Linear Model Selection when Covariates Contain Errors

Xinyu Zhang
Academy of Mathematics and Systems Science, Chinese Academy of Sciences, Beijing 100190, China, xinyu@amss.ac.cn

Haiying Wang
Department of Mathematics and Statistics, University of New Hampshire, Durham, NH 03824, haiying.wang@unh.edu

Yanyuan Ma
Department of Statistics, Penn State University, State College, PA 16802, yanyuanma@gmail.com
Raymond J. Carroll
Department of Statistics, Texas A&M University, 3143 TAMU, College Station, TX 77843-3143, and School of Mathematical and Physical Sciences, University of Technology Sydney, Broadway NSW 2007, carroll@stat.tamu.edu

ABSTRACT

Prediction precision is arguably the most relevant criterion of a model in practice and is often a sought after property. A common difficulty with covariates measured with errors is the impossibility of performing prediction evaluation on the data even if a model is completely given without any unknown parameters. We bypass this inherent difficulty by using special properties on moment relations in linear regression models with measurement errors. The end product is a model selection procedure that achieves the same optimality properties that are achieved in classical linear regression models without covariate measurement error. Asymptotically, the procedure selects the model with the minimum prediction error in general, and selects the smallest correct model if the regression relation is indeed linear. Our model selection procedure is useful in prediction when future covariates without measurement error become available, e.g., due to improved technology or better management and design of data collection procedures.

Some Key Words: Errors in covariates, Loss efficiency, Measurement error, Model selection, Selection consistency.

1 Introduction

Model selection is a much studied problem in the regression context. Familiar methods such as AIC and BIC have long existed and are well studied; see for example Shao (1997). However, when covariates are subject to measurement error, relatively little work has been conducted to study how to perform model selection. In fact, the only literature we are aware of are in terms of variable selection in partially linear models (Liang and Li, 2009) and in generalized partially linear measurement error models (Ma and Li, 2010). A tuning parameter is required in these methods and its effect is well studied in Zhang et al. (2010). Model selection in these papers is performed simultaneously with parameter estimation, and is achieved through shrinking small coefficients towards zero. As such, a critical requirement of these methods is sparsity, which implies that the true model is indeed included in the set of candidate models under consideration, and the true model is sufficiently simple. Here, we make an important distinction between the true model and a correct model. The true model refers to the smallest correct model, i.e. the correct model with the smallest dimension among all possible models under consideration. Of course, the true model may not be included in the set of candidate models for selection. As a result, it is unclear what these methods will yield if all the models under consideration are misspecified.

In this work, we systematically study the issue of model selection in the context of linear measurement error models. We evaluate the goodness-of-fit of a candidate model using its prediction error, which is arguably the most relevant criterion in practice (Efron, 2004). Here, by prediction, we imply predicting the response based on the chosen linear model and the error free covariates. A common perception in the measurement error model framework is that for prediction purposes, it is not necessary to account for measurement error. But this is only true in the situation that the measurement error structures used in the data analysis and used in prediction are exactly the same. In this case, a sensible thing to do in the prediction context is to simply use the observed data (Carroll et al., 2006). However, as soon as the measurement error structure changes, for example when the measurement error variance decreases in the data used for prediction, this shortcut no longer applies: an example in nonparametric regression is Carroll et al. (2009). In our problem studied in Section 3.4, in a small data set, the true covariates are available. Thus, studying the true underneath modeling and estimation between the response variable and the true covariates of original interest is highly relevant.

To this end, although prediction error is usually unobtainable in measurement error models, we are able to bypass this difficulty and estimate the cumulative effect of the prediction errors using linear measurement error model properties. Model complexity is evaluated via a degrees of freedom calculation, which is of course nonstandard because of measurement errors. Finally, we study the effect of the various sizes of the model complexity penalties and derive the properties of the model selection procedure, both when the candidate model set contains some correctly specified models and when it does not contain any correctly specified model.

2 Main Methodology and Theoretical Results

2.1 Models and notation

We now formally describe the problem we work on and the related notation. Consider a data generation process

$$Y_i = \mu_i + \epsilon_i \tag{1}$$

for $i = 1, \ldots, n$. Here Y_i is a univariate response variable, μ_i is the mean of Y_i, and ϵ_i is an error term with mean zero and variance σ^2. Write $\mathbf{Y} = (Y_1, \ldots, Y_n)^{\mathrm{T}}$, $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_n)^{\mathrm{T}}$, and $\boldsymbol{\epsilon} = (\epsilon_1, \ldots, \epsilon_n)^{\mathrm{T}}$. Let $\mathbf{X}_i$ be a p-dimensional covariate vector used to predict μ_i. However, some or all components of $\mathbf{X}_i$ are measured with error. Thus, instead of observing $\mathbf{X}_i$, we observe a p-dimensional random variable $\mathbf{W}_i$, where $\mathbf{W}_i = \mathbf{X}_i + \mathbf{U}_i$, and $\mathbf{U}_i$ is a mean zero normal random vector with variance-covariance matrix $\boldsymbol{\Sigma}$. We allow some components of $\mathbf{U}_i$ to be identically zero, therefore these components of $\mathbf{X}_i$ are precisely measured. This also allows us to include the constant 1 in $\mathbf{X}_i$. Without loss of generality, we assume that the first p^* components of $\mathbf{X}_i$

are subject to errors, while the remaining $p - p^*$ components are error free. Thus, the lower $p - p^*$ subvector of $\mathbf{U}_i$ is zero, and $\boldsymbol{\Sigma}$ is zero except for its upper-left $p^* \times p^*$ submatrix. We also assume that the measurement error vector $\mathbf{U}_i$ is independent of ϵ_i, and $(\mathbf{U}_i, \epsilon_i)$ are independent and identically distributed for $i = 1, \dots, n$, and independent of $\mathbf{X}_i$. Since $\mathbf{X}_i$ is independent of $(\mathbf{U}_i, \epsilon_i)$ for $i = 1, \dots, n$, we can treat $\mathbf{X}_i$'s as nonrandom covariate vectors.

Assume that we have S_n candidate models for μ_i. In the s^{th} candidate model, μ_i is modeled by $\mu_i = \mathbf{X}_{(s),i}^{\mathrm{T}} \boldsymbol{\beta}_{(s)}$, where $\mathbf{X}_{(s),i}$ is a $p_{(s)}$-dimensional sub-vector of $\mathbf{X}_i$, and $\boldsymbol{\beta}_{(s)}$ is its coefficient vector. Instead of observing $\mathbf{X}_{(s),i}$, we observe a random variable $\mathbf{W}_{(s),i}$, where $\mathbf{W}_{(s),i} = \mathbf{X}_{(s),i} + \mathbf{U}_{(s),i}$, and $\mathbf{U}_{(s),i}$ is a sub-vector of $\mathbf{U}_i$ with mean zero and variance-covariance matrix $\boldsymbol{\Sigma}_{(s)}$. Let $\mathbf{U}_{(s)} = (\mathbf{U}_{(s),1}, \dots, \mathbf{U}_{(s),n})^{\mathrm{T}}$, $\mathbf{X}_{(s)} = (\mathbf{X}_{(s),1}, \dots, \mathbf{X}_{(s),n})^{\mathrm{T}}$ and $\mathbf{W}_{(s)} = (\mathbf{W}_{(s),1}, \dots, \mathbf{W}_{(s),n})^{\mathrm{T}}$.

2.2 Parameter estimation and degrees of freedom

In the s^{th} candidate model, following Carroll et al. (2006), $\boldsymbol{\beta}_{(s)}$ can be estimated from

$$\widehat{\boldsymbol{\beta}}_{(s)} = (\mathbf{W}_{(s)}^{\mathrm{T}} \mathbf{W}_{(s)} - n\boldsymbol{\Sigma}_{(s)})^{-1} \mathbf{W}_{(s)}^{\mathrm{T}} \mathbf{Y}.$$

Subsequently, the corresponding estimator of the regression mean function $\boldsymbol{\mu}$ is

$$\widehat{\boldsymbol{\mu}}_{(s)} = (\widehat{\mu}_{(s),1}, \dots, \widehat{\mu}_{(s),n})^{\mathrm{T}} = \mathbf{X}_{(s)} \widehat{\boldsymbol{\beta}}_{(s)}.$$

The number of degrees of freedom is an important element in statistical analysis as a measure of the complexity of different models. The research literature on the construction of the degrees of freedom is large, see Hastie and Tibshirani (1990), Efron (2004), Zou et al. (2007), Mukherjee et al. (2015) and the references therein, in which definitions and unbiased estimators of degrees of freedom under different settings are one of the primary theoretical results. Following Efron (2004), we define the degrees of freedom of the s^{th} model as $df_{(s)} = \sigma^{-2} cov(\widehat{\boldsymbol{\mu}}_{(s)}^{\mathrm{T}}, \mathbf{Y}^{\mathrm{T}})$. It is easy to check that without measurement error, $df_{(s)} = p_{(s)}$. When some covariates are contaminated with measurement errors, we suggest to estimate $df_{(s)}$ using

$$\widehat{df}_{(s)} = p_{(s)} + \frac{t_{(s),2} + (t_{(s),1} - p_{(s)} - 1) t_{(s),1}}{n}, \tag{2}$$

where $t_{(s),k} = \mathrm{tr}(\mathbf{H}_{(s)}^k)$ for any positive integer k, $\mathbf{H}_{(s)} = \mathbf{W}_{(s)}\mathbf{G}_{(s)}$, $\mathbf{G}_{(s)} = \mathbf{\Lambda}_{(s)}\mathbf{W}_{(s)}^{\mathrm{T}}$, and $\mathbf{\Lambda}_{(s)} = (\mathbf{W}_{(s)}^{\mathrm{T}}\mathbf{W}_{(s)} - n\mathbf{\Sigma}_{(s)})^{-1}$.

Theorem 1. *For any $s \in \{1, \ldots, S_n\}$, $E(\widehat{df}_{(s)}) = df_{(s)}$.*

Theorem 1 indicates that $\widehat{df}_{(s)}$ is an unbiased estimator of $df_{(s)}$. Note that $\widehat{df}_{(s)}$ contains the matrix $\mathbf{\Sigma}_{(s)}$, which will be replaced by its estimate $\widehat{\mathbf{\Sigma}}_{(s)}$ if needed. Theorem 1 is proved in Appendix A.1.

