Variable Selection and Empirical Likelihood based Inference for Measurement Error Data[1]

Hua Liang and Runze Li

Abstract

Using nonconvex penalized least squares, we propose a class of variable selection procedures for linear models and partially linear models when the covariates are measured with additive error. The rate of convergence and the asymptotic normality of the resulting estimate are established. We further demonstrate that, with proper choice of penalty functions and the regularization parameter, the resulting estimate performs as well as an oracle procedure. A robust standard error formula is derived using a sandwich formula, and empirically tested. Local polynomial regression techniques are used to estimate the baseline function in the partially linear model. To avoid to estimate the asymptotic covariance in establishing confidence region of the parameter of interest, we further develop a statistic based on the empirical likelihood principle, and show that the statistic is asymptotically chi-squared distributed. Finite sample performance of the proposed inference procedures is assessed by Monte Carlo simulation studies. We further illustrate the proposed procedures by an application.

Key Words and Phrases: Empirical Likelihood; Estimating equation; LASSO; Local linear regression; SCAD; Variable Selection

Short title: Variable Selection for Measurement Error Data

[1]Corresponding Author: *Hua Liang, hliang@bst.rochester.edu*. Hua Liang (E-mail:hliang@bst.rochester.edu) is Associate Professor, Department of Biostatistics and Computational Biology, University of Rochester, NY 14642. Runze Li (E-mail:rli@stat.psu.edu) is Associate Professor, Department of Statistics, The Pennsylvania State University, University Park, PA 16802-2111. NIH/NIAID grants R01 AI62247 and R01 AI59773. Li's research was supported by a National Institute on Drug Abuse (NIDA) grant P50 DA10075 and NSF grants DMS-0348869.

Variable Selection and Empirical Likelihood based Inference for Measurement Error Data

Abstract

Using nonconvex penalized least squares, we propose a class of variable selection procedures for linear models and partially linear models when the covariates are measured with additive error. The rate of convergence and the asymptotic normality of the resulting estimate are established. We further demonstrate that, with proper choice of penalty functions and the regularization parameter, the resulting estimate performs as well as an oracle procedure. A robust standard error formula is derived using a sandwich formula, and empirically tested. Local polynomial regression techniques are used to estimate the baseline function in the partially linear model. To avoid to estimate the asymptotic covariance in establishing confidence region of the parameter of interest, we further develop a statistic based on the empirical likelihood principle, and show that the statistic is asymptotically chi-squared distributed. Finite sample performance of the proposed inference procedures is assessed by Monte Carlo simulation studies. We further illustrate the proposed procedures by an application.

Key Words and Phrases: Empirical Likelihood; Estimating equation; LASSO; Local linear regression; SCAD; Variable Selection

Short title: Variable Selection for Measurement Error Data

1 Introduction

In many situations, covariates are measured with additive errors in laboratory. Such data are referred as measurement error data or error-in-variable data. Various statistical procedures have been developed for measurement error data. Fuller (1987) and Carroll, Ruppert, and Stefanski (1995) give a systematic survey on this research topic and present many applications of measurement error data. To get insights into the challenge of statistical inference and variable selection for measurement error data, let us demonstrate a simple example. Consider linear regression models for measurement error data:

$$\begin{cases} Y &= X\beta + V\gamma + \varepsilon, \\ W &= X + U, \end{cases} \tag{1}$$

where X is the true but unobserved covariate vector, W is the surrogate of X, V is a covariate vector measured without an error, and ε is a model error with $E(\varepsilon|X) = 0$, U is the measurement error with mean zero and is independent of (Y, X, V). For simplicity of presentation, let us consider both X and V univariate case and assume that ε, U, X and V are independent of each other. Thus,

$$\gamma = \frac{E[\{V - E(V|X)\}\{Y - E(Y|X)\}]}{E\{V - E(V|X)\}^2}.$$

Suppose that $\{W_i, V_i, Y_i\}$, $i = 1, \cdots, n$, is a random sample from model (1). Then a naive estimator of γ is

$$\hat{\gamma} \;=\; \frac{\sum_{i=1}^n \{V_i - \widehat{E}(V_i|W_i)\}\{Y_i - \widehat{E}(Y_i|W_i)\}}{\sum_{i=1}^n \{V_i - \widehat{E}(V_i|W_i)\}^2}.$$

By some straightforward derivation,

$$\hat{\gamma} \to \frac{\text{cov}(V, Y|W)}{\text{var}(V|W)} = \gamma + (1 - \tau_1)\beta\tau_2,$$

where $\tau_1 = E\{\text{var}(X|V)\}/E\{\text{var}(W|V)\}$, and τ_2=cov(X, V)/var(V), the slope of regression X on V. This indicates that the naive estimator is inconsistent. This poses many challenges in variable selection for measurement error data. For example, when V is not significant, i.e., the

true value of γ equals 0, and should be excluded from the final selected model. However, ignoring measurement errors implies that the estimated value of γ is close to $(1-\tau_1)\beta\tau_2$. Existing variable selection procedures likely include this variable into their selected model when the magnitude of $(1-\tau_1)\beta\tau_2$ is large. On the other hand, when $\gamma + (1-\tau_1)\beta\tau_2$ closes to zero, the existing variable selection procedures may delete this variable, even though V is very significant. Thus, ignoring measurement errors, variable selection procedures may select a very misleading model. The purpose of this paper is to develop a class of variable selection procedure without ignoring measurement errors.

Variable selection plays an important role in analysis of measurement error data. In practice, many variables can be introduced to the initial analysis. Deciding which covariates to be kept in the final statistical model has always been a tricky task for data analysis. Stepwise regression and the best subset selection can be directly extended to analysis of data with measurement error. However, they suffer from several drawbacks, the most severe one of which is the lack of stability as pointed out by Breiman (1996). While they are useful in practice, the stepwise deletion and the best subset method ignore stochastic errors inherited in the stage of variable selection. Hence, their theoretical properties are somewhat hard to understand and the sampling properties of the resulting estimates are difficult to establish, even in the classical linear model. Consequently, the confidence intervals based on these methods may not necessarily be valid. Nonconvex penalized least squares approach has been proposed to select significant variables in linear regression model (Fan and Li, 2001). They are improvements of LASSO (Tibshirani, 1996). With suitable choices of penalty functions and regularization parameters (Fan and Li, 2001), the resulting estimates of the nonconvex penalized least squares approaches possess an oracle property. This encourages us to extend the methodology for measurement error data.

Partially linear model is good compromises of parametric models and nonparametric models. It retains the flexibility of nonparametric models for the intercept function and interpretation power of linear regression models. Hence it has been popular in the literature (Engle, *et al.*, 1986; Heckman, 1986; Robinson, 1988; Speckman, 1988). A comprehensive survey was given by Härdle,

Liang, and Gao (2000). The partially linear model has been used for longitudinal data with various statistical formulation (Zeger and Diggle, 1994; Moyeed and Diggle, 1994; Martinussen and Scheike, 1999, 2001; Lin and Ying, 2001; Fan and Li, 2004 and Liang, *et al.*, 2004). Liang, Härdle, and Carroll (1999) considered the partially linear models for measurement error data. There are some existing variable selection procedures for partially linear models (Bunea, 2004, Bunea and Wegkamp, 2004, Fan and Li, 2004). To our best knowledge, however, there is no existing variable selection procedures for measurement error data. Thus, it is desirable to develop effective variable selection procedures for measurement error data. In this paper, we first propose a class of nonconvex penalized least squares variable selection procedures for partially linear measurement error models. We then demonstrate how the rate of convergence of the resulting estimate depends on the regularization parameters of the nonconvex penalized least squares. With proper choice of penalty function and the regularization parameter, we show that the resulting estimate of the proposed procedure asymptotically performs as well as an oracle procedure.

Empirical likelihood has been used to construct accurate confidence intervals or regions in the literature. In this paper, we also investigate how to construct empirical likelihood based confidence intervals for regression coefficients in the partially linear measurement error models. We propose an effective method to construct confidence region for the parameter of interest using empirical likelihood principles. Through the use of the presence of auxiliary information, the newly proposed approach improves the confidence region and increases accuracy of coverage. Moreover, the proposed procedure is easily implemented since it avoids computing covariance matrix of the estimator for the parameters of interest. The estimating covariance matrix may be tedious although straightforward,

The rest of this paper is organized as follows. In Section 2 we propose a class of variable selection procedures for linear models and partially linear models via nonconvex penalized least squares. In Section 3, we develop an empirical likelihood statistic and derive its asymptotic distribution. Simulation study and application of the proposed procedures are presented in Section 4. Regularity conditions and technical proofs are given in the Appendix.

2 Variable selection

In this section, we first propose a class of variable selection procedures for linear measurement error models via nonconvex penalized least squares. We further extended the nonconvex penalized least squares for partially linear measurement error models. We also studied the sampling properties of the proposed procedures in Section 2.3. Issues related to practical implementation are addressed in Section 2.4.

2.1 Linear measurement error models

Consider the following linear measurement error model:

$$\begin{cases} Y &= X^{\mathrm{T}}\boldsymbol{\beta} + \varepsilon, \\ W &= X + U, \end{cases} \tag{2}$$

where X is the true but unobserved d-dimensional covariate vector, W is the surrogate of X, and ε is a model error with $E(\varepsilon|X) = 0$, U is the measurement error with mean zero and covariance matrix Σ_{uu}, and is independent of $\{X, Y\}$. Note that when a covariate is measured without an error, it can be treated as a error-in-variable covariate, in which the error is a random variable with zero mean and zero variance. In this paper, the covariance matrix Σ_{uu} is allowed to be singular in order for model (2) to include covariates measured without error. Suppose that $\{W_i, Y_i\}$, $i = 1, \cdots, n$, is a random sample from model (2). Define a penalized least squares for model (2):

$$\mathcal{L}(\boldsymbol{\beta}) = \sum_{i=1}^{n}(Y_i - W_i^{\mathrm{T}}\boldsymbol{\beta})^2 - n\boldsymbol{\beta}^{\mathrm{T}}\Sigma_{uu}\boldsymbol{\beta} + 2n\sum_{j=1}^{d} p_{\lambda_j}(|\beta_j|), \tag{3}$$

where $p_{\lambda_j}(\cdot)$ is a specified penalty function with a regularization parameter λ_j, which can chosen by data-driven method such as cross-validation and generalized cross validation (GCV, Craven and Wahba, 1979). It is worth noting that the penalties and regularization parameters are unnecessarily same for all coefficients. For instance, we want to keep important variables in the final model, and therefore we should not penalize their coefficients.

