_i}^\zeta - \phi_{t_i}\right]^{-1}(D_{t_i}^\zeta)^\top \Omega_{t_i} \mathbf{M} \right\}. \qquad (16)
\end{aligned}
$$

Note that the quantity in (13) is a sum of i.i.d. variables, each having mean $\mathbf{0}$ and variance $F = \mathbf{E}\Big(\big[\mathbf{W} - \mathbf{E}(\mathbf{X}\mathbf{Z}^\top|T)[\mathbf{E}(\mathbf{Z}\mathbf{Z}^\top|T)]^{-1}\boldsymbol{\zeta}\big](\varepsilon - \mathbf{U}^\top\boldsymbol{\beta}_0) + \Sigma_u\boldsymbol{\beta}_0\Big)^{\otimes 2}$. Substituting equation (12) into (14), we can see that (14) is equal to

$$
\sum_{i=1}^n \big[\mathbf{E}(\mathbf{X}_i \mathbf{Z}_i^\top | T_i)[\mathbf{E}(\mathbf{Z}_i \mathbf{Z}_i^\top | T_i)]^{-1} \boldsymbol{\zeta}_i\big](\varepsilon_i - \mathbf{U}_i^\top \boldsymbol{\beta}_0) O_P(c_n).
$$

From the Central Limit Theorem, we know that the quantity in (14) is of order $O_P(\sqrt{n})O_P(c_n)$ $= o_P(\sqrt{n})$. Similarly, we can show that (15) and (16) are also of order $o_P(\sqrt{n})$. So, by Slutsky's Theorem and the Central Limit Theorem, $J_2 \xrightarrow{d} 2\mathbf{u}^\top G$, where $G \sim N(0, F)$.

Last, let us consider J_3. For $\beta_{j0} \neq 0$ $(j = 1, ..., s)$, noting that $\hat{\boldsymbol{\beta}}^*$ is consistent, then for an arbitrarily small δ s.t. $0 < \delta < |\beta_{j0}|$, with probability tending to one,

$$
\lambda_n \frac{|\beta_{j0} + u_j/\sqrt{n}| - |\beta_{j0}|}{|\hat{\beta}_j^*|^\gamma} < \lambda_n \frac{|\beta_{j0} + u_j/\sqrt{n}| - |\beta_{j0}|}{(|\beta_{j0}| - \delta)^\gamma} \leq \frac{\lambda_n}{\sqrt{n}} \frac{|u_j|}{(|\beta_{j0}| - \delta)^\gamma} \to 0.
$$

Now, for $\beta_{j0} = 0$ $(j = s+1, ..., d)$, we have $\lambda_n \frac{|\beta_{j0} + u_j/\sqrt{n}| - |\beta_{j0}|}{|\hat{\beta}_j^*|^\gamma} = \frac{\lambda_n d_n^\gamma}{\sqrt{n}} \frac{|u_j|}{|d_n \hat{\beta}_j^*|^\gamma}$. This last quantity goes to ∞ for $u_j \neq 0$ because $\hat{\beta}_j^*$ is d_n-consistent and $\lambda_n d_n^\gamma / \sqrt{n} \to \infty$. Therefore, $J_3 \xrightarrow{p} 0$ if $u_j = 0$ for all $j = s+1, ..., d$, otherwise $J_3 \xrightarrow{p} \infty$. Combining J_1, J_2 and J_3, it can be shown that the limiting distribution of $\mathcal{Q}(\mathbf{u})$ is

$$
\begin{cases}
(\mathbf{u}^{(1)})^\top B^{(11)} \mathbf{u}^{(1)} - 2(\mathbf{u}^{(1)})^\top G^{(1)}, & \text{if } u_j = 0, j = s+1, ..., d, \\
\infty & \text{otherwise,}
\end{cases}
$$

where $\mathbf{u}^{(1)}$ is the vector that contains the first s components of $\mathbf{u}$, $\mathbf{u}^{(2)}$ contains the other components and $G^{(1)} \sim N(0, F^{(11)})$. Note that this limit is convex. Hence from the epi-convergence results of Geyer (1994) and Knight and Fu (2000), the estimators of $\mathbf{u}^{(1)}$ and $\mathbf{u}^{(2)}$ must satisfy

$$\hat{\mathbf{u}}^{(1)} \xrightarrow{d} (B^{(11)})^{-1} G^{(1)}, and$$
$$\hat{\mathbf{u}}^{(2)} \xrightarrow{d} \mathbf{0}.$$