Remark 1. The unbiasedness properties in this subsection is mainly on the basis on Stein Lemma (i.e., our Lemma 1). Although, this technique has also been used in other literature such as Zou et al. (2007) and Liang et al. (2008), our derivation is non-trivial because in our paper the lemma takes effect because of the normal measurement errors while in other literature it takes effect because of the normal model error The estimated degree of freedom will form an important part of the model selection criteria, as is demonstrated below.

2.3 Model selection criteria

To study model selection and establish selection criteria, we first define $L_{(s)} = \|\widehat{\boldsymbol{\mu}}_{(s)} - \boldsymbol{\mu}\|^2$ as the squared loss of $\widehat{\boldsymbol{\mu}}_{(s)}$, and let the risk be $R_{(s)} = E(L_{(s)})$. Our goal is to select a "best" model, in that it remains as parsimonious as possible, while minimizing the loss. To facilitate the minimization procedure, we further characterize $R_{(s)}$. We first note that the risk can be decomposed as

$$\begin{aligned} R_{(s)} &= E(L_{(s)}) = E\|\widehat{\boldsymbol{\mu}}_{(s)} - \boldsymbol{\mu}\|^2 = E\|\widehat{\boldsymbol{\mu}}_{(s)} - \mathbf{Y} + \boldsymbol{\epsilon}\|^2 \\ &= E\|\widehat{\boldsymbol{\mu}}_{(s)} - \mathbf{Y}\|^2 + E\|\boldsymbol{\epsilon}\|^2 + 2E\{(\widehat{\boldsymbol{\mu}}_{(s)} - \mathbf{Y})^{\mathrm{T}}\boldsymbol{\epsilon}\}. \end{aligned}$$

Obviously, $E\|\boldsymbol{\epsilon}\|^2 = n\sigma^2$, while the third component above can be calculated as

$$\begin{aligned} 2E\{(\widehat{\boldsymbol{\mu}}_{(s)} - \mathbf{Y})^{\mathrm{T}}\boldsymbol{\epsilon}\} &= 2E(\widehat{\boldsymbol{\mu}}_{(s)}^{\mathrm{T}}\boldsymbol{\epsilon}) - 2E(\mathbf{Y}^{\mathrm{T}}\boldsymbol{\epsilon}) \\ &= 2E[\{\widehat{\boldsymbol{\mu}}_{(s)} - E(\widehat{\boldsymbol{\mu}}_{(s)})\}^{\mathrm{T}}(\mathbf{Y} - \boldsymbol{\mu})] + 2E(\widehat{\boldsymbol{\mu}}_{(s)})^{\mathrm{T}}E(\boldsymbol{\epsilon}) - 2E(\mathbf{Y}^{\mathrm{T}}\boldsymbol{\epsilon}) \\ &= 2cov(\widehat{\boldsymbol{\mu}}_{(s)}^{\mathrm{T}}, \mathbf{Y}^{\mathrm{T}}) - 2n\sigma^2. \end{aligned}$$

This leads to the alternative expression of the risk

$$R_{(s)} = E\|\widehat{\boldsymbol{\mu}}_{(s)} - \mathbf{Y}\|^2 + 2\sigma^2 df_{(s)} - n\sigma^2, \tag{3}$$

Here $n\sigma^2$ does not change with the model choice, while in Section 2.2, we have shown that $\widehat{df}_{(s)}$ is an unbiased estimator of $df_{(s)}$. So $\|\widehat{\boldsymbol{\mu}}_{(s)} - \mathbf{Y}\|^2 + 2\sigma^2\widehat{df}_{(s)}$ is an unbiased estimator of $R_{(s)}$ plus a constant. However, $\widehat{\boldsymbol{\mu}}_{(s)}$ depends on the unobservable $\mathbf{X}_{(s)}$, hence we need to further develop an unbiased estimator of $R_{(s)}$ which depends on $\mathbf{W}_{(s)}$ instead of $\mathbf{X}_{(s)}$. To do this, we first define $a_{(s)} = t_{(s),1} - p_{(s)}$ and

$$\begin{aligned} B_{(s)} &= n^{-2}\{2na_{(s)} + a_{(s)}^2 - a_{(s)} + t_{(s),2} - t_{(s),1}\}\mathbf{Y}^{\mathrm{T}}\mathbf{Y} \\ &\quad + n^{-2}2\{n(1 - a_{(s)} - t_{(s),1}) - a_{(s)}^2 + a_{(s)}(3 - t_{(s),1}) - 2t_{(s),2} + 3t_{(s),1} - 4\}\mathbf{Y}^{\mathrm{T}}\mathbf{H}_{(s)}\mathbf{Y} \\ &\quad + n^{-2}\{n(2t_{(s),1} - 4) + a_{(s)}^2 + a_{(s)}(2t_{(s),1} - 9) - 7t_{(s),1} + 3t_{(s),2} + 28\}\mathbf{Y}^{\mathrm{T}}\mathbf{H}_{(s)}^2\mathbf{Y} \\ &\quad + n^{-2}2(n + 2a_{(s)} + t_{(s),1} - 16)\mathbf{Y}^{\mathrm{T}}\mathbf{H}_{(s)}^3\mathbf{Y} + n^{-2}12\mathbf{Y}^{\mathrm{T}}\mathbf{H}_{(s)}^4\mathbf{Y}. \end{aligned}$$

We then have the following result.

Theorem 2. *For any $s \in \{1, \ldots, S_n\}$,*

$$E\|\widehat{\boldsymbol{\mu}}_{(s)} - \mathbf{Y}\|^2 = E\{\mathbf{Y}^{\mathrm{T}}(\mathbf{I}_n - \mathbf{H}_{(s)})\mathbf{Y} + B_{(s)}\}, \tag{4}$$

where $\mathbf{I}_n$ is the $n \times n$ identity matrix.

Theorem 2 is proved in the supplementary material. From Theorems 1 and 2, combined with the results in (3), we know that

$$\mathbf{Y}^{\mathrm{T}}(\mathbf{I}_n - \mathbf{H}_{(s)})\mathbf{Y} + B_{(s)} + 2\sigma^2\widehat{df}_{(s)} \tag{5}$$

is an unbiased estimator of the risk $R_{(s)}$ up to a constant. Let $\widehat{\boldsymbol{\Sigma}}$ and $\widehat{\sigma}^2$ be the estimators of $\boldsymbol{\Sigma}$ and σ^2, and let $\widetilde{B}_{(s)}$, $\widetilde{\boldsymbol{\Lambda}}_{(s)}$, $\widetilde{\mathbf{G}}_{(s)}$, $\widetilde{\mathbf{H}}_{(s)}$ and $\widetilde{df}_{(s)}$ be the corresponding quantities with $\boldsymbol{\Sigma}$ substituted by $\widehat{\boldsymbol{\Sigma}}$. We then propose to approximate the risk $R_{(s)}$, up to a constant, using

$$C_2(s) = \mathbf{Y}^{\mathrm{T}}(\mathbf{I}_n - \widetilde{\mathbf{H}}_{(s)})\mathbf{Y} + \widetilde{B}_{(s)} + 2\widehat{\sigma}^2\widetilde{df}_{(s)}.$$

Based on the above analysis, we define a general model selection criterion

$$C_\lambda(s) = \mathbf{Y}^{\mathrm{T}}(\mathbf{I}_n - \widetilde{\mathbf{H}}_{(s)})\mathbf{Y} + \widetilde{B}_{(s)} + \lambda_n \widehat{\sigma}^2 \widetilde{df}_{(s)},$$

where λ_n is a tuning parameter. Our method is motivated by the **un**biased **e**stimator in (5) of the **r**isk, so we term the method the UBER information criterion (UBERIC). We write the corresponding selected model as the $\widehat{s}_\lambda^{\mathrm{th}}$ model, where

$$\widehat{s}_\lambda = \mathrm{argmin}_{s \in \{1,\ldots,S_n\}} C_\lambda(s).$$

In the following, we derive the asymptotic properties of $\widehat{s}_\lambda$.

<u>Remark</u> 2. In the measurement error literature, the estimate $\widehat{\mathbf{\Sigma}}$ can be obtained by two strategies: one is through using duplicate measurements corresponding to each $\mathbf{X}_i$; the other is through introducing instrumental variables. For instance, if we have duplicate measurements $\mathbf{W}_{i,j} = \mathbf{X}_i + \mathbf{U}_{i,j}$ for $j = 1, \ldots, J_i$ and $i = 1, \ldots, n$, then we can estimate $\mathbf{\Sigma}$ by

$$\widehat{\mathbf{\Sigma}} = \frac{\sum_{i=1}^n \sum_{j=1}^{J_i} (\mathbf{W}_{i,j} - \bar{\mathbf{W}}_i)(\mathbf{W}_{i,j} - \bar{\mathbf{W}}_i)^{\mathrm{T}}}{\sum_{i=1}^n (J_i - 1)}, \tag{6}$$

where $\bar{\mathbf{W}}_i = J_i^{-1} \sum_{j=1}^{J_i} \mathbf{W}_{i,j}$. By using the full model, i.e., the model containing all covariates $\mathbf{X}_i$ and following Carroll et al. (2006), we estimate σ^2 as $\widehat{\sigma}^2 = \{\|\mathbf{Y} - \mathbf{W}\widehat{\boldsymbol{\beta}}_{\text{full}}\|^2 - n\widehat{\boldsymbol{\beta}}_{\text{full}}^{\mathrm{T}} \widehat{\mathbf{\Sigma}} \widehat{\boldsymbol{\beta}}_{\text{full}}\}/(n - p)$, where $\mathbf{W} = (\mathbf{W}_1, \ldots, \mathbf{W}_n)^{\mathrm{T}}$ and $\widehat{\boldsymbol{\beta}}_{\text{full}} = (\mathbf{W}^{\mathrm{T}}\mathbf{W} - n\widehat{\mathbf{\Sigma}})^{-1}\mathbf{W}^{\mathrm{T}}\mathbf{Y}$.