The penalized least squares (3) provides a general framework of variable selection for linear measurement error models. Taking the penalty function to be L_0-penalty (also called the entropy

penalty in the literature), namely, $p_{\lambda_j}(|\beta_j|) = 1/2\lambda_j^2 I(|\beta_j| \neq 0)$, where $I\{\cdot\}$ is an indicator function, we may extend the traditional variable selection criteria, including the AIC (Akaike, 1973), BIC (Schwarz, 1978) and RIC (Foster and George, 1994), for linear models to linear measurement error models:

$$\sum_{i=1}^{n}(Y_i - W_i^{\mathrm{T}}\boldsymbol{\beta})^2 - n\boldsymbol{\beta}^{\mathrm{T}}\Sigma_{uu}\boldsymbol{\beta} + 2n\lambda_j^2\sum_{j=1}^{d} I\{|\beta_j| \neq 0\} \tag{4}$$

as $\sum_{j=1}^{d} I\{|\beta_j| \neq 0\}$ equals the size of the selected model. Specifically, the AIC, BIC and RIC correspond to $\lambda_j = (2/n)^{1/2}\sigma$, $\{\log(n)/n\}^{1/2}\sigma$, and $\{\log(d)/n\}^{1/2}\sigma$, respectively.

Note that the L_0-penalty is discontinuous, and therefore optimizing (4) requires exhaustive search over 2^d possible subsets. This poses much computation. Furthermore, as pointed out by Breiman (1996), the best subset variable selection suffers from several drawbacks, including its lack of stability. In the recent literature of variable selection for linear regression model without measurement error, many authors advocated the use of continuous and smooth penalty functions. Bridge regression (Frank and Friedman, 1993) corresponds to L_q-penalty $p_{\lambda_j}(|\beta_j|) = \lambda_j|\beta_j|^q$. In particular, the LASSO (Tibshirani, 1996) corresponds to the L_1-penalty. Fan and Li (2001) studied the choice of penalty functions in depth. They advocated the use of the smoothly clipped absolute deviation (SCAD) penalty, whose first derivative is given by

$$p'_\lambda(\beta) = \lambda\left\{I(\beta \leq \lambda) + \frac{(a\lambda - \beta)_+}{(a-1)\lambda}I(\beta > \lambda)\right\}, \text{ for } \beta > 0,$$

where $a = 3.7$ and $p_\lambda(0) = 0$. For simplicity of presentation, we will use the name "SCAD" for all procedures using the SCAD penalty. As demonstrated in Fan and Li (2001), the SCAD is an improvement of the LASSO in terms of modeling bias and of the bridge regression in terms of stability.

2.2 Partially linear measurement error models

We next propose a class of variable selection procedures for the following partially linear measurement error model:

$$\left\{ \begin{array}{lcl} Y & = & \nu(Z) + X^{\mathrm{T}}\boldsymbol{\beta} + \varepsilon, \\ W & = & X + U, \end{array} \right. \tag{5}$$

where Z is a univariate covariate measured without error, $E(\varepsilon|Z) = 0$, and the other notation is the same as that in (2).

Since $E(\varepsilon|Z) = 0$, then $E(Y|Z) = \nu(Z) + E(X|Z)^{\mathrm{T}}\boldsymbol{\beta}$. Thus,

$$Y - E(Y|Z) = \{X - E(X|Z)\}^{\mathrm{T}}\boldsymbol{\beta} + \varepsilon.$$

This motivates us to consider the following penalized least squares based on partial residuals. Partial residual approach has been proposed for partially linear models in Speckman (1988). Let $m_w(Z) = E(W|Z)$ and $m_y(Z) = E(Y|Z)$. Suppose that $\{W_i, Y_i, Z_i\}$, $i = 1, \cdots, n$ is a random sample from model (5). Define a penalized least squares

$$\begin{array}{rl} & \Psi\{m_w(\cdot), m_y(\cdot), \Sigma_{uu}, \boldsymbol{\beta}\} \\ = & \dfrac{1}{2}\displaystyle\sum_{i=1}^{n}[Y_i - m_y(Z_i) - \{W_i - m_w(Z_i)\}^{\mathrm{T}}\boldsymbol{\beta}]^2 - \dfrac{n}{2}\boldsymbol{\beta}^{\mathrm{T}}\Sigma_{uu}\boldsymbol{\beta} + n\displaystyle\sum_{j=1}^{d} p_{\lambda_j}(|\beta_j|) \end{array} \tag{6}$$

Note that $m_w(\cdot)$ and $m_y(\cdot)$ are unknown regression functions. However, they may be easily estimated using existing nonparametric smoothing methods. In this paper, we will employ local linear regression (Fan and Gijbels, 1996) to estimate $m_w(\cdot)$ and $m_y(\cdot)$. Denote $\hat{m}_w(\cdot)$ and $\hat{m}_y(\cdot)$ to be an estimate of $m_w(\cdot)$ and $m_y(\cdot)$, respectively. Substituting $m_w(\cdot)$ and $m_y(\cdot)$ with their estimate in (6) results in

$$\mathcal{L}_P(\Sigma_{uu}, \boldsymbol{\beta}) \stackrel{\mathrm{def}}{=} \frac{1}{2}\sum_{i=1}^{n}[Y_i - \hat{m}_y(Z_i) - \{W_i - \hat{m}_w(Z_i)\}^{\mathrm{T}}\boldsymbol{\beta}]^2 - \frac{n}{2}\boldsymbol{\beta}^{\mathrm{T}}\Sigma_{uu}\boldsymbol{\beta} + n\sum_{j=1}^{d} p_{\lambda_j}(|\beta_j|). \tag{7}$$

Minimizing $\mathcal{L}_P(\Sigma_{uu}, \boldsymbol{\beta})$ with respect to $\boldsymbol{\beta}$ results in a penalized least squares estimator $\hat{\boldsymbol{\beta}}$.

2.3 Sampling properties

We first assume that Σ_{uu} is known and will consider the case where Σ_{uu} is unknown later. Let $\boldsymbol{\beta}_0 = (\beta_{10}, \ldots, \beta_{d0})^{\mathrm{T}} = (\boldsymbol{\beta}_{10}^{\mathrm{T}}, \boldsymbol{\beta}_{20}^{\mathrm{T}})^{\mathrm{T}}$ be the true value of $\boldsymbol{\beta}$. Without loss of generality, assume that $\boldsymbol{\beta}_{10}$ consists of all nonzero components of $\boldsymbol{\beta}_0$, and $\boldsymbol{\beta}_{20} = \mathbf{0}$. Denote

$$a_n = \max_{1 \le j \le d} \{|p'_{\lambda_j}(|\beta_{j0}|)|, \beta_{j0} \neq 0\}, \text{ and } b_n = \max_{1 \le j \le d} \{|p''_{\lambda_j}(|\beta_{j0}|)|, \beta_{j0} \neq 0\}. \tag{8}$$

Denote the number of nonzero components of $\boldsymbol{\beta}_0$ by s and

$$\mathbf{b} = \{p'_\lambda(|\beta_{10}|)\mathrm{sgn}(\beta_{10}), \ldots, p'_\lambda(|\beta_{s0}|)\mathrm{sgn}(\beta_{s0})\}^{\mathrm{T}}, \text{ and } \Sigma_\lambda = \mathrm{diag}\{p''_\lambda(|\beta_{10}|), \ldots, p''_\lambda(|\beta_{s0}|)\}. \tag{9}$$

Denote U_{11} and X_{11} the vectors comprised by the first s elements of U_1 and X_1, respectively. Let $\Sigma_{uu1} = E(U_{11}U_{11}^{\mathrm{T}})$ and $\Sigma_{X|Z} = \mathrm{cov}\{X_{11} - E(X_{11}|Z)\}$.

For ease of presentation, for any random variable (vector) ζ, denote $\zeta - E(\zeta|Z)$ by $\widetilde{E}(\zeta|Z)$. For example, $\widetilde{W}_i = W_i - E(W_i|Z_i)$. Denote $\mathbf{Y} = (Y_1, \ldots, Y_n)^{\mathrm{T}}$, $\mathbf{Z} = (Z_1, \ldots, Z_n)^{\mathrm{T}}$, $\mathbf{X} = (X_1, \cdots, X_n)^{\mathrm{T}}$, and $\mathbf{W} = (W_1, \cdots, W_n)^{\mathrm{T}}$. Then, we have the following theorem whose proof is given in the Appendix.

Theorem 1 *Suppose that regularity conditions (a)—(e) in the Appendix hold. Then*

*(A) (***Convergence rate***) There exists a local maximizer $\hat{\boldsymbol{\beta}}$ of $\mathcal{L}_P(\Sigma_{uu}, \boldsymbol{\beta})$ defined in (7) such that its rate of convergence is $O_p(n^{-1/2} + a_n)$, where a_n is given in (8);*

*(B) (***Oracle property***) We further assume that $\lambda \to 0$, $n^{1/2}\lambda \to \infty$, and*

$$\liminf_{n\to\infty} \liminf_{u \to 0^+} p'_\lambda(u)/\lambda > 0. \tag{10}$$

The root n consistent estimator $\hat{\boldsymbol{\beta}}$ in (A) must satisfy that

*(a) (***Sparsity***) $\hat{\boldsymbol{\beta}}_2 = 0$; and*

*(b) (***Asymptotic normality***)*

$$\sqrt{n}(\Sigma_{X|Z} + \Sigma_\lambda)\{\widehat{\boldsymbol{\beta}}_1 - \boldsymbol{\beta}_{10} + (\Sigma_{X|Z} + \Sigma_\lambda)^{-1}\mathbf{b}\} \xrightarrow{D} N(0, \Gamma),$$

where $\Gamma = E\left\{\widetilde{X}_{11}(\varepsilon - U_{11}^{\mathrm{T}}\beta_{10}) + \varepsilon U_{11} + (U_{11}U_{11}^{\mathrm{T}} - \Sigma_{uu1})\beta_{10}\right\}^{\otimes 2}$.