The proof on the part of asymptotic normality is completed by recognizing that $\hat{\mathbf{u}}^{(1)} = \sqrt{n}(\hat{\boldsymbol{\beta}}_1 - \boldsymbol{\beta}_{10})$ and $G^{(1)} \sim N(0, F^{(11)})$. □

Proof of Theorem 1: Sparsity. To prove the sparsity part, it suffices to show that with probability tending to 1, for any $\beta_j = O(1/\sqrt{n})$, $j = s+1, ...d$,

$$\frac{\partial \mathcal{L}(\breve{\boldsymbol{\beta}})}{\partial \beta_j} > 0 \text{ when } \beta_j > 0, \qquad \text{and} \qquad \frac{\partial \mathcal{L}(\breve{\boldsymbol{\beta}})}{\partial \beta_j} < 0 \text{ when } \beta_j < 0, \tag{17}$$

where $\breve{\boldsymbol{\beta}} = (\hat{\boldsymbol{\beta}}_1^\top, \boldsymbol{\beta}_2^\top)^\top$. By direct calculations, we obtain

$$\begin{aligned}
\frac{1}{\sqrt{n}} \frac{\partial \mathcal{L}(\breve{\boldsymbol{\beta}})}{\partial \beta_j} &= -\frac{2\sum_{i=1}^n \left\{ \left[Y_i - \hat{Y}_i - (\mathbf{W}_i - \hat{\mathbf{W}}_i)^\top \breve{\boldsymbol{\beta}} \right] (\mathbf{W}_i - \hat{\mathbf{W}}_i) + \Sigma_u \breve{\boldsymbol{\beta}} \right\}_j}{\sqrt{n}} + \frac{\lambda_n \text{sgn}(\beta_j)}{\sqrt{n} |\hat{\beta}_j^*|^\gamma} \\
&= -\frac{2(\Delta_n)_j}{\sqrt{n}} + \left\{ \frac{2}{n} \sum_{i=1}^n \left[(\mathbf{W}_i - \hat{\mathbf{W}}_i)^{\otimes 2} - \Sigma_u \right] \times \sqrt{n}(\breve{\boldsymbol{\beta}} - \boldsymbol{\beta}_0) \right\}_j + \frac{\lambda_n \text{sgn}(\beta_j)}{\sqrt{n} |\hat{\beta}_j^*|^\gamma}.
\end{aligned} \tag{18}$$

From the previous proof, we know that the first and second terms in (18) are both of order $O_p(1)$. The third term can be written as $\frac{\lambda_n d_n^\gamma}{\sqrt{n}} \frac{\text{sgn}(\beta_j)}{|d_n \hat{\beta}_j^*|^\gamma} \xrightarrow{p} \text{sgn}(\beta_j)\infty$. This means the sign of the derivative is determined solely by the third term. However, the sign of this term is the same as that of β_j's. Thus, $\frac{\partial \mathcal{L}(\breve{\boldsymbol{\beta}})}{\partial \beta_j}$ and β_j have the same sign with probability tending to one. □

B Proof of the oracle property of the SCAD estimator

The SCAD estimator $\hat{\boldsymbol{\beta}}_S$ is the minimizer of

$$\mathcal{L}_S(\boldsymbol{\beta}) = \frac{1}{2} \sum_{i=1}^n \left\{ Y_i - \hat{Y}_i - [\mathbf{W}_i - \hat{\mathbf{W}}_i]^\top \boldsymbol{\beta} \right\}^2 - \frac{n}{2} \boldsymbol{\beta}^\top \Sigma_u \boldsymbol{\beta} + n \sum_{j=1}^d p_{\lambda_{nj}}(|\beta_j|), \tag{19}$$

where $p_\lambda(\cdot)$ is the SCAD penalty function. The derivative of $p_\lambda(\cdot)$ is

$$p'_\lambda(\beta) = \lambda \left\{ I(\beta \le \lambda) + \frac{(a\lambda - \beta)_+}{(a-1)\lambda} I(\beta > \lambda) \right\} \tag{20}$$

for $a = 3.7$ and $\beta > 0$, and $p'_\lambda(0) = 0$. Now, define

$$\begin{aligned}
a_n &= \max_j \{ p'_{\lambda_{nj}}(|\beta_{j0}|) : \beta_{j0} \neq 0 \}, \\
b_n &= \max_j \{ p''_{\lambda_{nj}}(|\beta_{j0}|) : \beta_{j0} \neq 0 \}, \\
\mathbf{b} &= \left(p'_{\lambda_{n1}}(|\beta_{10}|)\text{sgn}(\beta_{10}), ..., p'_{\lambda_{ns}}(|\beta_{s0}|)\text{sgn}(\beta_{s0}) \right)^\top, and \\
\Sigma_\lambda &= \text{diag}\{ p''_{\lambda_{n1}}(|\beta_{10}|), ..., p''_{\lambda_{ns}}(|\beta_{s0}|) \}.
\end{aligned}$$