2.4 Asymptotic loss efficiency and selection consistency

We name the s^{th} candidate model ($s \in \{1, \ldots, S_n\}$) as a correct model if indeed $\boldsymbol{\beta}_{(s)}$ satisfies that $\boldsymbol{\mu} = \mathbf{X}_{(s)}\boldsymbol{\beta}_{(s)}$. Let $\mathcal{S}_n^C$ be the set of correct candidate models, $\mathcal{S}_n^I$ be the set of incorrect candidate models, and $\widetilde{\mathbf{P}}_{(s)} = \mathbf{X}_{(s)}\widetilde{\mathbf{G}}_{(s)}$. Write the estimator of $\boldsymbol{\mu}$ as $\widetilde{\boldsymbol{\mu}}_{(s)} = \widetilde{\mathbf{P}}_{(s)}\mathbf{Y}$ and the estimated squared estimation loss as $\widetilde{L}_{(s)} = \|\widetilde{\boldsymbol{\mu}}_{(s)} - \boldsymbol{\mu}\|^2$. Let $\breve{\mathbf{P}}_{(s)} = \mathbf{X}_{(s)}(\mathbf{X}_{(s)}^{\mathrm{T}}\mathbf{X}_{(s)})^{-1}\mathbf{X}_{(s)}^{\mathrm{T}}$, $\breve{L}_{(s)} = \|\breve{\mathbf{P}}_{(s)}\mathbf{Y} - \boldsymbol{\mu}\|^2$, and $\breve{R}_{(s)} = E(\breve{L}_{(s)})$, the squared estimation risk without measurement error.

We first discuss two conditions, which play important roles in terms of the asymptotic properties of the UBERIC model selection procedure in Section 2.3. All the limiting properties in the conditions and throughout the text hold when $n \to \infty$.

Condition (C.1). $n^{1/2}p^2 / \min_{s \in \mathcal{S}_n^I} \breve{R}_{(s)} = o(1)$.

Condition (C.2). $\lambda_n p / \min_{s \in \mathcal{S}_n^I} \breve{R}_{(s)} = o(1)$.

Both Conditions (C.1) and (C.2) require the incorrect models to have sufficiently large risk. More specifically, Condition (C.1) implies that the minimum squared estimation risk of an incorrect candidate model increases faster than the rate $n^{1/2}p^2$. Based on Section S2.2 of Flynn et al. (2013), we know that typically $\min_{s \in \mathcal{S}_n^I} \breve{R}_{(s)}$ has order n, so Condition (C.1) is satisfied when $p = o(n^{1/4})$. Similarly, Condition (C.2) requires the risk to increase faster than $\lambda_n p$, which is typically satisfied as long as $\lambda_n p = o(n)$. Compared with the model selection procedures without involving covariate measurement errors studied in (Li, 1987) and (Shao, 1997), where it is only required that $\min_{s \in \mathcal{S}_n^I} \breve{R}_{(s)} \to \infty$, Conditions (C.1) and (C.2) are more specific and slightly stronger. This is the price we pay for handling the measurement errors. Intuitively, the presence of measurement errors blurs the assessment of the risk. Hence, only when the risks of the incorrect models are sufficiently worse than those of the correct ones, we can tell the incorrect models apart from the correct ones.

In addition to Conditions (C.1) and (C.2), we also need some more technical conditions, which are all quite mild. Denote by $\lambda_{\max}(\mathbf{M})$ and $\lambda_{\min}(\mathbf{M})$ the maximum and minimum singular values for a matrix M, respectively.

Condition (C.3). *There are constants* $0 < c_1 \leq c_2 < \infty$ *such that*

$$c_1 < n^{-1}\lambda_{\min}(\mathbf{X}^{\mathrm{T}}\mathbf{X}) \leq n^{-1}\lambda_{\max}(\mathbf{X}^{\mathrm{T}}\mathbf{X}) < c_2 \quad \textit{and} \quad \lambda_{\max}(\boldsymbol{\Sigma}) < c_2.$$

Condition (C.4). $\|\widehat{\boldsymbol{\Sigma}} - \boldsymbol{\Sigma}\| = O_p(n^{-1/2}p)$ *where* $\|\cdot\|$ *is the Frobenius norm, and* $\widehat{\sigma}^2 = O_p(1)$.

Condition (C.5). $\|\boldsymbol{\mu}\| = O(n^{1/2})$.

Condition (C.6). $p^2/n = o(1)$.

Remark 3. In Condition (C.3), $\mathbf{X}$ is assumed to be nonrandom. If it is random, we need the first set of inequalities to hold in probability, i.e., these inequalities hold with probability approaching one as $n \to \infty$.

Remark 4. The two assumptions in Condition (C.4) are quite mild. The first assumption requires that the estimator $\widehat{\boldsymbol{\Sigma}}$ is root-n consistent. If p is fixed, the estimator in (6) satisfies this requirement(Carroll et al., 2006). The second assumption requires that $\widehat{\sigma}^2$ is bounded in probability; it does not even require consistency of $\widehat{\sigma}^2$.

Theorem 3. Assume that Conditions (C.1)-(C.6) hold.
(1) (**Asymptotic Loss Efficiency**) If $\mathcal{S}_n^C$ is empty, then $\widehat{s}_\lambda$ is asymptotically loss efficient, i.e.,

$$\frac{\widetilde{L}_{\widehat{s}_\lambda}}{\min_{s \in \{1,\dots,S_n\}} \widetilde{L}_{(s)}} \to 1$$

in probability.
(2) (**Selection Consistency**) If $\mathcal{S}_n^C$ is not empty and $\lambda_n/(n^{1/2}p) \to \infty$, then model $\widehat{s}_\lambda$ is consistent, i.e. the probability that $\widehat{s}_\lambda$ is correct and has the smallest dimension goes to 1.

Theorem 3 is proved in Appendix A.2.

Remark 5. Theorem 3 establishes two properties of the model selection procedure: when none of the candidate models is correct, the selection procedure finds a model that minimizes the loss; when at least one correct model is included in the candidate models, the selection procedure finds the most concise correct model. For the first property to hold, the requirement of the tuning parameter λ_n is simply what is described in Condition (C.2). We can easily see that when $\lambda_n = 2$, i.e. for AIC, the method is asymptotically loss efficient given the other conditions listed. This finding agrees with that in the linear regression model case without measurement error. However, for the second property to hold, we need $\lambda_n/(n^{1/2}p) \to \infty$. This obviously eliminates BIC, which sets $\lambda_n = \log(n)$. This finding indicates that the common choice of $\lambda_n = \log(n)$ may not lead to a consistent selection criterion for measurement error models and thus new tuning parameters are needed to achieve selection consistency.

Remark 6. After a model is selected, we can proceed to perform coefficient estimation under the model. In the case when the candidate models include the true model (the smallest correct model), the selection consistency property ensures that when the sample size is large, the estimation will essentially be performed under the true model. If Lindeberg's condition is satisfied, then the estimator of coefficients has the usual consistency, root-n convergence rate and asymptotically normal distribution as if the true model is given. This indicates that the model selection procedure does not incur additional cost for the subsequent estimation procedure, in other words, for large samples, performing estimation following the model selection procedure is the same as performing estimation in the given true model. This is usually known as the oracle property.

Remark 7. Because we typically consider the case that the candidate model set contains all the $2^p - 1$ possible linear models formed by the p available variables, the candidate model set either contains the true model, or it does not contain any correct model at all. However, if the candidate model set does not contain all possible linear models, then an interesting situation is when the true model is not included but some other correct models are included. For example, if the true model contains only two variables, say X_1 and X_2, but the candidate set only contains models with at least three variables, then the true model is not included in the candidate model set but any model with X_1 and X_2 in it is a correct model. For this scenario, our results indicate that a correct model with the smallest dimension will be selected because of the selection consistency property. However, if there are multiple correct models with the same smallest dimension, our theory only ensures that one of them will be selected with probability approaching one. It is unclear which one among these multiple models will be selected. Intuitively, this is because one cannot tell which one of these correct models, all with the same dimension, is the best. Indeed it is challenging to define the best model in this situation. Even for linear models without errors in covariates, we are not aware of any results in this situation.

Remark 8. In the proof of Theorem 3, it is also shown that $\widehat{df}_{(s)} = p_{(s)} + o_p(1)$ regardless of whether $\mathbf{\Sigma}_{(s)}$ is replaced by its estimate $\widehat{\mathbf{\Sigma}}_{(s)}$ or not. Assuming the second term of $\widehat{df}_{(s)}$ defined in (2) is uniformly integrable, then we further have $df_{(s)} = E(\widehat{df}_{(s)}) = p_{(s)} + o(1)$ and thus $\widehat{df}_{(s)} - df_{(s)} = o_p(1)$, i.e., $\widehat{df}_{(s)}$ is a consistent estimator $df_{(s)}$.

2.5 Discussion on existence of interactions

A referee asked an interesting question about the applicability of UBERIC when there exists interaction terms in the model, i.e., some of the components in $\mathbf{X}_i$, $\mathbf{U}_i$ or $\mathbf{W}_i$ are in fact products of other components. We now discuss the three cases.

In case one, some of the components in $\mathbf{X}_i$ are products of other components, for example, $X_{i1} = X_{i2}X_{i3}$. Because we do not put any distributional assumption on $\mathbf{X}_i$, as long as $\mathbf{W}_i$ still satisfies $\mathbf{W}_i = \mathbf{X}_i + \mathbf{U}_i$, where $\mathbf{U}_i$ is normal, UBERIC can be applied without any modification.