In the absence of nonparametric component $\nu(Z)$, model (5) becomes model (2). Thus, under the conditions of Theorem 1, the resulting estimate of penalized least squares (3) possesses the rate of convergence $O_p(n^{-1/2} + a_n)$, and the oracle property: $\hat{\boldsymbol{\beta}}_2 = 0$, and

$$\sqrt{n}\{\text{cov}(X_{11}) + \Sigma_\lambda\}[\widehat{\boldsymbol{\beta}}_1 - \boldsymbol{\beta}_{10} + \{\text{cov}(X_{11}) + \Sigma_\lambda\}^{-1}\mathbf{b}] \xrightarrow{D} N(0, \Gamma_0),$$

where $\Gamma_0 = E\left\{X_{11}(\varepsilon - U_{11}^{\mathrm{T}}\boldsymbol{\beta}_{10}) + \varepsilon U_{11} + (U_{11}U_{11}^{\mathrm{T}} - \Sigma_{uu1})\boldsymbol{\beta}_{10}\right\}^{\otimes 2}$.

Condition (b) in the Appendix requires the bandwidths in estimating $m_w(\cdot)$ and $m_y(\cdot)$ are of order $n^{-1/5}$. All bandwidths with this rate lead to the same limiting distribution of $\hat{\boldsymbol{\beta}}$. Therefore the bandwidth selection can be done in a standard routine. In our implementation, we use the plug-in bandwidth selector proposed in Ruppert, Sheather, and Wand (1995) to select bandwidths for $m_w(\cdot)$ and $m_y(\cdot)$. In our empirical study, we have shifted bandwidths around the selected values, and found that the results were stable.

From Theorem 1, the asymptotic covariance of $\widehat{\boldsymbol{\beta}}_1$ is

$$\frac{1}{n}(\Sigma_{X|Z} + \Sigma_\lambda)^{-1}\Gamma(\Sigma_{X|Z} + \Sigma_\lambda)^{-1}.$$

Let X_1^* consist of the selected variable in the final model, and W_1^* be the corresponding vector of W-variable. Let $\dim(X_1^*)$ be the number of the components of X_1. A consistent estimate of $\Sigma_{X|Z}$ is defined as

$$\{n - \dim(X_1^*)\}^{-1}\sum_{i=1}^{n}\left\{W_{1i}^* - \widehat{E}(W_{i1}^*|Z_i)\right\}^{\otimes 2} - \Sigma_{uu} \stackrel{\text{def}}{=} \widehat{\Sigma}_{X|Z}.$$

Recall the function Ψ given in (6), one may construct a consistent sandwich–type estimate of Γ, namely

$$\widehat{\Gamma}_n = n^{-1}\sum_{i=1}^{n}\left\{\widetilde{W}_{i1}(\widetilde{Y}_i - \widetilde{W}_{i1}^{\mathrm{T}}\widehat{\boldsymbol{\beta}}_1) + \Sigma_{uu}\widehat{\boldsymbol{\beta}}_1\right\}^{\otimes 2}.$$

The covariance of the estimates $\hat{\boldsymbol{\beta}}_1$, the nonvanishing component of $\hat{\boldsymbol{\beta}}$, can be estimated by

$$n^{-1}\left\{\widehat{\Sigma}_{X|Z} + \Sigma_\lambda(\hat{\boldsymbol{\beta}}_1)\right\}^{-1}\widehat{\Gamma}_n\left\{\widehat{\Sigma}_{X|Z} + \Sigma_\lambda(\hat{\boldsymbol{\beta}}_1)\right\}^{-1}.$$

We next consider the case in which Σ_{uu} is unknown. The commonly used method of estimating Σ_{uu} (Carroll, et al., 1995, Chapter 3) is to assume that there are partial replicated observations, so

that we observe $W_{ij} = X_i + U_{ij}, \quad j = 1, ..., J_i$. Let $\overline{W}_i$ be the sample mean of the replicates. Then a consistent, unbiased method of moments estimate for Σ_{uu} is

$$\widehat{\Sigma}_{uu} = \sum_{i=1}^{n}\sum_{j=1}^{J_i}(W_{ij} - \overline{W}_i)^{\otimes 2} \Big/ \sum_{i=1}^{n}(J_i - 1).$$

Consequently the penalized least squares function is defined as

$$\mathcal{L}_P(\widehat{\Sigma}_{uu}, \boldsymbol{\beta}) = \frac{1}{2}\sum_{i=1}^{n}\{Y_i - \hat{m}_y(Z_i) - (\overline{W}_i - \hat{m}_{\overline{w}}(Z))^{\mathrm{T}}\boldsymbol{\beta}\}^2 - \frac{n}{2}\boldsymbol{\beta}^{\mathrm{T}}\widehat{\Sigma}_{uu}\boldsymbol{\beta} + n\sum_{j=1}^{d} p_{\lambda_n}(|\beta_j|), \quad (11)$$

Corollary 1 *Under the condition of Theorem 1, we still have the following conclusions.*

*(A) (***Convergence rate***) There exists a local maximizer $\hat{\boldsymbol{\beta}}$ of $\mathcal{L}_P(\widehat{\Sigma}_{uu}, \boldsymbol{\beta})$ defined in (11) such that its rate of convergence is $O_p(n^{-1/2} + a_n)$, where a_n is given in (8);*

*(B) (***Oracle property***) The root n consistent estimator $\hat{\boldsymbol{\beta}}$ in (A) must satisfy that*

*(a) (***Sparsity***) $\hat{\boldsymbol{\beta}}_2 = 0$,; and*

*(b) (***Asymptotic normality***)*

$$\sqrt{n}(\Sigma_{X|Z} + \Sigma_\lambda)\{\widehat{\boldsymbol{\beta}}_1 - \boldsymbol{\beta}_{10} + (\Sigma_{X|Z} + \Sigma_\lambda)^{-1}\mathbf{b}\} \xrightarrow{D} N(0, \Gamma^*),$$

where $\Gamma^* = E\left\{\widetilde{X}_{11}(\varepsilon - \overline{U}_{11}^{\mathrm{T}}\beta_{10}) + \varepsilon\overline{U}_{11} + (\overline{U}_{11}\overline{U}_{11}^{\mathrm{T}} - \Sigma_{uu1})\boldsymbol{\beta}_{10}\right\}^{\otimes 2}$, where $\overline{U}_{11}$ denotes the sample mean of U_{1j}s.

Because $\widehat{\Sigma}_{uu}$ is a consistent, unbiased method of moment estimator for Σ_{uu}, Corollary 1 can be proved in a similar way to Theorem 1 by replacing $\mathcal{L}_P(\Sigma_{uu}, \boldsymbol{\beta})$ by $\mathcal{L}_P(\widehat{\Sigma}_{uu}, \boldsymbol{\beta})$. To save space, we omit the details.

Standard error estimates can also be derived. A consistent estimate of $\Sigma_{X|Z}$ is defined as

$$\{n - \dim(X_{11})\}^{-1}\sum_{i=1}^{n}\left\{\overline{W}_{i1} - \widehat{E}(W_{i1}|Z_i)\right\}^{\otimes 2} - \frac{1}{2}\widehat{\Sigma}_{uu}.$$

Estimates of Γ^* can be also easily developed. Let

$$R_i = \widetilde{\overline{W}}_i(\widetilde{Y}_i - \widetilde{\overline{W}}_i^{\mathrm{T}}\widehat{\boldsymbol{\beta}}) + \widehat{\Sigma}_{uu}\widehat{\boldsymbol{\beta}}/J_i + \frac{\sum_{i=1}^{n} J_i^{-1}}{n(J_i - 1)}\left\{\frac{1}{2}(W_{i1} - W_{i2})^{\otimes 2} - \widehat{\Sigma}_{uu}\right\}\widehat{\boldsymbol{\beta}}.$$

Then a consistent estimate of Γ^* is the sample covariance matrix of the R_i's. See Liang, Härdle, and Carroll (1999) for a detailed discussion.

2.4 Issues related to practical implementation

In previous section, we propose variable selection procedures for linear measurement error models and partially linear measurement error models. There are two important issues related to practical implementation. One is how to optimize the penalized least squares in (3), (7) and (11) since the Newton-Raphson algorithm cannot be directly applied for these penalized least squares functions. The other one is how to select the regularization parameters λ_j. We next address these two issues.

Local quadratic approximation

Since penalty functions such as the smoothly clipped absolute deviation and the L_1 are singular at the origin, it is challenging to minimizing the penalized least squares functions. Following Fan and Li (2001), we will use a local quadratic approximation to the penalty function in our implementation. Suppose that we are given an initial value $\beta^{(0)}$ that is close to the true value of β. The penalty function is locally approximated by a quadratic function as

$$p_{\lambda_j}(|\beta_j|) \approx p_{\lambda_j}(|\beta_j^{(0)}|) + \frac{1}{2}\{p'_{\lambda_j}(|\beta_j^{(0)}|)/|\beta_j^{(0)}|\}(\beta_j^2 - \beta_j^{(0)2}).$$

When $\beta_j^{(0)}$ is very close to zero, the approximation is not stable. Hunter and Li (2005) proposed a perturbed version of the local quadratic approximation:

$$p_{\lambda_j}(|\beta_j|) \approx q_{\lambda_j}(|\beta_j|) \equiv p_{\lambda_j}(|\beta_j^{(0)}|) + \frac{1}{2}\{p'_{\lambda_j}(|\beta_j^{(0)}|)/(|\beta_j^{(0)}| + \eta)\}(\beta_j^2 - \beta_j^{(0)2}),$$

where η is a small positive number. Hunter and Li (2005) discussed how to determine the value of η in details. In our implementation, we adopt their strategy to choose η. With the aid of the local quadratic approximation, the Newton-Raphson algorithm can be applied to minimize the penalized least squares (3), (7) and (11). We set the unpenalized least squares estimate $\hat{\boldsymbol{\beta}}^u$ as the initial value of $\boldsymbol{\beta}$ since it is root n consistent by Theorem 1 with $\lambda_j = 0$.