Theorem 2. *Suppose that $a_n = O(\frac{1}{\sqrt{n}})$, $b_n \to 0$ and Assumptions 1-6 in Appendix A hold. Then we have:*

(i) With probability approaching one, there exists a local minimizer $\hat{\boldsymbol{\beta}}_S$ of $\mathcal{L}_S(\boldsymbol{\beta})$ which is $\sqrt{n}$-consistent.

If we further suppose $\liminf_{n\to\infty} \liminf_{\beta\to 0^+} \frac{p'_{\lambda_{nj}}(\beta)}{\lambda_{nj}} > 0$, $j = 1, ..., s$, *then we obtain the following results:*

(ii) Sparsity, i.e., $\hat{\boldsymbol{\beta}}_{S2} = \mathbf{0}$ with probability approaching one, where $\hat{\boldsymbol{\beta}}_{S2}$ is the SCAD estimator of $\boldsymbol{\beta}_{20}$.

(iii) Asymptotic normality, i.e.,

$$\sqrt{n}\left(B^{(11)} + \Sigma_\lambda\right)\left\{\hat{\boldsymbol{\beta}}_{S1} - \boldsymbol{\beta}_{10} + \left(B^{(11)} + \Sigma_\lambda\right)^{-1}\mathbf{b}\right\} \xrightarrow{d} N\Big(0,\ F^{(11)}\Big). \tag{21}$$

Proof of part (i). Following the idea of Fan and Huang (2005) and Liang and Li (2009), it suffices to show for any $\epsilon > 0$, there exists a large constant C such that

$$\Pr\left\{\inf_{\|\mathbf{u}\|=C} \mathcal{L}_S(\boldsymbol{\beta}_0 + \frac{\mathbf{u}}{\sqrt{n}}) > \mathcal{L}_S(\boldsymbol{\beta}_0)\right\} \geq 1 - \epsilon. \tag{22}$$

From the proof of Theorem 1,

$$\begin{aligned}
&\mathcal{L}_S(\boldsymbol{\beta}_0 + \frac{\mathbf{u}}{\sqrt{n}}) - \mathcal{L}_S(\boldsymbol{\beta}_0) \\
&= \frac{1}{2}J_1 - \frac{1}{2}J_2 + n\sum_{j=1}^{d}\left\{p_{\lambda_{nj}}(|\beta_{j0} + \frac{u_j}{\sqrt{n}}|) - p_{\lambda_{nj}}(|\beta_{j0}|)\right\} \\
&\geq \frac{1}{2}J_1 - \frac{1}{2}J_2 + n\sum_{j=1}^{s}\left\{p_{\lambda_{nj}}(|\beta_{j0} + \frac{u_j}{\sqrt{n}}|) - p_{\lambda_{nj}}(|\beta_{j0}|)\right\}
\end{aligned} \tag{23}$$

From Fan and Li (2001), the last term of (23) is dominated by J_1 when $a_n = O(\frac{1}{\sqrt{n}})$, $b_n \to 0$ and C is sufficiently large. Also, for sufficiently large C, J_1 dominates J_2. This proves (22). □

Proof of part (ii). We need to show the result in (17) holds if $\mathcal{L}(\cdot)$ is replaced by $\mathcal{L}_S(\cdot)$. To do this, we only need to replace the last term in (18) by $\sqrt{n}\lambda_{nj}\frac{p'_{\lambda_{nj}}(|\beta_j|)}{\lambda_{nj}}\text{sgn}(\beta_j)$, which goes to $\text{sgn}(\beta_j)\infty$ under our assumptions. □

Proof of part (iii). From the results in parts (i) and (ii), with probability approaching one, there exists a $\sqrt{n}$-consistent local minimizer $\hat{\boldsymbol{\beta}}_{S1}$ of $\mathcal{L}\left((\boldsymbol{\beta}_1^\top, 0^\top)^\top\right)$ such that

$$\frac{\partial \mathcal{L}_S\left((\hat{\boldsymbol{\beta}}_{S1}^\top, 0^\top)^\top\right)}{\partial \beta_j} = 0 \text{ for } j = 1, ..., s.$$