In case two, some of the components in $\mathbf{U}_i$ are products of other components, for example, $U_{i1} = U_{i2}U_{i3}$. This implies that U_{i1}, U_{i2} and U_{i3} cannot be all normal random variables. Unfortunately, the results in Theorems 1 and 2 heavily rely on the normality assumption of the measurement error distribution, hence the unbiasedness results will not be valid anymore. However, the results in Theorem 3 still hold, and thus UNERIC can still be applied which will still yield the desired properties. Specifically, there are two different scenarios. The first scenario is when $E(\mathbf{U}_i) = \mathbf{0}$. In this scenario, we can still compute the error variance matrix $\mathbf{\Sigma}$ and simply ignore the interaction in UNERIC. If a correct model is included in the candidate model set, UNERIC will still be consistent in the sense that it will select the correct model with the smallest dimension. The second scenario is when $E(\mathbf{U}_i) \neq \mathbf{0}$. In this case, we can view $\mathbf{W}_i = \widetilde{\mathbf{X}}_i + \widetilde{\mathbf{U}}_i$, where $\widetilde{\mathbf{X}}_i = \mathbf{X}_i + E(\mathbf{U}_i)$ and $\widetilde{\mathbf{U}}_i = \mathbf{U}_i - E(\mathbf{U}_i)$. Using UNERIC as it is in this scenario will yield an optimal model $\widetilde{\mathbf{X}}^{\mathrm{T}}\boldsymbol{\beta}$, which leads to the optimal model $\mathbf{X}_i^{\mathrm{T}}\boldsymbol{\beta} + E(\mathbf{U}_i)^{\mathrm{T}}\boldsymbol{\beta}$ in the original setting.

Third, some of the components in $\mathbf{W}_i$ are products of other components, for example, $W_{i1} = W_{i2}W_{i3}$. This implies $W_{i1} = (X_{i2} + U_{i2})(X_{i3} + U_{i3}) = X_{i2}X_{i3} + X_{i2}U_{i3} + X_{i3}U_{i2} + U_{i2}U_{i3} = X_{i1} + U_{i1}$, where $X_{i1} = X_{i2}X_{i3}$ and $U_{I1} = X_{i2}U_{i3} + X_{i3}U_{i2} + U_{i2}U_{i3}$. This unfortunately leads to the dependence between U_{i1} and X_{i1}, which violates the model assumption, hence none of the results will not apply although UNERIC can be implemented.

In measurement error models, the procedure and the subsequent statistical properties can be very different for different error distributions, hence it is not a surprise that UNERIC does not always have the established properties. Linear models with independent normal error provides a starting point for the literature in terms of estimation of measurement error models in general, and it also serves as a starting point for model selection in these models. Much more investigation remains to be carried out in the general measurement error model context.

3 Numerical experiments

3.1 Overview

We now perform numerical experiments to demonstrate the finite sample performance of UBERIC. We consider model selection by minimizing $C_\lambda(s)$ with $\lambda_n = 2$ and $\lambda_n = \log(n)pn^{1/2}$. The former corresponds to the unbiased estimator of $R_{(s)}$ shown in (5), which is a finite sample size property, so it is expected that it has good performance when the sample size is small. The latter is motivated by the condition $\lambda_n/(n^{1/2}p) \to \infty$ in Theorem 3, which is a large sample property, so it is expected that it has good performance when the sample size is large. For comparison, we include the naive AIC and BIC methods ignoring the measurement errors. These are labeled as AIC$_0$ and BIC$_0$. Specifically, AIC$_0$ and BIC$_0$ minimize $n\log\{n^{-1}\|\mathbf{W}_{(s)}(\mathbf{W}_{(s)}^{\mathrm{T}}\mathbf{W}_{(s)})^{-1}\mathbf{W}_{(s)}^{\mathrm{T}}\mathbf{Y} - \mathbf{Y}\|^2\} + 2p_{(s)}$ and $n\log\{n^{-1}\|\mathbf{W}_{(s)}(\mathbf{W}_{(s)}^{\mathrm{T}}\mathbf{W}_{(s)})^{-1}\mathbf{W}_{(s)}^{\mathrm{T}}\mathbf{Y} - \mathbf{Y}\|^2\} + \log(n)p_{(s)}$ respectively.

To handle covariate measurement errors in model selection, Liang and Li (2009) and Wang et al. (2012) proposed AIC$_1$ and BIC$_1$ through minimizing $\|\mathbf{W}_{(s)}\widehat{\boldsymbol{\beta}}_{(s)} - \mathbf{Y}\|^2 - n\widehat{\boldsymbol{\beta}}_{(s)}^{\mathrm{T}}\widehat{\boldsymbol{\Sigma}}_{(s)}\widehat{\boldsymbol{\beta}}_{(s)} + 2\widehat{\sigma}^2 p_{(s)}$ and $\|\mathbf{W}_{(s)}\widehat{\boldsymbol{\beta}}_{(s)} - \mathbf{Y}\|^2 - n\widehat{\boldsymbol{\beta}}_{(s)}^{\mathrm{T}}\widehat{\boldsymbol{\Sigma}}_{(s)}\widehat{\boldsymbol{\beta}}_{(s)} + \log(n)\widehat{\sigma}^2 p_{(s)}$ respectively. We include both AIC$_1$ and BIC$_1$ in our numerical experiment. We also include the penalized least squares variable selection procedure proposed by Liang and Li (2009) for comparison, where the SCAD penalty (Fan and Li, 2001) is implemented. The regularization parameter in the penalized function is chosen respectively by generalized cross-validation (GCV) and BIC of Wang et al. (2007); the latter is supported by Liang and Li (2009).

We consider two simulation examples. In the first example, the candidate model set does not contain any correct model, thus we evaluate the performance of the competing methods through comparing their squared losses $\widetilde{L}_{\widehat{s}_\lambda} = \|\widetilde{\boldsymbol{\mu}}_{\widehat{s}_\lambda} - \boldsymbol{\mu}\|^2$. In the second example, the candidate model set includes some correct models. Thus, we evaluate the performance of the competing methods through inspecting the frequency of selecting the smallest correct model.

3.2 Example I

We generated data from model (1) with $\mu_i = \mathbf{X}_i^{\mathrm{T}}\boldsymbol{\beta} + \nu_i$ and normal additive errors. Here $\nu_i \sim \text{Normal}(0.2, 0.5^2)$, and is independent of $\mathbf{X}_i$. We set n = 25, 50, 100 and 200, $p = [6n^{1/5}/5] + 2$, and generated $\mathbf{X}_i = (X_{i,1}, \ldots, X_{i,p})^{\mathrm{T}}$ from a normal distribution with mean 0 and covariance $\sigma_x^2 0.5^{|j_1 - j_2|}$ between X_{i,j_1} and X_{i,j_2}, $j_1, j_2 = 1, \ldots, p$. We set $\sigma = 0.5$ and 1, $\boldsymbol{\Sigma} = \rho \mathbf{I}_p$, $\rho = 0.5$, and $\boldsymbol{\beta} = (2, 1.5, 0, \ldots, 0)^{\mathrm{T}}$ to generate $\mathbf{U}_i, \mathbf{W}_i$ and Y_i. The variance σ_x^2 is chosen such that the reliability ratio, defined by $\tau = \sigma_x^2/(\sigma_x^2 + \rho)$ (Carroll et al., 2006) varies in $(0.85, 0.95)$. The candidate model set consists of all the linear models that are submodels of $\mathbf{X}_i^{\mathrm{T}}\boldsymbol{\beta}$, thus there are a total of $S_n = 2^p - 1$ candidate models. However, because ν_i is excluded from all the candidate models, none of these models is correct. We conducted 500 replicates, and provide the means of $\widetilde{L}_{\widehat{s}_\lambda} / \min_{s \in \{1,\ldots,S_n\}} \widetilde{L}_{(s)}$ in Table 1.

We see that in general, UBERIC outperforms the naive methods and the existing methods, in that the best relative loss is typically achieved by a procedure based on it. In addition, when the sample size is small, the choice $\lambda = 2$ is often the winner, while as the sample size increases, the choice of $\lambda = \log(n)pn^{1/2}$ catches up and leads the performance among all methods. In addition, when sample sizes increase, the best relative loss approaches 1.

To further check the performance of $\widehat{df}_{(s)}$ in estimating $df_{(s)}$, we present the box plot of the difference $\widehat{df}_{(s)} - df_{(s)}$ in Figure 1, for $s = 2^{p-1}$. Here, $\widehat{df}_{(s)}$ is computed using the average of

$\mathrm{tr}(\mathbf{X}_{(s)}\mathbf{G}_{(s)})$ from 500 replications. It is seen that all the differences are reasonably close to zero, and when n or τ increases, the differences become closer to zero. This performance indicates that $\widehat{df}_{(s)}$ estimates $df_{(s)}$ quite well. Box plots with other s values show similar patterns.

3.3 Example II

We repeat the same experiment as in Example I, except that ν_i is now excluded from μ_i when generating the data. Thus, among the $2^p - 1$ candidate models, the 2^{p-2} submodels that contain the first two covariates are correct models, and the submodel with only the first two variables is the smallest correct model. Table 2 shows the frequency in selecting the smallest correct model from 500 replicates. It is clear that UBERIC with $\lambda = \log(n)pn^{1/2}$ has the best performance, and the frequency of selecting the smallest correct model is very close to 1 when sample sizes becomes large ($n = 100, 200$). Figure 2 shows the box plot of difference $\widehat{df}_{(s)} - df_{(s)}$ with $s = 2^{p-1}$. Similar to the finding in Example I, $\widehat{df}_{(s)}$ also estimates $df_{(s)}$ well.

In the above two examples, the optimization involved in UBERIC is performed in a brute-force way, which is feasible for small p. To facilitate the computation when p is large, we can combine UBERIC and the penalized least squares (PLS) variable selection procedure proposed by Liang and Li (2009), which yields desirable results. Please see the supplementary document for further descriptions.

3.4 Empirical data example

For illustration, we applied our model selection procedure to a data set from the Women's Interview Study of Health (WISH) (Brinton et al., 1995; Potischman et al., 1999). This study was designed as a case-control study and consists of middle aged women who developed breast cancer as well as who did not. To avoid the fact that case-control studies are not random samples from a single population and are thus a case biased sampling, we make a rare disease approximation and analyzed the subset of data formed by 1209 women who did not develop breast cancer (controls), because with rare disease, the controls are quite representative of the entire population.