Choice of regularization parameters

Similarly to Fan and Li (2001), we suggest selecting the tuning parameters λ_j's using generalized cross validation. Let $\ell(\boldsymbol{\beta})$ be either

$$\ell(\boldsymbol{\beta}) = \frac{1}{2}\sum_{i=1}^{n}(Y_i - W_i^{\mathrm{T}}\boldsymbol{\beta})^2 - n\boldsymbol{\beta}^{\mathrm{T}}\Sigma_{uu}\boldsymbol{\beta},$$

or

$$\ell(\boldsymbol{\beta}) = \frac{1}{2}\sum_{i=1}^{n}[Y_i - m_y(Z_i) - \{W_i - m_w(Z_i)\}^{\mathrm{T}}\boldsymbol{\beta}]^2 - \frac{n}{2}\boldsymbol{\beta}^{\mathrm{T}}\Sigma_{uu}\boldsymbol{\beta}$$

when Σ_{uu} is known, or

$$\ell(\boldsymbol{\beta}) = \frac{1}{2}\sum_{i=1}^{n}\{Y_i - \hat{m}_y(Z_i) - (\overline{W}_i - \hat{m}_{\overline{w}}(Z))^{\mathrm{T}}\boldsymbol{\beta}\}^2 - \frac{n}{2}\boldsymbol{\beta}^{\mathrm{T}}\widehat{\Sigma}_{uu}\boldsymbol{\beta}.$$

when Σ_{uu} is unknown. Define

$$\mathcal{L}_a(\boldsymbol{\beta}) = \ell(\boldsymbol{\beta}) + n\sum_{j=1}^{n} q_{\lambda_j}(|\beta_j|).$$

In the last step of the Newton-Raphson iteration, we may compute the effective number of parameters:

$$e(\lambda_1, \cdots, \lambda_d) = \mathrm{tr}[\{\mathcal{L}_a''(\hat{\boldsymbol{\beta}})\}^{-1}\ell''(\hat{\boldsymbol{\beta}})].$$

The generalized cross validation statistic is defined by

$$\mathrm{GCV}(\lambda_1, \cdots, \lambda_d) = \frac{\mathrm{RSS}_{\lambda_1,\cdots,\lambda_d}}{n\{1 - e(\lambda_1, \cdots, \lambda_d)/n\}^2},$$

where $\mathrm{RSS}_{\lambda_1,\cdots,\lambda_d}$ is the residual sum of squares corresponding to the model selected by penalized least squares with tuning parameters $\lambda_1, \cdots, \lambda_d$. The minimization problem over a d-dimensional space is difficult. However, it is expected that the magnitude of λ_j should be proportional to the standard error of the unpenalized least squares estimator of β_j. Denoted by $\hat{\beta}_j^u$. In practice, we suggest taking $\lambda_j = \lambda * SE(\hat{\beta}_j^u)$, where $SE(\hat{\beta}_j^u)$ is the estimated standard error of $\hat{\beta}_j^u$. Such a choice of λ_j works well from our simulation experience. Thus, the minimization problem will reduce to a one-dimensional problem, and the tuning parameter can be estimated by a grid search.

3 Empirical-likelihood-based inference

In the previous section we have derived estimators for the standard error of the resulting estimate. Confidence region/interval based on normal approximation or it bootstrap version becomes available. Although we have confirmed that the estimator of the standard errors given in Section 2 is

consistent, its finite-sample behavior may not be optimistic because we need to plug in several estimated terms and the resulting confidence intervals is symmetrical about the resulting estimate. This may not be true since the finite sample distribution may not be symmetric. An alternative method to avoid these weaknesses is empirical likelihood, which was originally proposed by Owen (1988, 90), and further contributed by Owen (1991), Qin (1994, 99), Qin and Lawless(1994), Chen (1993, 1994). Owen (2000) gave a comprehensive survey of the empirical likelihood and the related topics.

In this section, we use the empirical likelihood principle to develop confidence region to avoid estimating covariance matrix and using normal approximation. In the literature, empirical likelihood principle have been extensively recently because of its appealing advantages (see Owen, 2000). For example, the empirical likelihood approach has been used to construct confidence interval for linear models (Owen 1991, Chen 1993, 1994), for generalized linear models (Kolaczyk 1994), and for general estimating equation (Qin and Lawless 1994). The structure of measurement error data poses many challenges. Thus, our current statistical settings are much more complicated. So the generalization is not trivial. In this section, it is assumed that ε_i are independent and identically distributed and independent of (W_i, Z_i). We first propose our empirical likelihood ratio statistic, and show that its asymptotic distribution is a chi-squared distribution. In this section, we will not consider the task of variable selection and will work on the full model.

Let F be the distribution function which assigns probability p_i at points (W_i, Y_i, Z_i). Then $\sum_{i=1}^n p_i = 1$ and $p_i \geq 0$ for each i. Our semiparametric empirical likelihood ratio is defined as follows. Note that

$$E\{\widetilde{W}(\widetilde{Y} - \widetilde{W}^{\mathrm{T}}\boldsymbol{\beta}) + \Sigma_{uu}\boldsymbol{\beta}\} = 0.$$

The empirical likelihood ratio function for $\boldsymbol{\beta}$ may be defined by

$$\mathcal{R}(\boldsymbol{\beta}) = \max\left\{\prod_{i=1}^n np_i \;\Big|\; \sum_{i=1}^n p_i\left\{\widetilde{W}_i(\widetilde{Y}_i - \widetilde{W}_i^{\mathrm{T}}\boldsymbol{\beta}) + \Sigma_{uu}\boldsymbol{\beta}\right\} = 0, p_i \geq 0, \sum_{i=1}^n p_i = 1\right\},$$

if $m_w(x)$, $m_y(z)$, and Σ_{uu} are known. We first assume that Σ_{uu} are known. In our model setting, a

modified empirical likelihood ratio function is defined as

$$\mathcal{R}_n(\boldsymbol{\beta}) = \max\left\{\prod_{i=1}^{n} np_i \;\Big|\; \sum_{i=1}^{n} p_i\Big(\{W_i - \widehat{m}_w(Z_i)\}[Y_i - \widehat{m}_y(Z_i) - \{W_i - \widehat{m}_w(Z_i)\}^{\mathrm{T}}\boldsymbol{\beta}] + \Sigma_{uu}\boldsymbol{\beta}\Big) = 0, p_i \geq 0, \sum_{i=1}^{n} p_i = 1\right\} \tag{12}$$

Theorem 2 *Under regularity conditions given in the Appendix and that the third moments of the elements of X, ε and U are finite, $-2\log\{\mathcal{R}_n(\boldsymbol{\beta})\}$ is asymptotically chi-squared distributed with degree freedom p.*

Based on this result, a confidence region of $\boldsymbol{\beta}$ can be given as $\{\boldsymbol{\beta} : -2\log\{\mathcal{R}_n(\boldsymbol{\beta})\} \leq c_\alpha\}$, where c_α denotes the α quantile of the chi-squared distribution. When Σ_{uu} is unknown, replace W_i by $\overline{W}_i$ and $\Sigma_{uu}\boldsymbol{\beta}$ by $1/2\widehat{\Sigma}_{uu}\boldsymbol{\beta}$. Theorem 2 still holds.

4 Simulation study and application

In this section, we investigate the finite sample performance of the proposed procedure by Monte Carlo simulation, and illustrate the proposed methodology by an analysis of a real data set.

4.1 Prediction error and generalize mean squared error

The prediction error is defined as the average error in the prediction of the dependent variable given the independent variables for future cases that are not used in the construction of a prediction equation. Let $\{X^*, Y^*, Z^*\}$ be a new observation from a regression model with mean $\mu(X, Z)$. The prediction error (PE) of a predictor $\hat{\mu}(X, Z)$ is defined as

$$\mathrm{PE}(\hat{\mu}) = E_*\{Y^* - \hat{\mu}(X^*, Z^*)\}^2$$

where the expectation E_* is a conditional expectation given the data used in constructing the prediction procedure. The PE can be decomposed as

$$\mathrm{PE}(\hat{\mu}) = E_*\{Y^* - \mu(X^*, Z^*)\}^2 + E_*\{\hat{\mu}(X^*, Z^*) - \mu(X^*, Z^*)\}^2.$$

The first component is the inherent prediction error due to noise. The second one is due to lack of fit with an underlying model. This component is termed *model error* (ME). For the linear model (2), the $\hat{\mu} = X^{\mathrm{T}}\hat{\boldsymbol{\beta}}$, and hence, the ME equals $(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})^{\mathrm{T}} E_* X^* X^{*\mathrm{T}} (\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})$. For the partial linear model (5), the ME of $\hat{\mu}(X, Z) = X^{\mathrm{T}}\hat{\boldsymbol{\beta}} + \hat{\nu}(Z)$ can be further decomposed

$$\begin{aligned} \mathrm{ME}(\hat{\mu}) \equiv \mathrm{ME}(\hat{\boldsymbol{\beta}}, \hat{\nu}) &= (\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})^{\mathrm{T}} E_* X^* X^{*\mathrm{T}} (\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}) + E_*\{\hat{\nu}(Z^*) - \nu(Z^*)\}^2 \\ &\quad + 2E_*\{\hat{\nu}(Z^*) - \nu(Z^*)\} X^{*\mathrm{T}} (\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}). \end{aligned} \tag{13}$$

The first term and the second term in (13) measures how well the parametric component and non-parametric component fit, respectively. The third term vanishes when Z^* and X^* are independent and $EX^* = 0$. Thus, the third term equals zero in our simulation setting. Therefore, we will summarize our simulation results using *generalized mean squared error* (GMSE) defined by

$$\mathrm{GMSE}(\hat{\boldsymbol{\beta}}) = (\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})^{\mathrm{T}} E_* X^* X^{*\mathrm{T}} (\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}).$$

4.2 Simulation studies

Example 1 We generate 1000 data sets, each consisting of $n = 200$ random samples, from the following linear measurement error model:

$$\begin{cases} Y &= X^{\mathrm{T}}\boldsymbol{\beta} + \varepsilon, \\ W &= X + U, \end{cases}$$

where $\boldsymbol{\beta} = (0, 2, 1, 0, 0, 0, 1.5, 0)^{\mathrm{T}}$, $\varepsilon \sim N(0, 1)$, and X is 8-dimensional normal distribution with mean zero and covariance matrix $(\sigma_{ij})_{8\times 8}$ with $\sigma_{ij} = 0.5^{|i-j|}$. In our simulation, $m = 2$, and the first 5 components of X are measured with errors $U \sim N(\mathbf{0}, \Sigma_{uu})$ and $EU_iU_j = 0.6$ for $i = j$ and 0.12 for $i \neq j$. The last three components of X are measured without errors. To estimate Σ_{uu}, two replicates of W are generated.