By direct calculations and the Taylor series expansion, for $j = 1, ...s$,

$$\begin{aligned}
\frac{\partial \mathcal{L}_S\left((\hat{\boldsymbol{\beta}}_{S1}^\top, 0^\top)^\top\right)}{\partial \beta_j} =& -(\Delta_n)_j + \left\{\sum_{i=1}^{n}\left[(\mathbf{W}_i - \hat{\mathbf{W}}_i)^{\otimes 2} - \Sigma_u\right]\left((\hat{\boldsymbol{\beta}}_{S1}^\top, 0^\top)^\top - \boldsymbol{\beta}_0\right)\right\}_j \\
&+ n\left\{p'_{\lambda_{nj}}(|\beta_{j0}|)\text{sgn}(\beta_{j0}) + p''_{\lambda_{nj}}(|\beta_{j0}|)[1 + o_P(1)](\hat{\boldsymbol{\beta}}_{S1} - \boldsymbol{\beta}_{10})_j\right\}.
\end{aligned}$$

Note that $\Delta_n/\sqrt{n} \xrightarrow{d} G \sim N(0, F)$. The result then follows from Slutsky's theorem and the Central Limit Theorem.

□

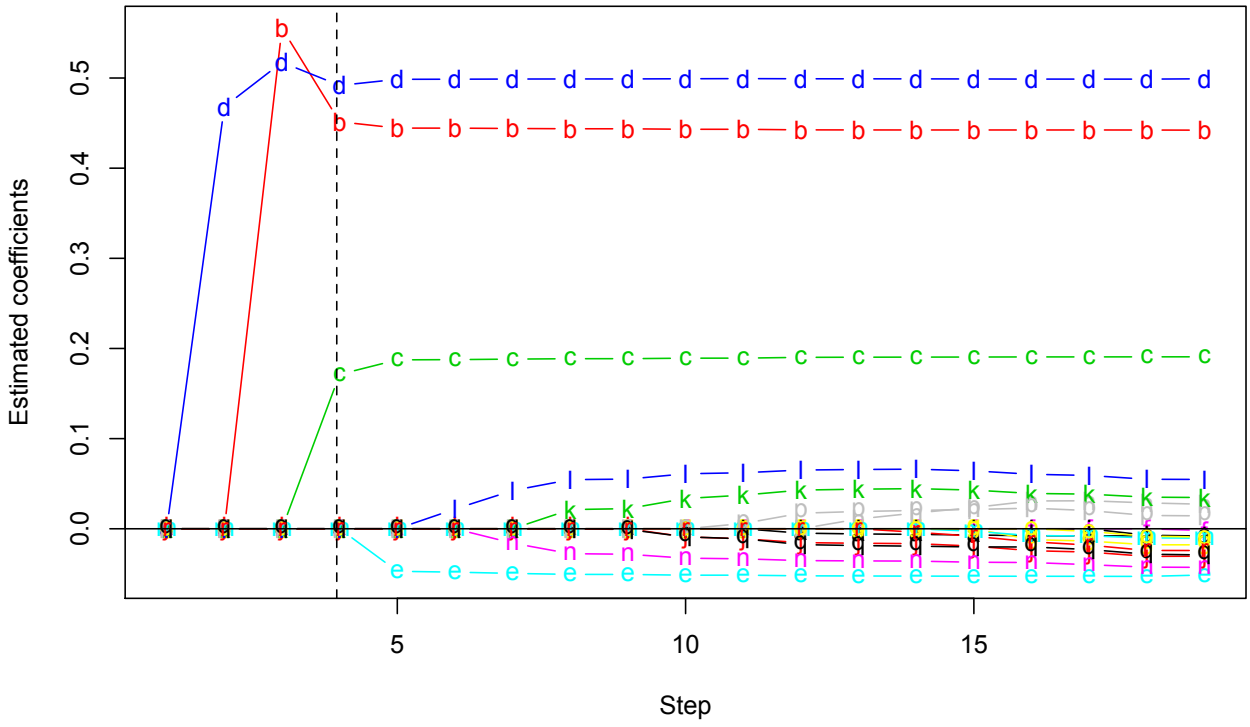

Figure 1: Solution paths of linear coefficients for CSFII data

Table 1: MSEs of estimators

Scenario		$n = 100$					$n = 200$				
	$\sigma_u =$	0.0	0.1	0.3	0.5	1.0	0.0	0.1	0.3	0.5	1.0
S1	ALASSO	0.065	0.071	0.100	0.171	1.450	0.024	0.025	0.036	0.066	0.460
	SCAD	0.050	0.054	0.087	0.194	1.822	0.017	0.019	0.029	0.061	1.096
	LASSO	0.086	0.091	0.127	0.206	0.861	0.033	0.035	0.050	0.092	0.502
	ORACLE	0.040	0.043	0.066	0.121	0.610	0.015	0.016	0.025	0.046	0.211
	full model	0.141	0.152	0.237	0.464	9.691	0.054	0.058	0.093	0.181	1.171
S2	ALASSO	0.073	0.075	0.085	0.106	0.419	0.025	0.025	0.027	0.033	0.107
	SCAD	0.077	0.079	0.097	0.137	0.403	0.022	0.023	0.031	0.051	0.195
	LASSO	0.086	0.087	0.096	0.114	0.249	0.033	0.034	0.036	0.043	0.108