For the main study , the study collected the measurements of daily intakes of protein (W_1), fat (W_2) and carbohydrate (W_3). These measurements are based on a food frequency questionnaire, hence contains substantial measurement errors. To better understand the mechanism of the measurement error associated with the food frequency questionnaire, the study also collected a validation data set based on a subset of the main study. In the subset, daily intakes were measured by 24-hour recalls as well as dietary records, and these results were combined to provide "true" measurements of protein, fat and carbohydrate intakes; see, for example, Nusser et al. (1996), Spiegelman et al. (2001), and Yi et al. (2015). The validation study consists of 178 observations. Based on these two studies, we obtain the measurement error variance-covariance matrix associated with $(W_1, W_2, W_3)^{\mathrm{T}}$. Finally, age ($W_4$) and smoking status (discrete variable with 3 levels, generated two discrete covariates W_5 and W_6) were also included as covariates without measurement errors. We take body mass index (BMI) as the response variable. Thus, we have a total of six covariates and $2^6 - 1 = 63$ candidate linear regression models.

Table 3 contains the models selected by different methods. For example, UBERIC method with $\lambda_n = 2$ selects the model with the second and fifth covariates. It is clear that all methods consider fat intake as an important factor that is highly relevant with BMI, while UBERIC with $\lambda_n = \log(n)pn^{1/2}$ selects the most parsimonious model. Taking advantage of the validation data, we further evaluated the performance of these methods using the prediction error, and Figure 3 shows clearly that our methods based on UBERIC have the smallest squared prediction errors among all competing methods.

4 Concluding remarks

In the linear measurement error model context, we have developed a model selection criterion based on minimizing prediction errors, despite the fact that individual predictions are not computable. With probability approaching one, the procedure selects the most parsimonious model among all the correct models if there are correct models included in the candidate models, and achieves the minimum prediction error if no correct model is included.

One interesting question is how to generalize the procedure to the case where $p > n$. This will be impossible if all p covariates contain error because the error variability will dominate the risk. However, if the number of covariates with measurement errors is much smaller than n, it may be possible to identify the best model. Of course, possibly solving this problem will involve additional assumptions and techniques that can be nontrivial.

It will also be interesting yet challenging to generalize the above procedure to more complex models in the measurement error context. Much work is needed in the model selection area in measurement error models and we hope this work can be a starting point in this research domain.

Acknowledgments

Zhang was supported by the grants from the National Natural Science Foundation of China (Grant Nos. 71522004, 11471324 and 11271355). Ma was supported by the National Science Foundation (DMS-1608540). Carroll was supported by a grant from the National Cancer Institute (U01-CA057030). The authors are grateful to the co-Editor, an associate editor and two referees for their constructive comments.

Appendix: Proofs

A.1 Proof of Theorem 1

To prove Theorem 1, we first establish the following two lemmas. The first lemma is the Lemma 1 in Stein (1981). We list it here for completeness and skip the proof. The proof of Lemma 2 is in the supplementary materials.

Lemma 1. *(Stein, 1981): Let* a *be a Normal*$(0,1)$ *random variable and* $g(a) : \mathcal{R} \to \mathcal{R}$ *be an indefinite integral of the Lebesgue measurable function* $\dot{g}(a)$*, essentially the derivative of* $g(a)$*. Suppose also that* $E|\dot{g}(a)| < \infty$*. Then* $E\{\dot{g}(a)\} = E\{ag(a)\}$*.*

Lemma 2. *Let* $\mathbf{M}_1$ *and* $\mathbf{M}_2$ *be* $n \times p_{(s)}$ *matrices,* $\mathbf{M}_3$ *be an* $n \times n$ *matrix, and* $\mathbf{M}_4$ *be a* $p_{(s)} \times p_{(s)}$ *matrix, consisting of functions of* $\breve{\mathbf{U}}_{(s)}$*, where* $\breve{\mathbf{U}}_{(s)}$ *is a* n *by* $p_{(s)}$ *matrix consisting of independent standard normal random variables. The following equations hold,*

$$Etr(\mathbf{M}_1\breve{\mathbf{U}}_{(s)}^{\mathrm{T}}) = Etr(\mathrm{D}\mathbf{M}_1^{\mathrm{T}}), \tag{7}$$

$$tr\{(\mathbf{M}_1 \otimes \mathbf{M}_2^{\mathrm{T}})K_{p_{(s)}n}\} = tr\{K_{np_{(s)}}(\mathbf{M}_2^{\mathrm{T}} \otimes \mathbf{M}_1)\} = tr(\mathbf{M}_1\mathbf{M}_2^{\mathrm{T}}), \tag{8}$$

$$tr[\{\mathbf{M}_4\mathrm{d}\breve{\mathbf{U}}_{(s)}^{\mathrm{T}}\mathbf{M}_3\}/\mathrm{d}\breve{\mathbf{U}}_{(s)}^{\mathrm{T}}] = tr(\mathbf{M}_3)tr(\mathbf{M}_4), \tag{9}$$

$$tr[\{\mathbf{M}_1^{\mathrm{T}}\mathrm{d}\breve{\mathbf{U}}_{(s)}\mathbf{M}_2^{\mathrm{T}}\}/\mathrm{d}\breve{\mathbf{U}}_{(s)}^{\mathrm{T}}] = tr(\mathbf{M}_1^{\mathrm{T}}\mathbf{M}_2), \tag{10}$$

$$tr[\{tr(\mathbf{M}_1\mathrm{d}\breve{\mathbf{U}}_{(s)}^{\mathrm{T}}\mathbf{M}_3)\mathbf{M}_2^{\mathrm{T}}\}/\mathrm{d}\breve{\mathbf{U}}_{(s)}^{\mathrm{T}}] = tr(\mathbf{M}_3\mathbf{M}_1\mathbf{M}_2^{\mathrm{T}}), \tag{11}$$

where $K_{np_{(s)}}$ *and* $K_{p_{(s)}n}$ *are the* $np_{(s)} \times np_{(s)}$ *commutation matrices (Magnus and Neudecker, 1979).*

Proof of Theorem 1:

Here we adopt the matrix differential theory in Magnus and Neudecker (2007), and denote $\mathrm{d}\mathbf{M}$ as the differential of $\mathbf{M}$ with respect to $\breve{\mathbf{U}}_{(s)}^{\mathrm{T}}$, where $\mathbf{M}$ is a matrix of differentiable functions of $\breve{\mathbf{U}}_{(s)}^{\mathrm{T}}$. Denote the Jacobian matrix of $\mathbf{M}$ at $\breve{\mathbf{U}}_{(s)}^{\mathrm{T}}$ as

$$\mathrm{D}\mathbf{M} = \frac{\mathrm{d}\mathbf{M}}{\mathrm{d}\breve{\mathbf{U}}_{(s)}^{\mathrm{T}}} = \frac{\partial \mathrm{vec}\mathbf{M}}{\partial(\mathrm{vec}\breve{\mathbf{U}}_{(s)}^{\mathrm{T}})^{\mathrm{T}}}. \tag{12}$$

For simplicity, we use $\frac{\mathbf{M}}{\breve{\mathbf{U}}_{(s)}^{\mathrm{T}}}$ or $\mathbf{M}/\breve{\mathbf{U}}_{(s)}^{\mathrm{T}}$ to denote the matrix $[(\mathrm{vec}\mathbf{M})_i/\{(\mathrm{vec}\breve{\mathbf{U}}_{(s)}^{\mathrm{T}})^{\mathrm{T}}\}_j]$. This helps present matrix differentiation. For example, $\mathbf{M}\mathrm{d}\breve{\mathbf{U}}_{(s)}^{\mathrm{T}}/\mathrm{d}\breve{\mathbf{U}}_{(s)}^{\mathrm{T}}$ is the Jacobian matrix of $\mathbf{M}\breve{\mathbf{U}}_{(s)}^{\mathrm{T}}$ at $\breve{\mathbf{U}}_{(s)}^{\mathrm{T}}$ with $\mathbf{M}$ treated as a constant matrix although it can be a matrix of functions of $\breve{\mathbf{U}}_{(s)}^{\mathrm{T}}$, i.e., $(\mathbf{M}\mathrm{d}\breve{\mathbf{U}}_{(s)}^{\mathrm{T}})/\mathrm{d}\breve{\mathbf{U}}_{(s)}^{\mathrm{T}} = \{\mathbf{I}_n \otimes \mathbf{M}\mathrm{vec}(\breve{\mathbf{U}}_{(s)}^{\mathrm{T}})\}/\mathrm{d}\breve{\mathbf{U}}_{(s)}^{\mathrm{T}} = \mathbf{I}_n \otimes \mathbf{M}$. Equipped with Lemmas 1 and 2, we now provide the proof of Theorem 1.

Let $\breve{\mathbf{U}}_{(s)} = \mathbf{U}_{(s)}\mathbf{\Sigma}_{(s)}^{-1/2}$, an n by $p_{(s)}$ matrix consisting of independent standard normal random variables. For the case that some covariates are error free, $\mathbf{\Sigma}_{(s)}^{-1/2}$ is defined as a matrix with its upper-left submatrix being the nonzero upper-left submatrix of $\mathbf{\Sigma}_{(s)}$ to the power of $-1/2$ and other elements being 0. Then,

$$\begin{aligned}
&\sigma^{-2}cov(\widehat{\boldsymbol{\mu}}_{(s)}^{\mathrm{T}}, \mathbf{Y}^{\mathrm{T}}) \\
&= \sigma^{-2}E[\{\widehat{\boldsymbol{\mu}}_{(s)} - E(\widehat{\boldsymbol{\mu}}_{(s)})\}^{\mathrm{T}}(\mathbf{Y} - \boldsymbol{\mu})] = \sigma^{-2}E(\widehat{\boldsymbol{\mu}}_{(s)}^{\mathrm{T}}\boldsymbol{\epsilon}) = \sigma^{-2}E\{\boldsymbol{\epsilon}^{\mathrm{T}}\mathbf{X}_{(s)}\mathbf{G}_{(s)}(\boldsymbol{\mu} + \boldsymbol{\epsilon})\} \\
&= E\{\mathrm{tr}(\mathbf{X}_{(s)}\mathbf{G}_{(s)})\} = E\{\mathrm{tr}(\mathbf{W}_{(s)}\mathbf{G}_{(s)} - \mathbf{U}_{(s)}\mathbf{G}_{(s)})\} \\
&= E\{\mathrm{tr}(\mathbf{H}_{(s)}) - \mathrm{tr}(\mathbf{U}_{(s)}\mathbf{G}_{(s)})\} = E\{\mathrm{tr}(\mathbf{H}_{(s)}) - \mathrm{tr}(\mathbf{G}_{(s)}^{\mathrm{T}}\mathbf{\Sigma}_{(s)}^{1/2}\breve{\mathbf{U}}^{\mathrm{T}})\} \\
&= E[\mathrm{tr}(\mathbf{H}_{(s)}) - \mathrm{tr}\{\mathrm{D}(\mathbf{\Sigma}_{(s)}^{1/2}\mathbf{G}_{(s)})\}], \qquad (13)
\end{aligned}$$