In this example, we also investigate the impact of ignoring the measurement error when the covariates are measured with error. Define relative generalized mean squared error (RGMSE) to be the ratio of GMSE of an underlying model to that of the full model taking account into the measurement error. Table 1 gives the median and median of absolute deviation of RGMSE

Table 1: Simulation Results for Example 1

Model	RGMSE	Zero Coefficient		RGMSE	Zero Coefficient	
	Median(MAD)	C	I	Median(MAD)	C	I
	Incorporate measurement error			Ignore measurement error		
Full	1.0000(0)	0	0	2.0705(1.4072)	0	0
SCAD	0.3931(0.2456)	4.501	0.005	1.8306(1.2757)	3.621	0
L_1	0.9235(0.7753)	4.410	0.016	2.2196(1.5272)	2.608	0
AIC	0.6269(0.3151)	3.400	0	2.0454(1.3725)	2.771	0
BIC	0.5238(0.2853)	4.189	0	1.9439(1.3593)	3.982	0
RIC	0.5604(0.3038)	3.969	0	1.9843(1.3551)	3.648	0
Oracle	0.3357(0.2509)	5.000	0	1.7224(1.2504)	5.000	0

over 1000 simulations. The average number of zero coefficients demonstrates how the proposed procedure reduces model complexity and is reported in Table 1, in which the column labeled 'C' stands for the average restricted only to the true zero coefficients, while the column label 'I' depicts the average of coefficients erroneously estimated as 0. In Table 1, the row with label 'Full' corresponds to the full model, SCAD, L_1, AIC, BIC, and RIC stand for the penalized least squares procedures with the smoothly clipped absolute deviation, L_1, AIC, BIC and RIC penalties, as defined in Section 2, respectively, and 'Oracle' for the oracle procedure. Since the entropy penalty is discontinuous, the solutions for AIC, BIC, and RIC are obtained by exhaustively searching over all possible subsets. Thus, the resulting subsets are the best subsets for the corresponding criterion, and the computational cost for these procedures is much more expensive than that for the smoothly clipped absolute deviation and the L_1 methods. Table 1 shows that model fitting and model complexity taking account into measurement error are much better than those ignoring measurement error. From Table 1, we can see that the SCAD outperforms the other variable selection procedures in terms of model error and model complexity. Furthermore, its RGMSE is very close to that of the oracle estimator, which is consistent with the result in Theorem 1, and the method reduces the model complexity almost as effectively as the oracle procedure.

Table 2: Bias, Standard Deviation and Standard Error

	β_2				β_7			
Model	bias	Std	SE(std(SE))	CP(95%)	bias	Std	SE(std(SE))	CP
Incorporate measurement error								
SCAD	.0280	.2073	.1787(.0298)	.920	-.0009	.1256	.1197(.0117)	.934
Oracle	.0198	.1935	.1768(.0294)	.931	-.0009	.1214	.1180(.0114)	.939
Ignore measurement error								
SCAD	-.4757	.1127	.1074(.0111)	.009	.0217	.1200	.1111(.0113)	.938
Oracle	-.4510	.1074	.1053(.0108)	.012	.0249	.1106	.1085(.0105)	.943

We next test the accuracy of the standard error formula proposed in Section 2. The standard deviation of the 1000 estimated coefficients from the 1000 simulations can be regarded as the true standard error except for Monte Carlo error. The mean and standard deviation of the 1000 estimated standard errors gauge the overall performance of the standard error formula. The coverage probability (CP) indicates how accurate the confidence interval is. In Table 2, we present only the simulation results of $\hat{\beta}_2$ and $\hat{\beta}_7$ using the SCAD and the oracle procedure. For other cases, the results are similar. From Table 2, the difference between the true standard error and the mean of the estimated standard errors is less than one standard deviation of the estimated standard errors. This implies that the proposed standard error formula is accurate. The coverage probability (CP) of 95% confidence interval of $\hat{\beta}_j$ is also presented. Table 2 also depicts the bias of the 1000 estimate coefficients from the 1000 simulations. Taking account into measurement error, the biases for both $\hat{\beta}_2$ and $\hat{\beta}_7$ are close to zero. As expected, ignoring measurement error yields a significant bias for $\hat{\beta}_2$, and therefore, its 95% CP has very poor coverage probability since x_2 is measured with additive measurement error. The bias for $\hat{\beta}_7$ is still small because x_7 is measured without an error.

Example 2. In this example we simulate 1000 data sets, each consists of $n = 400$ random samples, from the partially linear measurement error model

$$Y \;=\; \nu(Z) + X^{\mathrm{T}}\boldsymbol{\beta} + \varepsilon,$$

Table 3: Simulation Results for Example 2 ($n = 400$)

Model	Homogeneous Error			Heteroscedastic Error		
	RGMSE	Zero Coefficient		RGMSE	Zero Coefficient	
	Median(MAD)	C	I	Median(MAD)	C	I
SCAD	0.4659(0.1962)	4.167	0.001	0.4699(0.2035)	4.260	0
L_1	0.6241(0.2978)	3.383	0.001	0.6037(0.2893)	3.415	0
AIC	0.6971(0.2100)	3.336	0	0.7097(0.2181)	2.526	0
BIC	0.5380(0.2146)	4.339	0	0.6945(0.2157)	2.746	0
RIC	0.5970(0.2145)	4.006	0	0.7026(0.2181)	2.651	0
Oracle	0.3360(0.1829)	5.000	0	0.3640(0.1893)	5.000	0

$$W \;=\; X + U,$$

where $Z \sim$ uniform$(0, 1)$, $\nu(z) = 2\sin(2\pi z^3)$, $\boldsymbol{\beta} = (0, 1.5, 0.75, 0, 0, 0, 1, 0)^{\mathrm{T}}$, and the others are the same as in Example 1. In our simulation study, we use local linear regression to estimate $\nu(\cdot)$ with $K(u) = 0.75(1 - u^2)I_{(|u|\le 1)}$. The bandwidth is selected using the plug-in method (Ruppert, Sheather and Wand, 1995).

In this example, we consider two scenarios: (i) ε follows $N(0, 1)$, i.e., homogeneous error; and (ii) the error follows $|\sin\{2\pi(X^{\mathrm{T}}\boldsymbol{\beta})^2 + 0.2Z\}|(\mathcal{X}_2^2 - 2)$, where $\mathcal{X}_2^2$ denotes the chi-squared distribution of 2 degree of freedom. This case is meant to see the effect of asymmetric and heteroscedastic error on the estimators and confidence intervals. The simulated results are summarized in Table 3, in which notation is the same as that in Table 1. From Table 3, we can see that the SCAD outperforms the other variable selection procedures in terms of model error and model complexity. We have also tested the accuracy of the standard error formula proposed in Section 2. To save space, we opt not to present the results here. In general, the sandwich formula gives us accurate estimates of standard errors and the coverage probabilities which are close to the nominal level.

We next compare the length and coverage probability of confidence intervals (CI) constructed by traditional sandwich formula and by empirical likelihood based method. Table 4 presents results

Table 4: Confidence Intervals

	β_2		β_3		β_7	
	95%CI	95% CP	95%CI	95% CP	95%CI	95% CP
Homogeneous Error						
Convention	(1.2695, 1.7649)	0.929	(0.5147, 0.9847)	0.930	(0.8055, 1.1704)	0.945
EL	(1.4103, 1.6571)	0.946	(0.6516, 0.8620)	0.934	(0.9054, 1.0963)	0.956
Heteroscedastic Error						
Convention	(1.3274, 1.7079)	0.903	(0.5804, 0.9256)	0.901	(0.8581, 1.1265)	0.944
EL	(1.4316, 1.6314)	0.962	(0.6791, 0.8358)	0.940	(0.9328, 1.0675)	0.948

for the non-zero coefficients. In Table 4, The row with label 'convention' corresponds to the CI constructed by sandwich formula proposed in Section 2. The row with label 'EL' corresponds to the CI constructed by empirical likelihood method. Compared with the traditional ones, Table 4 shows the empirical likelihood based CI is more accurate, and have better coverage probability. Table 4 also indicates that the empirical likelihood based CI has more gain for coefficients (such as β_2, β_3) of variables measured with error than those of variables (such as β_7) measured without error.

4.3 An application

Now we illustrate the proposed procedures by an analysis of a real data from nutritional epidemiology. This data set has been analyzed in Carroll et al. (1998). In nutrition research, the assessment of an individual's usual intake diet is difficult but important in studying the relation between diet and cancer and in monitoring dietary behavior. Food Frequency Questionnaires (FFQ) are frequently administered. FFQ's are thought to often involve a systematic bias (i.e., under- or over-reporting at the level of the individual).

Two commonly used instruments are the 24-hour food recall and the multiple-day food record

Table 5: Estimated coefficients and standard errors for the NHS data set

Variable	Full Model	EL-based 95%CI	Selected Model
x_1	-0.0200 (0.0162)	(-0.0482, 0.0570)	-0.0217 (0.0171)
x_2	0.2694 (0.0248)	(0.2523, 0.3183)	0.2729 (0.0252)
x_3	-0.0551 (0.0232)	(-0.1744, 0.0894)	-0.0536 (0.0242)
x_4	0.0420 (0.0284)	(0.0830, 0.1286)	0.0349 (0.0284)
x_2^2	-0.0264 (0.0170)	(-0.1118, 0.0411)	-0.0224 (0.0146)
x_2x_3	0.0060 (0.0227)	(-0.0816, 0.1639)	0 (0.0000)
x_2x_4	0.0298 (0.0285)	(-0.0523, 0.0310)	0.0117 (0.0082)
x_3^2	-0.0203 (0.0134)	(-0.0644, -0.0095)	-0.0123 (0.0095)
x_3x_4	0.0297 (0.0324)	(-0.0686, 0.2467)	0 (0.0000)
x_4^2	-0.0216 (0.0103)	(-0.0823, -0.0112)	-0.0174 (0.0092)

(FR). Each of these FR's is more work-intensive and more costly, but is thought to involve considerably less bias than a FFQ.

We consider data from the Nurses' Health Study (NHS), which is a calibration study of size $n = 168$ woman, all of whom completed a single FFQ and four multiple-day food diaries ($m = 4$ in our notation). Other variables from this study include 4-day energy and Vitamin A (VA), body mass index (BMI) and age. Denote

y: the intake of a nutrient reported on a FFQ,

x_1 : the true Vitamin A intake,

x_2: the long-term usual intake,

x_3 : the energy intake,

x_4 : Body mass index,

z: age.