	ORACLE	0.040	0.041	0.047	0.060	0.165	0.015	0.015	0.017	0.021	0.053
	full model	0.141	0.144	0.170	0.233	2.155	0.054	0.055	0.063	0.084	0.288
S3	ALASSO	0.109	0.119	0.195	0.404	3.959	0.040	0.044	0.073	0.150	1.114
	SCAD	0.114	0.124	0.219	0.526	3.016	0.039	0.042	0.073	0.168	1.611
	LASSO	0.165	0.177	0.277	0.457	1.920	0.063	0.067	0.104	0.199	0.902
	ORACLE	0.089	0.096	0.156	0.305	2.091	0.034	0.037	0.062	0.122	0.656
	full model	0.141	0.153	0.252	0.513	12.473	0.054	0.059	0.100	0.203	1.338
S4	ALASSO	0.096	0.103	0.161	0.287	1.539	0.031	0.034	0.053	0.102	0.569
	SCAD	0.098	0.105	0.155	0.265	1.523	0.030	0.033	0.058	0.110	0.923
	LASSO	0.098	0.105	0.148	0.240	0.911	0.037	0.039	0.058	0.102	0.490
	ORACLE	0.058	0.061	0.091	0.162	0.828	0.022	0.023	0.036	0.064	0.284
	full model	0.141	0.151	0.231	0.442	5.630	0.054	0.058	0.089	0.167	0.958
S5	ALASSO	0.106	0.114	0.189	0.459	3.057	0.034	0.036	0.063	0.212	2.191
	SCAD	0.104	0.115	0.213	0.722	3.980	0.031	0.034	0.066	0.466	2.993
	LASSO	0.128	0.137	0.207	0.433	1.992	0.049	0.052	0.090	0.244	1.814
	ORACLE	0.058	0.062	0.094	0.179	1.243	0.022	0.023	0.035	0.068	0.417
	full model	0.141	0.151	0.235	0.470	12.145	0.054	0.057	0.090	0.176	1.361
S6	ALASSO	0.098	0.104	0.161	0.300	2.555	0.032	0.034	0.054	0.146	1.367
	SCAD	0.097	0.104	0.168	0.444	3.586	0.030	0.033	0.062	0.310	2.413
	LASSO	0.115	0.123	0.180	0.337	1.535	0.044	0.047	0.079	0.204	1.270
	ORACLE	0.058	0.062	0.095	0.180	1.220	0.022	0.023	0.036	0.069	0.415
	full model	0.141	0.152	0.237	0.477	11.445	0.054	0.057	0.091	0.179	1.365