where the last equation is from (7) of Lemma 2. Note that

$$\begin{aligned}
\mathrm{d}(\mathbf{\Sigma}_{(s)}^{1/2}\mathbf{G}_{(s)}) &= \mathbf{\Sigma}_{(s)}^{1/2}\{\mathbf{\Lambda}_{(s)}\mathrm{d}\mathbf{W}_{(s)}^{\mathrm{T}} + \mathrm{d}\mathbf{\Lambda}_{(s)}\mathbf{W}_{(s)}^{\mathrm{T}}\} \\
&= \mathbf{\Sigma}_{(s)}^{1/2}\{\mathbf{\Lambda}_{(s)}\mathbf{\Sigma}_{(s)}^{1/2}\mathrm{d}\breve{\mathbf{U}}^{\mathrm{T}} - \mathbf{\Lambda}_{(s)}\mathrm{d}(\mathbf{W}_{(s)}^{\mathrm{T}}\mathbf{W}_{(s)})\mathbf{G}_{(s)}\} \\
&= \mathbf{\Sigma}_{(s)}^{1/2}\{\mathbf{\Lambda}_{(s)}\mathbf{\Sigma}_{(s)}^{1/2}\mathrm{d}\breve{\mathbf{U}}^{\mathrm{T}} - \mathbf{\Lambda}_{(s)}\mathbf{\Sigma}_{(s)}^{1/2}\mathrm{d}\breve{\mathbf{U}}^{\mathrm{T}}\mathbf{H}_{(s)} - \mathbf{G}_{(s)}\mathrm{d}\breve{\mathbf{U}}\mathbf{\Sigma}_{(s)}^{1/2}\mathbf{G}_{(s)}\} \\
&= \mathbf{\Sigma}_{(s)}^{1/2}\mathbf{\Lambda}_{(s)}\mathbf{\Sigma}_{(s)}^{1/2}\mathrm{d}\breve{\mathbf{U}}^{\mathrm{T}} - \mathbf{\Sigma}_{(s)}^{1/2}\mathbf{\Lambda}_{(s)}\mathbf{\Sigma}_{(s)}^{1/2}\mathrm{d}\breve{\mathbf{U}}^{\mathrm{T}}\mathbf{H}_{(s)} - \mathbf{\Sigma}_{(s)}^{1/2}\mathbf{G}_{(s)}\mathrm{d}\breve{\mathbf{U}}\mathbf{\Sigma}_{(s)}^{1/2}\mathbf{G}_{(s)}.
\end{aligned}$$

So from (9) and (10) of Lemma 2 we have

$$\begin{aligned}
\mathrm{tr}\{\mathrm{D}(\mathbf{\Sigma}_{(s)}^{1/2}\mathbf{G}_{(s)})\} &= \mathrm{tr}\{\mathrm{d}(\mathbf{\Sigma}_{(s)}^{1/2}\mathbf{G}_{(s)})/\mathrm{d}\breve{\mathbf{U}}_{(s)}^{\mathrm{T}}\} \\
&= \mathrm{tr}(\mathbf{\Sigma}_{(s)}^{1/2}\mathbf{\Lambda}_{(s)}\mathbf{\Sigma}_{(s)}^{1/2})\mathrm{tr}(\mathbf{I}_n) - \mathrm{tr}(\mathbf{\Sigma}_{(s)}^{1/2}\mathbf{\Lambda}_{(s)}\mathbf{\Sigma}_{(s)}^{1/2})\mathrm{tr}(\mathbf{H}_{(s)}) - \mathrm{tr}(\mathbf{G}_{(s)}^{\mathrm{T}}\mathbf{\Sigma}_{(s)}\mathbf{G}_{(s)}) \\
&= n\mathrm{tr}(\mathbf{\Sigma}_{(s)}\mathbf{\Lambda}_{(s)}) - \mathrm{tr}(\mathbf{H}_{(s)})\mathrm{tr}(\mathbf{\Sigma}_{(s)}\mathbf{\Lambda}_{(s)}) - \mathrm{tr}(\mathbf{G}_{(s)}^{\mathrm{T}}\mathbf{\Sigma}_{(s)}\mathbf{G}_{(s)}). \qquad (14)
\end{aligned}$$

Note further that

$$n\mathbf{\Sigma}_{(s)}\mathbf{\Lambda}_{(s)} = \{\mathbf{W}_{(s)}^{\mathrm{T}}\mathbf{W}_{(s)} - (\mathbf{W}_{(s)}^{\mathrm{T}}\mathbf{W}_{(s)} - n\mathbf{\Sigma}_{(s)})\}\mathbf{\Lambda}_{(s)} = \mathbf{W}_{(s)}^{\mathrm{T}}\mathbf{W}_{(s)}\mathbf{\Lambda}_{(s)} - \mathbf{I}_{p_{(s)}} \qquad (15)$$

and

$$n\mathbf{G}_{(s)}^{\mathrm{T}}\mathbf{\Sigma}_{(s)}\mathbf{G}_{(s)} = \mathbf{W}_{(s)}\mathbf{\Lambda}_{(s)}\{\mathbf{W}_{(s)}^{\mathrm{T}}\mathbf{W}_{(s)} - (\mathbf{W}_{(s)}^{\mathrm{T}}\mathbf{W}_{(s)} - n\mathbf{\Sigma}_{(s)})\}\mathbf{\Lambda}_{(s)}\mathbf{W}_{(s)}^{\mathrm{T}}$$

$$= \mathbf{H}_{(s)}^2 - \mathbf{H}_{(s)}. \tag{16}$$

This leads to

$$\mathrm{tr}(\mathbf{\Sigma}_{(s)}\mathbf{\Lambda}_{(s)}) = \frac{\mathrm{tr}(\mathbf{W}_{(s)}^{\mathrm{T}}\mathbf{W}_{(s)}\mathbf{\Lambda}_{(s)} - \mathbf{I}_{p_{(s)}})}{n} = \frac{t_{(s),1} - p_{(s)}}{n} \tag{17}$$

and

$$\mathrm{tr}(\mathbf{G}_{(s)}^{\mathrm{T}}\mathbf{\Sigma}_{(s)}\mathbf{G}_{(s)}) = \frac{\mathbf{H}_{(s)}^2 - \mathbf{H}_{(s)}}{n} = \frac{t_{(s),2} - t_{(s),1}}{n}. \tag{18}$$

The proof is completed by combining the results in (13), (14), (17) and (18). □

A.2 Proof of Theorem 3

We use these results frequently in the proofs. For two matrices $\mathbf{M}_1$ and $\mathbf{M}_2$ that are comfortable for multiplication,

$$\|\mathbf{M}_1\mathbf{M}_2\| = \mathrm{tr}^{1/2}(\mathbf{M}_2^{\mathrm{T}}\mathbf{M}_1^{\mathrm{T}}\mathbf{M}_1\mathbf{M}_2) \le \lambda_{\max}(\mathbf{M}_1)\mathrm{tr}^{1/2}(\mathbf{M}_2^{\mathrm{T}}\mathbf{M}_2) = \lambda_{\max}(\mathbf{M}_1)\|\mathbf{M}_2\|. \tag{19}$$

Furthermore, if $\mathbf{M}_1$ and $\mathbf{M}_2$ are square matrices and $\mathbf{M}_2 \ge 0$, then

$$\mathrm{tr}(\mathbf{M}_1\mathbf{M}_2) = \mathrm{tr}(\mathbf{M}_2^{1/2}\mathbf{M}_1\mathbf{M}_2^{1/2}) \le \lambda_{\max}(\mathbf{M}_1)\mathrm{tr}(\mathbf{M}_2). \tag{20}$$

From Markov's inequality and Conditions (C.3) and (C.5), we have

$$P\left(\frac{\|\mathbf{U}_{(s)}^{\mathrm{T}}\mathbf{Y}\|}{\sqrt{np}} > M\right) \le \frac{E\|\mathbf{U}_{(s)}^{\mathrm{T}}\mathbf{Y}\|^2}{npM^2} = \frac{E(\mathbf{Y}^{\mathrm{T}}\mathbf{U}_{(s)}\mathbf{U}_{(s)}^{\mathrm{T}}\mathbf{Y})}{npM^2} = \frac{(n\sigma^2 + \|\boldsymbol{\mu}\|^2)\mathrm{tr}(\Sigma_{(s)})}{npM^2} \to 0,$$

as $M \to \infty$. So $\|\mathbf{U}_{(s)}^{\mathrm{T}}\mathbf{Y}\| = O_p\{(np)^{1/2}\}$ uniformly in s. Similarly, under Conditions (C.3), (C.5) and (C.6), $\|\mathbf{W}^{\mathrm{T}}\mathbf{W}\| = O_p(np)$, $\|\mathbf{U}^{\mathrm{T}}\mathbf{X}\| = O_p(n^{1/2}p)$, $\|\mathbf{W}^{\mathrm{T}}\mathbf{X}\| = O_p(np)$, $\boldsymbol{\mu}^{\mathrm{T}}\boldsymbol{\epsilon} = O_p(n^{1/2})$, $\|\mathbf{Y}\|^2 = O_p(n)$, $\|\widehat{\boldsymbol{\epsilon}}_{(s)}\|^2 = O_p(n)$, $\|\mathbf{U}_{(s)}^{\mathrm{T}}\mathbf{U}_{(s)} - n\mathbf{\Sigma}_{(s)}\| = O_p(n^{1/2}p)$, and

$$\begin{aligned}\|\mathbf{W}_{(s)}^{\mathrm{T}}\mathbf{Y}\| \le& \|\mathbf{X}_{(s)}^{\mathrm{T}}\boldsymbol{\mu}\| + \|\mathbf{X}_{(s)}^{\mathrm{T}}\boldsymbol{\epsilon}\| + \|\mathbf{U}_{(s)}^{\mathrm{T}}\mathbf{Y}\| \\ \le& \lambda_{\max}(\mathbf{X}_{(s)})\|\mathbf{Y}\| + \|\mathbf{U}_{(s)}^{\mathrm{T}}\mathbf{Y}\| = O_p(n) + O_p\{(np)^{1/2}\} = O_p(n).\end{aligned}$$