Because of measurement errors, x_1 is not directly observable. Four replicates of x_1, denoted by $w_1, \cdots, w_4$, are collected for each subject. Of interest here is to investigate the relation between

FFQ and FR, and other four factors. As an illustration, the following partially linear model with measurement error is considered:

$$y = \nu(z) + \sum_{k=1}^{4} \beta_k x_k + \sum_{u=2}^{4} \sum_{v=u}^{4} \beta_{uv} x_u x_v + \varepsilon, \tag{14}$$

and for each subject,

$$w_j = x_1 + u_j, \quad j = 1, \cdots, 4$$

are observed. The data were fitted by this model. The estimated coefficients and their standard errors are reported in the 2nd column of Table 5. Furthermore, the 95% empirical likelihood based confidence intervals of the coefficients are shown in the 3rd column of Table 5. We next apply the SCAD variable selection procedure to the model. Since linear component in (14) is of hierarchical order. In the implementation of the SCAD variable selection procedure. We do not penalize the linear effects β_j, $j = 1, \cdots, 4$, but penalize all other β's. The selected λ is 0.7243 by minimizing the GCV scores. With the selected λ, the final selected model is presented in the 4th column of Table 5, from which it can be seen that the selected model excludes two interaction terms, x_2x_3 and x_3x_4 from the full model.

After estimated the coefficient βs, one can easily estimate $\nu(z)$ by smoothing partial residuals $y_i - \mathbf{x}_i^T \hat{\boldsymbol{\beta}}$ over z_i since it is a 1-dimensional smoothing problem. The direct plug-in bandwidth selector proposed by Ruppert, Sheather and Wand (1995) is used to select a bandwidth for $\nu(z)$. The selected bandwidth equals 4.7726, and the estimated $\nu(z)$ and its 95% pointwise confidence interval are depicted in Figure 1, from which the effect of age seems not significant as the curve looks like a constant over age.

5 Discussion

We proposed a class of variable selection procedures for linear measurement error models and partially linear measurement error models. The procedures are derived by using the profile-kernel method (Speckman, 1988) to presmooth $m_w(z)$ and $m_y(z)$ for the penalized least squares objective function. An alternative method in estimation and inference of partially linear models is backfitting

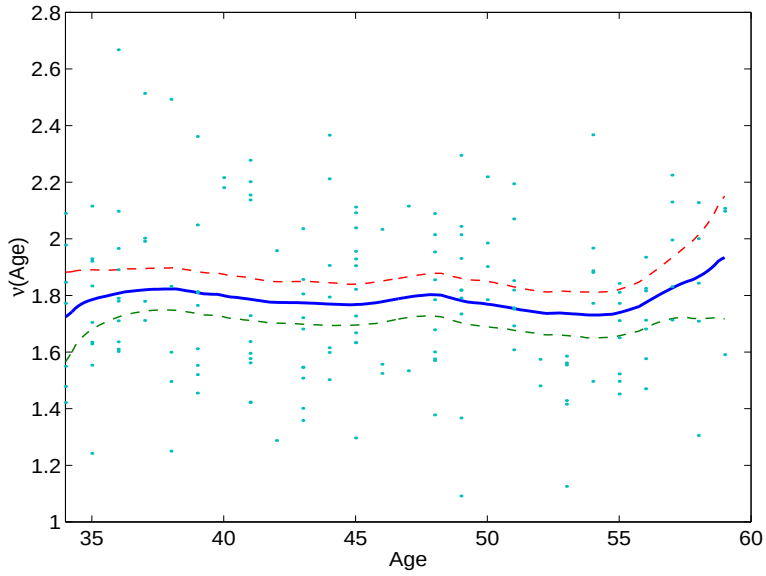

Figure 1: Estimated curve and its confidence interval of the nonparametric component based on the final model and a consideration of measurement error for the NHS data. The dots are the partial residuals $r_i = y_i - X_i^{\mathrm{T}}\hat{\boldsymbol{\beta}}$.

(Opsomer and Ruppert, 1999). How to give a penalized objective function based on backfitting approach is of interest and needs a further study. It is of interest to extend the proposed methodology for generalized partially linear measurement error model:

$$g\{E(Y|X,Z)\} = \nu(Z) + X^{\mathrm{T}}\boldsymbol{\beta},$$

where g is a link function. The study of this topic requires additional efforts, but it is beyond the scope of this article.

Appendix

To prove Theorems 1 and Corollary 1, we need the following regularity conditions.

(a) $\Sigma_{X|Z}$ is a positive-definite matrix, $E(\varepsilon|X,Z) = 0$, and $E(|\varepsilon|^3|X,Z) < \infty$;

(b) The bandwidths in estimating $m_w(z)$ and $m_y(z)$ are of order $n^{-1/5}$;

(c) $K(\cdot)$ is a bounded symmetric density function with compact support and satisfies that $\int K(u)du = 1$, $\int K(u)udu = 0$ and $\int u^2K(u)du = 1$;

(d) The density function of Z, $f_Z(z)$, and the density function of (Y, Z) are bounded away from 0 and have bounded continuous second derivative;

(e) $m_y(z)$ and $m_w(z)$ have bounded and continuous second derivatives;

We first point out a fact, which is assured by Conditions (b)-(e), that the local polynomial regression yields estimates of $m_y(\cdot)$ and $m_w(\cdot)$ satisfied with

$$\sup_z |\hat{m}_y(z) - m_y(z)| = o_p(n^{-1/4}), \text{ and } \sup_z |\hat{m}_{w,j}(z) - m_{w,j}(z)| = o_p(n^{-1/4}) \tag{A.1}$$

for $j = 1, \cdots, d$, where $m_{w,j}(\cdot)$ and $\hat{m}_{w,j}(\cdot)$ are the j-th component of $m_w(\cdot)$ and $\hat{m}_w(\cdot)$. See Mack and Silverman (1982) for a detailed discussion of availability of (A.1), which will repeatedly be used in our proof.

Lemma 1 Assume random variables $a_i(W_i, Z_i, Y_i)$ and $b_i(X_i, Z_i, Y_i)$, denoted by a_i and b_i, satisfy $Ea_i = 0$ and $\|b_i\| = o_p(n^{-1/4})$. Then

$$\sum_{i=1}^n a_i b_i \xi_i = o_p(n^{1/2}),$$

where ξ_i are independent variables with zero conditional mean and finite variance.

Proof. Denote the event $\|b_i\| \leq c_n n^{-1/4}$ by Ω_i, where c_n is a sequence of constants converges to infinity at a slow enough rate. Note that

$$P\left\{\left|\sum_{i=1}^n a_i b_i \xi_i\right| > \eta_n n^{-1/2}\right\} \leq P\left\{\left|\sum_{i=1}^n a_i b_i \xi_i\right| > \eta_n n^{-1/2}, \|b_i\| \leq c_n n^{-1/4}\right\} + P\left\{\|b_i\| > c_n n^{-1/4}\right\}.$$

The second term is $o_p(1)$. For the first term, using Chebyschev's inequality yields that

$$\begin{aligned} P\left\{\left|\sum_{i=1}^n a_i b_i \xi_i\right| > \eta_n n^{-1/2}, \|b_i\| \leq c_n n^{-1/4}\right\} &\leq \eta_n^{-2} n^{-1} \sum_{i=1}^n E[a_i \xi_i b_i \{I(\Omega_i) = 1\}]^2 \\ &+ \eta_n^{-2} n^{-1} \sum_{i \neq k}^n E[a_i a_k \xi_i \xi_k b_i b_k \{I(\Omega_k) = 1 \forall k\}]. \end{aligned} \tag{A.2}$$

Because $|b_i|\{I(\Omega_i) = 1\} \leq c_n n^{-1/4}$ is independent of $a_i \xi_i$, the first term in (A.2) is of order $O(\zeta_n^{-2} c_n^2 n^{-1/2}) = o_p(1)$. The second term in (A.2) is easily shown to equal zero.

Lemma 2 Under the condition of Theorem 1,

$$\mathcal{L}_P(\Sigma_{uu}, \boldsymbol{\beta}) - \Psi\{m_w(\cdot), m_y(\cdot), \boldsymbol{\beta}\} = o_p(1)$$

holds uniformly in $\boldsymbol{\beta} \in U(\boldsymbol{\beta}_0, \delta)$, a neighbor of the true value, denoted by $\boldsymbol{\beta}_0$, of $\boldsymbol{\beta}$. The solution of (7) is therefore equivalent to the solution of the penalized likelihood based on the penalized function $\mathcal{L}_P(m_w, m_y, \boldsymbol{\beta})$.

Proof. Note that $\mathcal{L}_P(\boldsymbol{\beta}) - \Psi\{m_w(\cdot), m_y(\cdot), \boldsymbol{\beta}\}$ equals

$$\sum_{i=1}^{n} 2[\{\widehat{m}_y(Z_i) - m_y(Z_i)\} - \{\widehat{m}_w(Z_i) - m_w(Z_i)\}^{\mathrm{T}}\boldsymbol{\beta}][Y_i - m_y(Z_i) - \{W_i - m_w(Z_i)\}^{\mathrm{T}}\boldsymbol{\beta}]$$
$$+\sum_{i=1}^{n}[\{\widehat{m}_y(Z_i) - m_y(Z_i)\} - \{\widehat{m}_w(Z_i) - m_w(Z_i)\}^{\mathrm{T}}\boldsymbol{\beta}]^2,$$

which equals to $o_p(1)$ by (A.1) and Lemma 1. The proof follows.

A.1 Proof of Theorem 1

Let $\alpha_n = n^{-1/2} + a_n$. Due to Lemma 2, we need only to show that for any given $\zeta > 0$, there exists a large constant C such that

$$P\left\{\inf_{\|\mathbf{v}\|=C} \mathcal{L}_P(m_w, m_y, \boldsymbol{\beta}_0 + \alpha_n\mathbf{v}) > \mathcal{L}_P(m_w, m_y, \boldsymbol{\beta}_0)\right\} \geq 1 - \zeta, \tag{A.3}$$

where $\|\mathbf{v}\|$ denotes the Euclidean norm for the vector $\mathbf{v}$. Note that

$$\begin{aligned} D_n(\mathbf{v}) &\equiv \mathcal{L}_P(m_w, m_y, \boldsymbol{\beta}_0 + \alpha_n\mathbf{v}) - \mathcal{L}_P(m_w, m_y, \boldsymbol{\beta}_0) \\ &\geq \sum_{i=1}^{n}\left[\left\{\widetilde{Y}_i - \widetilde{W}_i^{\mathrm{T}}(\boldsymbol{\beta}_0 + \alpha_n\mathbf{v})\right\}^2 - \left(\widetilde{Y}_i - \widetilde{W}_i^{\mathrm{T}}\boldsymbol{\beta}_0\right)^2\right] \\ &\quad -n\{(\boldsymbol{\beta}_0 + \alpha_n\mathbf{v})^{\mathrm{T}}\Sigma_u(\boldsymbol{\beta}_0 + \alpha_n\mathbf{v}) - \boldsymbol{\beta}_0^{\mathrm{T}}\Sigma_u\boldsymbol{\beta}_0\} \\ &\quad +n\sum_{j=1}^{s}\{p_{\lambda_n}(|\beta_{j0} + \alpha_n v_j|) - p_{\lambda_n}(|\beta_{j0}|)\}, \end{aligned}$$

where s is the number of components of $\boldsymbol{\beta}_{10}$.