Table 2: Average numbers of correct (C) and incorrect (I) zeros

Scenario		$\sigma_u = 0$		$\sigma_u = 0.1$		$\sigma_u = 0.3$		$\sigma_u = 0.5$		$\sigma_u = 1$	
		C	I	C	I	C	I	C	I	C	I
						$n = 100$					
S1	ALASSO	4.330	0	4.312	0	4.436	0	4.610	0	4.739	0.210
	SCAD	4.419	0	4.408	0	4.382	0	4.412	0	4.514	0.185
	LASSO	2.513	0	2.591	0	3.054	0	3.646	0	4.366	0.016
S2	ALASSO	4.078	0	4.112	0	4.156	0	4.270	0.001	4.532	0.216
	SCAD	4.112	0	4.160	0	4.247	0	4.344	0.001	4.393	0.140
	LASSO	2.512	0	2.576	0	2.800	0	3.195	0	3.962	0.007
S3	ALASSO	1.671	0	1.664	0	1.653	0	1.627	0.008	1.723	1.192
	SCAD	1.799	0	1.791	0	1.755	0	1.693	0.014	1.555	1.023
	LASSO	1.136	0.004	1.141	0.004	1.225	0.010	1.318	0.009	1.452	0.330
S4	ALASSO	3.053	0.013	3.051	0.019	3.077	0.059	3.273	0.207	3.718	1.151
	SCAD	3.329	0.009	3.355	0.012	3.358	0.041	3.408	0.129	3.559	1.027
	LASSO	1.906	0	1.921	0	2.242	0.003	2.622	0.037	3.393	0.504
S5	ALASSO	2.882	0.017	2.850	0.022	3.088	0.104	3.583	0.556	3.954	1.845
	SCAD	3.153	0.020	3.132	0.024	3.190	0.140	3.674	0.790	3.792	1.740
	LASSO	1.416	0.003	1.444	0.005	1.903	0.021	2.603	0.206	3.579	1.158
S6	ALASSO	3.014	0.017	2.996	0.018	3.236	0.069	3.626	0.266	3.889	1.156
	SCAD	3.313	0.011	3.329	0.010	3.460	0.051	3.667	0.169	3.645	0.899
	LASSO	1.525	0	1.615	0	2.145	0.003	2.816	0.038	3.434	0.334
						$n = 200$					
S1	ALASSO	4.441	0	4.482	0	4.602	0	4.799	0	4.904	0.020
	SCAD	4.629	0	4.615	0	4.625	0	4.600	0	4.741	0.019
	LASSO	2.552	0	2.653	0	3.327	0	4.029	0	4.632	0
S2	ALASSO	4.369	0	4.383	0	4.445	0	4.624	0	4.838	0.004
	SCAD	4.510	0	4.526	0	4.621	0	4.696	0	4.619	0.002
	LASSO	2.548	0	2.593	0	2.926	0	3.464	0	4.311	0
S3	ALASSO	1.750	0	1.750	0	1.765	0	1.778	0	1.725	0.268
	SCAD	1.901	0	1.878	0	1.853	0	1.810	0	1.652	0.368
	LASSO	1.115	0	1.144	0	1.250	0	1.386	0	1.508	0.056
S4	ALASSO	3.403	0	3.396	0	3.369	0.002	3.423	0.036	3.812	0.653
	SCAD	3.496	0	3.516	0	3.539	0	3.591	0.017	3.756	0.645
	LASSO	1.952	0	1.987	0	2.466	0	2.929	0.001	3.527	0.251
S5	ALASSO	3.327	0	3.322	0	3.396	0.007	3.709	0.204	3.998	1.769
	SCAD	3.457	0	3.474	0.001	3.451	0.008	3.723	0.484	3.910	1.461
	LASSO	1.472	0	1.568	0	2.215	0	2.948	0.052	3.834	1.090
S6	ALASSO	3.375	0	3.372	0	3.543	0	3.834	0.083	3.974	0.844
	SCAD	3.469	0	3.477	0	3.586	0	3.885	0.054	3.832	0.405
	LASSO	1.618	0	1.756	0	2.480	0	3.239	0.001	3.758	0.171

Table 3: Calculated versus simulated SEs

Scenario		$\hat{\beta}_1$		$\hat{\beta}_2$		$\hat{\beta}_3$	
		$\widehat{SE}$	SE	$\widehat{SE}$	SE	$\widehat{SE}$	SE
S1							
$\sigma_u = 0.1$	$n = 100$	0.111	0.129	0.113	0.135	0.100	0.116
	$n = 200$	0.075	0.076	0.076	0.083	0.067	0.071
$\sigma_u = 0.3$	$n = 100$	0.136	0.161	0.137	0.167	0.120	0.142
	$n = 200$	0.093	0.096	0.093	0.100	0.082	0.087
$\sigma_u = 0.5$	$n = 100$	0.181	0.222	0.178	0.222	0.156	0.189
	$n = 200$	0.125	0.133	0.122	0.130	0.107	0.118

Table 4: Results of real data example

Variable		ALASSO		Full Model	
		Estimate	SE	Estimate	SE
x_1	Body mass index	0	0	-0.0080	0.0043
x_2	Fat	0.4521	0.0079	0.4421	0.0090
x_3	Protein	0.1702	0.0054	0.1908	0.0125
x_4	Carbohydrates	0.4919	0.0046	0.4992	0.0054
x_5	Vitamin A	0	0	-0.0512	0.0467
x_6	Vitamin C	0	0	-0.0019	0.0172
x_7	Hispanic category 1	0	0	-0.0178	0.0271
x_8	Hispanic category 2	0	0	0.0272	0.0564
x_9	Hispanic category 3	0	0	-0.0081	0.0255
x_{10}	Hispanic category 4	0	0	-0.0304	0.0249
x_{11}	Hispanic category 5	0	0	0.0347	0.0491
x_{12}	Hispanic category 6	0	0	0.0547	0.0538
x_{13}	Hispanic category 7	0	0	-0.0104	0.0212
x_{14}	Race category 1	0	0	-0.0428	0.0273
x_{15}	Race category 2	0	0	-0.0092	0.0204
x_{16}	Race category 3	0	0	0.0143	0.0300
x_{17}	Race category 4	0	0	-0.0290	0.0236
x_{18}	Race category 5	0	0	-0.0241	0.0345