From Condition (C.3), and Theorems 3.21 and 4.21 of Schott (2005), uniformly in s,

$$\begin{aligned}
&\lambda_{\max}(\mathbf{W}_{(s)}^{\mathrm{T}}\mathbf{W}_{(s)}) \\
&= \lambda_{\max}(\mathbf{X}_{(s)}^{\mathrm{T}}\mathbf{X}_{(s)} + \mathbf{U}_{(s)}^{\mathrm{T}}\mathbf{X}_{(s)} + \mathbf{X}_{(s)}^{\mathrm{T}}\mathbf{U}_{(s)} + \mathbf{U}_{(s)}^{\mathrm{T}}\mathbf{U}_{(s)}) \\
&\le \lambda_{\max}(\mathbf{X}_{(s)}^{\mathrm{T}}\mathbf{X}_{(s)}) + \lambda_{\max}(n\mathbf{\Sigma}_{(s)}) + \lambda_{\max}(\mathbf{U}_{(s)}^{\mathrm{T}}\mathbf{X}_{(s)} + \mathbf{X}_{(s)}^{\mathrm{T}}\mathbf{U}_{(s)} + \mathbf{U}_{(s)}^{\mathrm{T}}\mathbf{U}_{(s)} - n\mathbf{\Sigma}_{(s)}) \\
&\le 2nc_2 + 2\|\mathbf{X}_{(s)}^{\mathrm{T}}\mathbf{U}_{(s)}\| + \|\mathbf{U}_{(s)}^{\mathrm{T}}\mathbf{U}_{(s)} - n\mathbf{\Sigma}_{(s)}\| \\
&= 2nc_2 + O_p(n^{1/2}p),
\end{aligned}$$

and $\lambda_{\min}(\mathbf{W}_{(s)}^{\mathrm{T}}\mathbf{W}_{(s)})$

$$\begin{aligned}
&= \lambda_{\min}(\mathbf{X}_{(s)}^{\mathrm{T}}\mathbf{X}_{(s)} + \mathbf{U}_{(s)}^{\mathrm{T}}\mathbf{X}_{(s)} + \mathbf{X}_{(s)}^{\mathrm{T}}\mathbf{U}_{(s)} + \mathbf{U}_{(s)}^{\mathrm{T}}\mathbf{U}_{(s)}) \\
&\ge \lambda_{\min}(\mathbf{X}_{(s)}^{\mathrm{T}}\mathbf{X}_{(s)}) + \lambda_{\min}(n\mathbf{\Sigma}_{(s)}) + \lambda_{\min}(\mathbf{U}_{(s)}^{\mathrm{T}}\mathbf{X}_{(s)} + \mathbf{X}_{(s)}^{\mathrm{T}}\mathbf{U}_{(s)} + \mathbf{U}_{(s)}^{\mathrm{T}}\mathbf{U}_{(s)} - n\mathbf{\Sigma}_{(s)}) \\
&\ge 2nc_1 - 2\|\mathbf{X}_{(s)}^{\mathrm{T}}\mathbf{U}_{(s)}\| - \|\mathbf{U}_{(s)}^{\mathrm{T}}\mathbf{U}_{(s)} - n\mathbf{\Sigma}_{(s)}\| \\
&= 2nc_1 + O_p(n^{1/2}p).
\end{aligned}$$

So under Conditions (C.3) and (C.6),

$$2c_1 + o_p(1) < \frac{\lambda_{\min}(\mathbf{W}_{(s)}^{\mathrm{T}}\mathbf{W})}{n} \le \frac{\lambda_{\max}(\mathbf{W}_{(s)}^{\mathrm{T}}\mathbf{W})}{n} < 2c_2 + o_p(1). \tag{21}$$

Similarly, it can be shown that under Conditions (C.3), (C.4) and (C.6),

$$c_1 + o_p(1) < \frac{\lambda_{\min}(\mathbf{W}_{(s)}^{\mathrm{T}}\mathbf{W}_{(s)} - n\widehat{\mathbf{\Sigma}}_{(s)})}{n} \le \frac{\lambda_{\max}(\mathbf{W}_{(s)}^{\mathrm{T}}\mathbf{W}_{(s)} - n\widehat{\mathbf{\Sigma}}_{(s)})}{n} < c_2 + o_p(1),$$

which implies

$$\lambda_{\max}\{(\mathbf{W}_{(s)}^{\mathrm{T}}\mathbf{W}_{(s)} - n\widehat{\mathbf{\Sigma}}_{(s)})^{-1}\} = O_p\left(n^{-1}\right). \tag{22}$$

From (19), (21) and (22), we obtain that uniformly in s,

$$\begin{aligned}
&\|(\mathbf{X}_{(s)}^{\mathrm{T}}\mathbf{X}_{(s)})^{-1} - (\mathbf{W}_{(s)}^{\mathrm{T}}\mathbf{W}_{(s)} - n\widehat{\mathbf{\Sigma}}_{(s)})^{-1}\| \\
=&\|(\mathbf{X}_{(s)}^{\mathrm{T}}\mathbf{X}_{(s)})^{-1}(\mathbf{W}_{(s)}^{\mathrm{T}}\mathbf{W}_{(s)} - n\widehat{\mathbf{\Sigma}}_{(s)} - \mathbf{X}_{(s)}^{\mathrm{T}}\mathbf{X}_{(s)})(\mathbf{W}_{(s)}^{\mathrm{T}}\mathbf{W}_{(s)} - n\widehat{\mathbf{\Sigma}}_{(s)})^{-1}\| \\
=&\|(\mathbf{X}_{(s)}^{\mathrm{T}}\mathbf{X}_{(s)})^{-1}(\mathbf{U}_{(s)}^{\mathrm{T}}\mathbf{U}_{(s)} - n\widehat{\mathbf{\Sigma}}_{(s)} - \mathbf{X}_{(s)}^{\mathrm{T}}\mathbf{U}_{(s)} - \mathbf{U}_{(s)}^{\mathrm{T}}\mathbf{X}_{(s)})(\mathbf{W}_{(s)}^{\mathrm{T}}\mathbf{W}_{(s)} - n\widehat{\mathbf{\Sigma}}_{(s)})^{-1}\|
\end{aligned}$$

$$=O(1/n)O_p(n^{1/2}p)O_p(1/n) = O_p(p/n^{3/2}). \tag{23}$$

With Condition (C.3), using (19), (22) and (23) we have

$$\begin{aligned}
&\|\{\widetilde{\mathbf{P}}_{(s)} - \breve{\mathbf{P}}_{(s)}\}\mathbf{Y}\| \\
&\le \|\mathbf{X}_{(s)}\{(\mathbf{X}_{(s)}^{\mathrm{T}}\mathbf{X}_{(s)})^{-1} - (\mathbf{W}_{(s)}^{\mathrm{T}}\mathbf{W}_{(s)} - n\widehat{\mathbf{\Sigma}}_{(s)})^{-1}\}\mathbf{X}_{(s)}^{\mathrm{T}}\mathbf{Y}\| + \|\mathbf{X}_{(s)}(\mathbf{W}_{(s)}^{\mathrm{T}}\mathbf{W}_{(s)} - n\widehat{\mathbf{\Sigma}}_{(s)})^{-1}\mathbf{U}_{(s)}^{\mathrm{T}}\mathbf{Y}\| \\
&= O(n^{1/2})O_p(p/n^{3/2})O_p(n) + O(n^{1/2})O_p(n^{-1})O_p(n^{1/2}p^{1/2}) \\
&= O_p(p)
\end{aligned}$$

and

$$\begin{aligned}
\|\{\widetilde{\mathbf{P}}_{(s)} + \breve{\mathbf{P}}_{(s)}\}\mathbf{Y}\| &\le \|\mathbf{X}_{(s)}(\mathbf{W}_{(s)}^{\mathrm{T}}\mathbf{W}_{(s)} - n\widehat{\mathbf{\Sigma}}_{(s)})^{-1}\mathbf{W}_{(s)}^{\mathrm{T}}\mathbf{Y}\| + \|\mathbf{X}_{(s)}(\mathbf{X}_{(s)}^{\mathrm{T}}\mathbf{X}_{(s)})^{-1}\mathbf{X}_{(s)}^{\mathrm{T}}\mathbf{Y}\| \\
&= O(n^{1/2})O_p(n^{-1})O_p(n) = O_p(n^{1/2}).
\end{aligned}$$

So

$$\begin{aligned}
|\widetilde{L}_{(s)} - \breve{L}_{(s)}| &= |\|\widetilde{\mathbf{P}}_{(s)}\mathbf{Y}\|^2 - \|\breve{\mathbf{P}}_{(s)}\mathbf{Y}\|^2 - 2\boldsymbol{\mu}^{\mathrm{T}}\{\widetilde{\mathbf{P}}_{(s)} - \breve{\mathbf{P}}_{(s)}\}\mathbf{Y}| \\
&\le |\mathbf{Y}^{\mathrm{T}}\{\widetilde{\mathbf{P}}_{(s)} - \breve{\mathbf{P}}_{(s)}\}^{\mathrm{T}}\{\widetilde{\mathbf{P}}_{(s)} + \breve{\mathbf{P}}_{(s)}\}\mathbf{Y}| + 2|\boldsymbol{\mu}^{\mathrm{T}}\{\widetilde{\mathbf{P}}_{(s)} - \breve{\mathbf{P}}_{(s)}\}\mathbf{Y}| \\
&\le \|\{\widetilde{\mathbf{P}}_{(s)} - \breve{\mathbf{P}}_{(s)}\}\mathbf{Y}\|\|(\widetilde{\mathbf{P}}_{(s)} + \breve{\mathbf{P}}_{(s)})\mathbf{Y}\| + 2\|\{\widetilde{\mathbf{P}}_{(s)} - \breve{\mathbf{P}}_{(s)}\}\mathbf{Y}\|\|\boldsymbol{\mu}\| \\
&= O_p(n^{1/2}p).
\end{aligned} \tag{24}$$