The first two terms, denoted by $J_n(\mathbf{v})$, can be expressed as

$$J_n(\mathbf{v}) = \sum_{i=1}^{n}\left(-2\alpha_n\widetilde{Y}_i\widetilde{W}_i^{\mathrm{T}}\mathbf{v} + 2\alpha_n\widetilde{W}_i^{\mathrm{T}}\boldsymbol{\beta}_0\mathbf{v}^{\mathrm{T}}\widetilde{W}_i + \alpha_n^2\widetilde{W}_i^{\mathrm{T}}\mathbf{v}\mathbf{v}^{\mathrm{T}}\widetilde{W}_i\right) - n\alpha_n\boldsymbol{\beta}_0^{\mathrm{T}}\mathbf{v} - n\alpha_n\mathbf{v}^{\mathrm{T}}\boldsymbol{\beta}_0 - n\alpha_n^2\mathbf{v}^{\mathrm{T}}\mathbf{v}.$$

Recall that $E\widetilde{Y}_i = E\widetilde{W}_i = E(\widetilde{W}_i + \widetilde{Y}_i) = 0$. It follows from the Central Limiting Theorem that $J_n(\mathbf{v}) = O_p(n^{1/2}\alpha_n) + O_p(n\alpha_n^2)$. The third term can be expressed as

$$\sum_{j=1}^{s} \left[n\alpha_n p'_{\lambda_n}(|\beta_{j0}|)\text{sgn}(\beta_{j0})v_j + n\alpha_n^2 p''_{\lambda_n}(|\beta_{j0}|)v_j^2\{1+o(1)\} \right].$$

This term is bounded by

$$\sqrt{s}n\alpha_n a_n\|\mathbf{v}\| + n\alpha_n^2 \max\{|p''_{\lambda_n}(|\beta_{j0}|) : \beta_{j0} \neq 0\}\|\mathbf{v}\|^2 = o_p(1).$$

A combination of these results yields that $D_n(\mathbf{v}) = O_p(n^{1/2}\alpha_n) + O_p(n\alpha_n^2) + o_p(1)$. We therefore complete the proof of (A).

We will follow the same strategy of Fan and Li (2001) to prove the sparsity, it suffices to show that for any given $\boldsymbol{\beta}_1$ satisfying $\|\boldsymbol{\beta}_1 - \boldsymbol{\beta}_{10}\| = O_p(n^{-1/2})$ and $j = s+1, \ldots, d$, such that $\frac{\partial \mathcal{L}_P(\Sigma_{uu},\beta)}{\partial \beta_j} > 0$ for $0 < \beta_j < \epsilon_n = Cn^{-1/2}$, and $\frac{\partial \mathcal{L}_P(\Sigma_{uu},\beta)}{\partial \beta_j} < 0$ for $-\epsilon_n < \beta_j < 0$.

By the Taylor expansion, we have

$$\frac{\partial \mathcal{L}_P(\Sigma_{uu},\beta)}{\partial \beta_j} = -2[Y_i - \widehat{m}_y(Z_i) - \{W_i - \widehat{m}_w(Z_i)\}^{\mathrm{T}}\beta]\{W_i - \widehat{m}_w(Z_i)\}_j + np'_{\lambda_j}(|\beta_j|)\text{sgn}(\beta_j).$$

The first term can be shown to be $O_p(n^{-1/2})$ by (A.1) and that $\boldsymbol{\beta}_1 - \boldsymbol{\beta}_{10} = O_p(n^{-1/2})$. Recall that $n^{1/2}\lambda \to \infty$, and $\liminf_{n\to\infty} \liminf_{u\to 0^+} p'_\lambda(u)/\lambda > 0$. We know that the sign of $\frac{\partial \mathcal{L}_P(\Sigma_{uu},\beta)}{\partial \beta_j}$ is determined by that of β_j, and is negative for $0 < \beta_j < \epsilon_n$ and positive for $-\epsilon_n > \beta_j > 0$. It follows that $\hat{\boldsymbol{\beta}}_2 = 0$.

We will use results in Newey (1994) to show the asymptotic normality of $\hat{\boldsymbol{\beta}}_1$. In what follows, we denote by $S_{\star 1}$ the elements of $S_\star$ with respect to β_{10} for any random/function vector $S_\star$. Note that the estimators $\widehat{\beta}_1$ based on the penalized likelihood function given in (7) is equivalent to the solution of the estimating equation:

$$\sum_{i=1}^{n} \Phi(\widehat{m}_{w1}, \widehat{m}_y, \boldsymbol{\beta}_1, Y_i, W_{i1}, Z_i) - n\boldsymbol{\zeta}_1 = 0, \tag{A.4}$$

where $\Phi(m_{w1}, m_y, \boldsymbol{\beta}_1, Y, W_1, Z) = \{W - m_{w1}(Z)\}\left[Y - m_y(Z) - \{W_1 - m_{w1}(Z)\}^{\mathrm{T}}\boldsymbol{\beta}_1\right] - \Sigma_{uu1}\boldsymbol{\beta}_1$ and $\boldsymbol{\zeta}_1 = \{p'_\lambda(|\boldsymbol{\beta}_1|)\text{sgn}(\boldsymbol{\beta}_1), \ldots, p'_\lambda(|\boldsymbol{\beta}_s|)\text{sgn}(\boldsymbol{\beta}_s)\}^{\mathrm{T}}$.

It follows from (A.1) that $\|\widehat{m}_{w1}(\bullet) - m_{w1}(\bullet)\| = o_p(n^{-1/4})$ and $\widehat{m}_y(\bullet) - m_y(\bullet) = o_p(n^{-1/4})$. Thus, Assumption 5.1(ii) in Newey (1994) holds. Let

$$D(m^*_{w1} - m_{w1}, m^*_y - m_y) = \frac{\partial \Phi}{\partial m_{w1}}(m^*_{w1} - m_{w1}) + \frac{\partial \Phi}{\partial m_y}(m^*_y - m_y).$$

where $\frac{\partial \Phi}{\partial m_{w1}}$ and $\frac{\partial \Phi}{\partial m_y}$ are the Frechet derivatives. A direct but cumbersome calculation derives that $E\left(\frac{\partial \Phi}{\partial m_{w1}}\right) = 0$ and $E\left(\frac{\partial \Phi}{\partial m_y}\right) = 0$. Furthermore,

$$\begin{aligned} &\|\Phi(m^*_{w1}, m^*_y, \boldsymbol{\beta}_1, Y, W_1, Z) - \Phi(m_{w1}, m_y, \boldsymbol{\beta}_1, Y, W_1, Z) - D(m^*_{w1} - m_{w1}, m^*_y - m_y)\| \\ &= O_p\left(\|m^*_{w1} - m_{w1}\|^2 + \|m^*_y - m_y\|^2\right). \end{aligned} \tag{A.5}$$

This indicates that Assumption 5.1(i) in Newey (1994) is valid. Assumption 5.2 in Newey (1994) holds by the expression of $D(\cdot, \cdot)$. In addition, it follows from the above statements that for any (m^*_{w1}, m^*_y),

$$E\left\{D(m^*_{w1} - m_{w1}, m^*_y - m_y)\right\} = 0,$$

and Newey's Assumption 5.3, $\alpha(T) = 0$, holds. By Lemma 5.1 in Newey (1994), it follows that $\hat{\boldsymbol{\beta}}_1$ has the same distribution as the solution to the equation

$$0 = \sum_{i=1}^{n} \Phi(m_{w1}, m_y, \boldsymbol{\beta}_1, Y_i, W_{i1}, Z_i) - n\boldsymbol{\zeta}_1. \tag{A.6}$$

A direct simplification yields that

$$\widetilde{X}_{11}\widetilde{X}_{11}^{\mathrm{T}}\sqrt{n}(\widehat{\boldsymbol{\beta}}_1 - \boldsymbol{\beta}_{01}) + n^{1/2}\mathbf{b} = n^{-1/2}\sum_{i=1}^{n}\{\widetilde{X}_{i1}(\varepsilon_i - U_i^{\mathrm{T}}\boldsymbol{\beta}_{10}) + U_i\varepsilon_i + (U_iU_i^{\mathrm{T}} - \Sigma_{uu1})\boldsymbol{\beta}_{10}\} + o_p(1).$$

This completes the proof.

A.2 Proof of Theorem 2

Denote $\boldsymbol{\Xi}_i = \{W_i - \widehat{m}_w(Z_i)\}[Y_i - \widehat{m}_y(Z_i) - \{W_i - \widehat{m}_w(Z_i)\}^{\mathrm{T}}\boldsymbol{\beta}] + \Sigma_{uu}\boldsymbol{\beta}$. By the standard argument presented in Owen (2000), it follows that

$$p_i = \frac{1}{n(1 + \mathbf{a}^{\mathrm{T}}\boldsymbol{\Xi}_i)} \text{ for } i = 1, \cdots, n, \tag{A.7}$$

where $\mathbf{a} = (a_1, \ldots, a_n)^{\mathrm{T}}$ is the solution of the equation

$$\frac{1}{n}\sum_{i=1}^{n}\frac{\boldsymbol{\Xi}_i}{1+\mathbf{a}^{\mathrm{T}}\boldsymbol{\Xi}_i} = 0. \tag{A.8}$$

Using the same strategy as the proof of Theorem 3.2 in Owen (2000), we have

$$\|\mathbf{a}\| = O_p(n^{-1/2}). \tag{A.9}$$

On the other hand, based on the assumptions, the result of Theorem 1, and the strong law of large number, we have

$$\max_{1\le i\le n}\|\boldsymbol{\Xi}_i\| = o_p(n^{1/2}). \tag{A.10}$$

It follows from $\sum_{i=1}^{n} p_i = 1$ and (A.8) that

$$\frac{1}{n}\sum_{i=1}^{n}\frac{\boldsymbol{\Xi}_i}{1+\mathbf{a}^{\mathrm{T}}\boldsymbol{\Xi}_i} = \frac{1}{n}\sum_{i=1}^{n}\boldsymbol{\Xi}_i(1-\mathbf{a}^{\mathrm{T}}\boldsymbol{\Xi}_i) + \frac{1}{n}\sum_{i=1}^{n}\frac{(\mathbf{a}^{\mathrm{T}}\boldsymbol{\Xi}_i)^2\boldsymbol{\Xi}_i}{1+\mathbf{a}^{\mathrm{T}}\boldsymbol{\Xi}_i}.$$