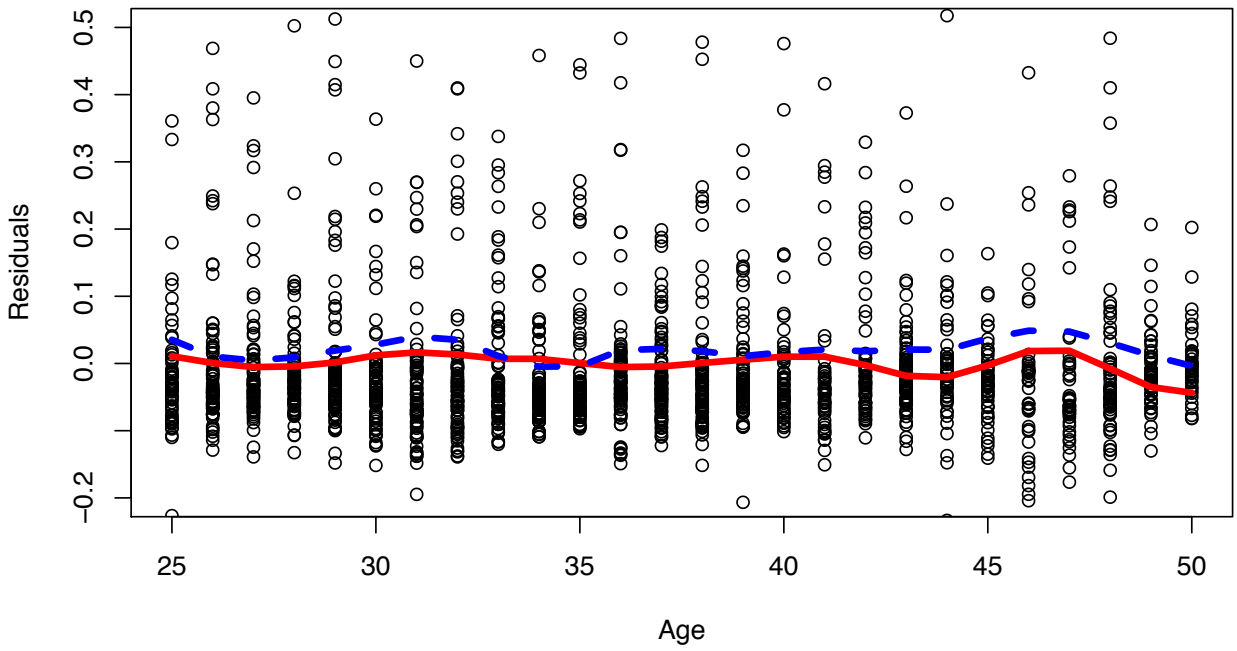

Figure 2: Residual plot: the solid line and the dashed line are estimated curve for $f_0(\cdot)$ and $f_1(\cdot)$ respectively

References

Cai, J., Fan, J., Li, R., and Zhou, H. (2005). Variable selection for multivariate failure time data. *Biometrika* **92**, 303–316.

Carroll, R. J., Ruppert, D., Stefanski, L. A., and Crainiceanu, C. M. (2006). *Measurement Error in Nonlinear Models: A Modern Perspective (2nd ed.)*. Chapman and Hall, New York.

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

Fan, J. and Huang, T. (2005). Profile likelihood inferences on semiparametric varying-coefficient partially linear models. *Bernoulli* **11**, 1031–1057.

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.

Fan, J. and Li, R. (2002). Variable selection for cox's proportional hazards model and frailty model. *The Annals of Statistics* **30**, 74–99.

Fan, J. and Li, R. (2004). New estimation and model selection procedures for semiparametric modeling in longitudinal data analysis. *Journal of the American Statistical Association* **99**, 710–723.

Fan, J. and Peng, H. (2004). Nonconcave penalized likelihood with a diverging number of parameters. *The Annals of Statistics* **32**, 928–961.

Frank, I. E. and Friedman, J. H. (1993). A statistical view of some chemometrics regression tools. *Technometrics* **35**, 109–135.