The second term is $o_p(n^{-1/2})$ by (A.9) and (A.10). It follows that

$$\mathbf{a} = \left(\sum_{i=1}^{n}\boldsymbol{\Xi}_i\boldsymbol{\Xi}_i^{\mathrm{T}}\right)^{-1}\sum_{i=1}^{n}\boldsymbol{\Xi}_i + o_p(n^{-1/2}). \tag{A.11}$$

A similar argument yields that

$$\frac{1}{n}\sum_{i=1}^{n}\frac{\mathbf{a}^{\mathrm{T}}\boldsymbol{\Xi}_i}{1+\mathbf{a}^{\mathrm{T}}\boldsymbol{\Xi}_i} = \frac{1}{n}\sum_{i=1}^{n}\mathbf{a}^{\mathrm{T}}\boldsymbol{\Xi}_i - \frac{1}{n}\sum_{i=1}^{n}(\mathbf{a}^{\mathrm{T}}\boldsymbol{\Xi}_i)^2 + o_p(n^{-1}).$$

This indicates that

$$\sum_{i=1}^{n}\mathbf{a}^{\mathrm{T}}\boldsymbol{\Xi}_i = \sum_{i=1}^{n}(\mathbf{a}^{\mathrm{T}}\boldsymbol{\Xi}_i)^2 + o_p(1). \tag{A.12}$$

Now let us consider $\mathcal{R}_n(\boldsymbol{\beta})$. By the Taylor expansion, we have

$$\begin{aligned} -\log\{\mathcal{R}_n(\boldsymbol{\beta})\} &= \sum_{i=1}^{n}\log(1+\mathbf{a}^{\mathrm{T}}\boldsymbol{\Xi}_i) \\ &= \sum_{i=1}^{n}\left\{\mathbf{a}^{\mathrm{T}}\boldsymbol{\Xi}_i - \frac{1}{2}(\mathbf{a}^{\mathrm{T}}\boldsymbol{\Xi}_i)^2\right\} + Q_n. \end{aligned}$$

The remainder term Q_n is bounded by $\|\mathbf{a}\|^2 \max_{1\le i\le n}\|\mathbf{a}^{\mathrm{T}}\boldsymbol{\Xi}_i\| \sum_{i=1}^n \|\boldsymbol{\Xi}_i\|^2$. Note that $\|\mathbf{a}\|^2 = O_p(n^{-1})$, $\max_{1\le i\le n}\|\mathbf{a}^{\mathrm{T}}\boldsymbol{\Xi}_i\| = o_p(1)$, and $n^{-1}\sum_{i=1}^n \|\boldsymbol{\Xi}_i\|^2 = O_p(1)$. We have that $Q_n = o_p(1)$. This argument with (A.12) and (A.11) implies that

$$\begin{aligned} -2\log\{\mathcal{R}_n(\boldsymbol{\beta})\} &= \sum_{i=1}^n \mathbf{a}^{\mathrm{T}}\boldsymbol{\Xi}_i\boldsymbol{\Xi}_i^{\mathrm{T}}\mathbf{a} + o_p(1) \\ &= \left(\frac{1}{\sqrt{n}}\sum_{i=1}^n \boldsymbol{\Xi}_i\right)^{\mathrm{T}} \left(\frac{1}{n}\sum_{i=1}^n \boldsymbol{\Xi}_i\boldsymbol{\Xi}_i^{\mathrm{T}}\right)^{-1} \left(\frac{1}{\sqrt{n}}\sum_{i=1}^n \boldsymbol{\Xi}_i\right) + o_p(1). \end{aligned}$$

The proof is completed.

References

Akaike, H. (1973), "Maximum Likelihood Identification of Gaussian Autoregressive Moving Average Models," *Biometrika*, 60, 255-65.

Breiman, L. (1996), "Heuristics of Instability and Stabilization in Model Selection," *The Annals of Statistics*, 24, 2350-2383.

Bunea, F. (2004), "Consistent Covariate Selection and Post Model Selection Inference in Semiparametric Regression," *The Annals of Statistics*, 32, 898-927

Bunea, F. and Wegkamp, M. (2004), "Two-stage Model Selection Procedures in Partially Linear Regression," *The Canadian Journal of Statistics*, 32, 105-118.

Carroll, R. J., Freedman, L. S., Kipnis, V. and Li, L. (1998), "A New Class of Measurement Error Models, with Applications to Estimating the Distribution of Usual Intake," *The Canadian Journal of Statistics*, 26, 467-477.

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

Chen, S.X. (1993), "On the Accuracy of Empirical Likelihood Confidence Regions for Linear Regression Model," *Annals of the Institute of Statistical Mathematics*, 45, 621-637.

Chen, S.X. (1994), "Empirical Likelihood Confidence Intervals for Linear Regression Coefficients," *Journal of Multivariate Analysis*, **49**, 24-40.

Craven, P. and Wahba, G. (1979), "Smoothing Noisy Data with Spline Functions: Estimating the Correct Degree of Smoothing by the Method of Generalized Cross-validation," *Numer. Math.*, 31, 377–403.

Engle, R.F., Granger, C.W.J., Rice, J., and Weiss, A. (1986), "Semiparametric Estimates of the Relation between Weather and Electricity Sales," *Journal of the American Statistical Association*, 81, 310–320.

Fan, J. and Gijbels, I. (1996), *Local Polynomial Modelling and Its Applications*. New York: Chapman and Hall.

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. (2004), "New Estimation and Model Selection Procedures for Semiparametric Modeling in Longitudinal Data Analysis," *Journal of the American Statistical Association*, 99, 710-723.

Foster, D. P. and George, E. I. (1994), "The Risk Inflation Criterion for Multiple Regression," *The Annals of Statistics*, 22, 1947-75.

Frank, I.E. and Friedman, J. H. (1993), "A Statistical View of Some Chemometrics Regression Tools," *Technometrics*, 35, 109-148.

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

Härdle, W., Liang, H. and Gao, J. (2000), *Partially Linear Models*, Heidelberg: Springer Physica-Verlag.

Heckman, N.E. (1986), "Spline Smoothing in Partly Linear Models," *Journal of the Royal Statistical Society, Series B*, 48, 244–248.

Hunter, D. and Li, R. (2005), "Variable Selection Using MM Algorithms," *The Annals of Statistics*, 33, 1617-1642.

Kolaczyk (1994), "Empirical Likelihood and Generalized Linear Models," *Statistica Sinica*, 4, 199-218.

Liang, H., Härdle, W., and Carroll, R.J. (1999), "Estimation in a Semiparametric Partially Linear Errors-in-variables Model," *The Annals of Statistics*, 27, 1519-1935.

Liang, H., Wang, S.J., Robins, J.M., and Carroll, R.J. (2004), "Estimation in Partially Linear Models with Missing Covariates," *Journal of the American Statistical Association*, 99, 357-367.

Lin, D.Y. and Ying, Z. (2001), "Semiparametric and Nonparametric Regression Analysis of Longitudinal Data (with discussion)," *Journal of the American Statistical Association*, 96, 103-113.

Mack, Y. and Silverman, B. (1982), "Weak and Strong Uniform Consistency of Kernel Regression Estimates," *Zeitschrift für Wahrscheinlichkeitstheorie und Verwandte Gebiete*, 60, 405–415.

Martinussen, T. and Scheike, T. H. (1999), "A semiparametric Additive Regression Model for Longitudinal Data," *Biometrika*, 86, 691-702.

Martinussen, T. and Scheike, T. H. (2001), "Sampling Adjusted Analysis of Dynamic Additive Regression Models for Longitudinal Data," *Scandinavian Journal of Statistics*, 28, 303-323.

Moyeed, R. A. and Diggle, P. J. (1994), "Rates of Convergence in Semi-parametric Modeling of Longitudinal Data," *Australia Journal of Statistics*, 36, 7593.

Newey, W. K. (1994), "The Asymptotic Variance of Semiparametric Estimators," *Econometrica*, 62, 1349–1382.

Opsomer, J.D. and Ruppert, D. (1999), "A Root-n Consistent Backfitting Estimator for Semiparametric Additive Modelling," *Journal of Computational and Graphical Statistics*, 8, 715-732.

Owen, A.B. (1988), "Empirical Likelihood Ratio Confidence Intervals for a Single Functional," *Biometrika*, 75, 237-249.

Owen, A.B. (1990), "Empirical Likelihood Confidence Regions," *The Annals of Statistics*, 18, 90-120.

Owen, A.B. (1991), "Empirical Likelihood for Linear Models," *The Annals of Statistics*, 19, 1725-1747.

Owen, A.B. (2000), *Empirical Likelihood*, London: Chapman and Hall/CRC.

Qin, J. (1994), "Semi-empirical Likelihood Ratio Confidence Intervals for the Difference of Two Sample Means," *The Annals of Statistics*, 46, 117-26.

Qin, J. (1999), "Empirical Likelihood Ratio Based Confidence Intervals for Mixture Proportions," *The Annals of Statistics*, 27, 1368-1384.

Qin, J. and Lawless, J. (1994), "Empirical Likelihood and General Estimating Equations," *The Annals of Statistics*, 22, 300-325.

Robinson, P. M. (1988), "Root-n-Consistent Semiparametric Regression," *Econometrica*, 56, 931–954.

Ruppert, D., Sheather, S. J. and Wand, M. P. (1995), "An Effective Bandwidth Selector for Local Least Squares Regression," *Journal of the American Statistical Association*, 90, 1257–1270.

Schwarz, G. (1978), "Estimating the Dimension of a Model," *The Annals of Statistics*, 6, 461-464.

Speckman, P. (1988), "Kernel Smoothing in Partial Linear Models," *Journal of the Royal Statistical Society, Series B*, 50, 413–436.

Tibshirani, R. (1996), "Regression Shrinkage and Selection via the LASSO," *Journal of the Royal Statistical Society, Series B*, 58, 267-288.

Zeger, S. L. and Diggle, P. J. (1994), "Semiparametric Models for Longitudinal Data with Application to CD4 Cell Numbers in HIV Seroconverters," *Biometrics*, 50, 689–699.