Geyer, C. J. (1994). On the asymptotics of constrained m-estimation. *The Annals of Statistics* **22**, 1993–2010.

Hall, P. and Ma, Y. (2007). Semiparametric estimators of functional measurement error models with unknown error. *Journal of the Royal Statistical Society (Series B)* **69**, 429–446.

Huang, Z. and Zhang, R. (2009). Empirical likelihood for nonparametric parts in semiparametric varying-coefficient partially linear models. *Statistics & Probability Letters* **79**, 1798–1808.

Hunter, D. R. and Li, R. (2005). Variable selection using mm algorithms. *The Annals of Statistics* **33**, 1617–1642.

Knight, K. and Fu, W. (2000). Asymptotics for lasso-type estimators. *The Annals of Statistics* **28**, 1356–1378.

Li, R. and Liang, H. (2008). Variable selection in semiparametric regression model. *The Annals of Statistics* **36**, 261–286.

Liang, H. (2000). Asymptotic normality of parametric part in partially linear models with measurement error in the nonparametric part. *Journal of Statistical Planning and Inference* **86**, 51–62.

Liang, H., Härdle, W., and Carroll, R. J. (1999). Estimation in a semiparametric partially linear errors-in-variables model. *The Annals of Statistics* **27**, 1519–1535.

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

Liang, H., Wang, S., and Carroll, R. J. (2007). Partially linear models with missing response variables and error-prone covariates. *Biometrika* **94**, 185–198.

Ma, Y. and Carroll, R. J. (2006). Locally efficient estimators for semiparametric models with measurement error. *Journal of the American Statistical Association* **101**, 1465–1474.

Ma, Y. and Li, R. (2009). Variable selection in measurement error models. *Bernoulli* **16**, 274–300.

Mack, Y. P. and Silverman, B. W. (1982). Weak and strong uniform consistency of kernel regression estimates. *Zeitschrift für Wahrscheinlichkeitstheorie und verwandte Gebiete* **61**, 405–415.

Osborne, M. R., Presnell, B., and Turlach, B. A. (2000). A new approach to variable selection in least squares problems. *IMA Journal of Numerical Analysis* **20**, 389–403.

Thompson, E. E., Sowers, M., Frongillo, E., and Parpia, B. (1992). Sources of fiber and fat in diets of u.s. women aged 19 to 50: implications for nutrition education and policy. *American Journal of Public Health* **82**, 695–702.

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

Tsiatis, A. A. and Ma, Y. (2004). Locally efficient semiparametric estimators for functional measurement error models. *Biometrika* **91**, 835–848.

Wang, H. and Leng, C. (2008). A note on adaptive group lasso. *Computational Statistics & Data Analysis* **52**, 5277–5286.

Wang, H., Li, R., and Tsai, C. L. (2007). Tuning parameter selectors for the smoothly clipped absolute deviation method. *Biometrika* **94**, 553–568.

You, J. and Chen, G. (2006). Estimation of a semiparametric varying-coefficient partially linear errors-in-variables model. *Journal of Multivariate Analysis* **97**, 324–341.

You, J. and Zhou, Y. (2006). Empirical likelihood for semiparametric varying-coefficient partially linear regression models. *Statistics & Probability Letters* **76**, 412–422.

You, J., Zhou, Y., and Chen, G. (2006). Corrected local polynomial estimation in varying-coefficient models with measurement errors. *The Canadian Journal of Statistics* **34**, 391–410.

Yuan, M. and Lin, Y. (2006). Model selection and estimation in regression with grouped variables. *Journal of the Royal Statistical Society (Series B)* **68**, 49–67.

Zhang, H. H. and Lu, W. (2007). Adaptive lasso for cox's proportional hazards model. *Biometrika* **94**, 691–703.

Zhang, W., Lee, S.-Y., and Song, X. (2002). Local polynomial fitting in semivarying coefficient model. *Journal of Multivariate Analysis* **82**, 166–188.

Zhao, P. and Xue, L. (2009). Variable selection for semiparametric varying coefficient partially linear models. *Statistics & Probability Letters* **79**, 2148–2157.

Zhao, P. and Xue, L. (2010). Variable selection for semiparametric varying coefficient partially linear errors-in-variables models. *Journal of Multivariate Analysis* **101**, 1872–1883.

Zhao, P. and Xue, L. (2011). Variable selection for semiparametric varying coefficient partially linear models with measurement errors. *working paper, Hechi University and Beijing University of Technology* .

Zou, H. (2006). The adaptive lasso and its oracle properties. *Journal of the American Statistical Association* **110**, 1418–1429.