From (21) and (22), we obtain that uniformly in s,

$$\begin{aligned}
\lambda_{\max}(\widetilde{\mathbf{H}}_{(s)}) =&\lambda_{\max}\{\mathbf{W}_{(s)}(\mathbf{W}_{(s)}^{\mathrm{T}}\mathbf{W}_{(s)} - n\widehat{\mathbf{\Sigma}}_{(s)})^{-1}\mathbf{W}_{(s)}^{\mathrm{T}}\} \\
\le&|\lambda_{\max}\{(\mathbf{W}_{(s)}^{\mathrm{T}}\mathbf{W}_{(s)} - n\widehat{\mathbf{\Sigma}}_{(s)})^{-1}\}|\lambda_{\max}(\mathbf{W}_{(s)}\mathbf{W}_{(s)}^{\mathrm{T}}) = O_p(1).
\end{aligned} \tag{25}$$

From (20) and (25), we have

$$\widetilde{t}_{(s),k} = \mathrm{tr}(\widetilde{\mathbf{H}}_{(s)}^k) \le \lambda_{\max}(\widetilde{\mathbf{H}}_{(s)}^k)\mathrm{tr}(\mathbf{I}_{p_{(s)}}) = p_{(s)}\lambda_{\max}^k(\widetilde{\mathbf{H}}_{(s)}) = O_p(p_{(s)}) \tag{26}$$

and

$$\mathbf{Y}^{\mathrm{T}}\widetilde{\mathbf{H}}_{(s)}^k\mathbf{Y} \le \lambda_{\max}^k(\widetilde{\mathbf{H}}_{(s)})\|\mathbf{Y}\|^2 = O_p(n),$$

so

$$\widetilde{B}_{(s)} = O_p(p_{(s)}) \text{ and } \widetilde{df}_{(s)} = p_{(s)} + O_p(p_{(s)}^2/n). \tag{27}$$

Now we use the above results to consider the term $\mathbf{Y}^{\mathrm{T}}(\mathbf{I}_n - \widetilde{\mathbf{H}}_{(s)})\mathbf{Y}$ for $\mathcal{S}_n^I$ and $\mathcal{S}_n^C$ separately. For $\mathcal{S}_n^I$, it follows from (19) and (23) that

$$\begin{aligned}
&\mathbf{Y}^{\mathrm{T}}(\mathbf{I}_n - \widetilde{\mathbf{H}}_{(s)})\mathbf{Y}\\
=&\mathbf{Y}^{\mathrm{T}}\mathbf{Y} - \mathbf{Y}^{\mathrm{T}}\mathbf{W}_{(s)}(\mathbf{X}_{(s)}^{\mathrm{T}}\mathbf{X}_{(s)})^{-1}\mathbf{W}_{(s)}^{\mathrm{T}}\mathbf{Y}\\
&- \mathbf{Y}^{\mathrm{T}}\mathbf{W}_{(s)}\{(\mathbf{W}_{(s)}^{\mathrm{T}}\mathbf{W}_{(s)} - n\widehat{\boldsymbol{\Sigma}}_{(s)})^{-1} - (\mathbf{X}_{(s)}^{\mathrm{T}}\mathbf{X}_{(s)})^{-1}\}\mathbf{W}_{(s)}^{\mathrm{T}}\mathbf{Y}\\
\leq&\mathbf{Y}^{\mathrm{T}}\mathbf{Y} - \mathbf{Y}^{\mathrm{T}}\mathbf{W}_{(s)}(\mathbf{X}_{(s)}^{\mathrm{T}}\mathbf{X}_{(s)})^{-1}\mathbf{W}_{(s)}^{\mathrm{T}}\mathbf{Y} + O_p(np^{1/2})O_p(p/n^{3/2})O_p(np^{1/2})\\
=&\mathbf{Y}^{\mathrm{T}}\mathbf{Y} - \mathbf{Y}^{\mathrm{T}}\mathbf{X}_{(s)}(\mathbf{X}_{(s)}^{\mathrm{T}}\mathbf{X}_{(s)})^{-1}\mathbf{X}_{(s)}^{\mathrm{T}}\mathbf{Y} + O_p(n^{1/2}p^2)\\
=&\|\mathbf{Y} - \breve{\mathbf{P}}_{(s)}\mathbf{Y}\|^2 + O_p(n^{1/2}p^2)\\
=&\|\boldsymbol{\epsilon}\|^2 + \breve{L}_{(s)} + 2\boldsymbol{\epsilon}^{\mathrm{T}}(\boldsymbol{\mu} - \breve{\mathbf{P}}_{(s)}\mathbf{Y}) + O_p(n^{1/2}p^2)\\
=&\|\boldsymbol{\epsilon}\|^2 + \breve{L}_{(s)} + O_p(n^{1/2}p^2).
\end{aligned} \tag{28}$$

For $\mathcal{S}_n^C$, it follows from (19), (20) and (23) that

$$\begin{aligned}
&\mathbf{Y}^{\mathrm{T}}(\mathbf{I}_n - \widetilde{\mathbf{H}}_{(s)})\mathbf{Y}\\
=&\mathbf{Y}^{\mathrm{T}}\mathbf{Y} - \mathbf{Y}^{\mathrm{T}}\mathbf{W}_{(s)}\{(\mathbf{W}_{(s)}^{\mathrm{T}}\mathbf{W}_{(s)} - n\widehat{\boldsymbol{\Sigma}}_{(s)})^{-1} - (\mathbf{X}_{(s)}^{\mathrm{T}}\mathbf{X}_{(s)})^{-1}\}\mathbf{W}_{(s)}^{\mathrm{T}}\mathbf{Y}\\
&- \mathbf{Y}^{\mathrm{T}}\mathbf{W}_{(s)}(\mathbf{X}_{(s)}^{\mathrm{T}}\mathbf{X}_{(s)})^{-1}\mathbf{W}_{(s)}^{\mathrm{T}}\mathbf{Y}\\
=&\mathbf{Y}^{\mathrm{T}}\mathbf{Y} - \mathbf{Y}^{\mathrm{T}}\mathbf{X}_{(s)}\{(\mathbf{W}_{(s)}^{\mathrm{T}}\mathbf{W}_{(s)} - n\widehat{\boldsymbol{\Sigma}}_{(s)})^{-1} - (\mathbf{X}_{(s)}^{\mathrm{T}}\mathbf{X}_{(s)})^{-1}\}\mathbf{X}_{(s)}^{\mathrm{T}}\mathbf{Y}\\
&- \mathbf{Y}^{\mathrm{T}}\mathbf{X}_{(s)}\{(\mathbf{W}_{(s)}^{\mathrm{T}}\mathbf{W}_{(s)} - n\widehat{\boldsymbol{\Sigma}}_{(s)})^{-1} - (\mathbf{X}_{(s)}^{\mathrm{T}}\mathbf{X}_{(s)})^{-1}\}\mathbf{U}_{(s)}^{\mathrm{T}}\mathbf{Y}\\
&- \mathbf{Y}^{\mathrm{T}}\mathbf{U}_{(s)}\{(\mathbf{W}_{(s)}^{\mathrm{T}}\mathbf{W}_{(s)} - n\widehat{\boldsymbol{\Sigma}}_{(s)})^{-1} - (\mathbf{X}_{(s)}^{\mathrm{T}}\mathbf{X}_{(s)})^{-1}\}\mathbf{W}_{(s)}^{\mathrm{T}}\mathbf{Y}\\
&- \mathbf{Y}^{\mathrm{T}}\mathbf{X}_{(s)}(\mathbf{X}_{(s)}^{\mathrm{T}}\mathbf{X}_{(s)})^{-1}\mathbf{X}_{(s)}^{\mathrm{T}}\mathbf{Y} - 2\mathbf{Y}^{\mathrm{T}}\mathbf{X}_{(s)}(\mathbf{X}_{(s)}^{\mathrm{T}}\mathbf{X}_{(s)})^{-1}\mathbf{U}_{(s)}^{\mathrm{T}}\mathbf{Y} - \mathbf{Y}^{\mathrm{T}}\mathbf{U}_{(s)}(\mathbf{X}_{(s)}^{\mathrm{T}}\mathbf{X}_{(s)})^{-1}\mathbf{U}_{(s)}^{\mathrm{T}}\mathbf{Y}\\
=&\mathbf{Y}^{\mathrm{T}}\mathbf{Y} + \mathbf{Y}^{\mathrm{T}}\mathbf{X}_{(s)}(\mathbf{W}_{(s)}^{\mathrm{T}}\mathbf{W}_{(s)} - n\widehat{\boldsymbol{\Sigma}}_{(s)})^{-1}\{\mathbf{W}_{(s)}^{\mathrm{T}}\mathbf{W}_{(s)} - n\widehat{\boldsymbol{\Sigma}}_{(s)} - \mathbf{X}_{(s)}^{\mathrm{T}}\mathbf{X}_{(s)}\}(\mathbf{X}_{(s)}^{\mathrm{T}}\mathbf{X}_{(s)})^{-1}\mathbf{X}_{(s)}^{\mathrm{T}}\mathbf{Y}\\
&- O_p(n)O_p(p/n^{3/2})O_p\{(np)^{1/2}\} - O_p\{(np)^{1/2}\}O_p(p/n^{3/2})O_p(n)\\
&- \mathbf{Y}^{\mathrm{T}}\mathbf{X}_{(s)}(\mathbf{X}_{(s)}^{\mathrm{T}}\mathbf{X}_{(s)})^{-1}\mathbf{X}_{(s)}^{\mathrm{T}}\mathbf{Y} - 2\mathbf{Y}^{\mathrm{T}}\mathbf{X}_{(s)}(\mathbf{X}_{(s)}^{\mathrm{T}}\mathbf{X}_{(s)})^{-1}\mathbf{U}_{(s)}^{\mathrm{T}}\mathbf{Y} + O_p(n)O_p\{(np)^{1/2}\}O_p\{(np)^{1/2}\}
\end{aligned}$$

$$
\begin{aligned}
=&\mathbf{Y}^{\mathrm{T}}\mathbf{Y} + \mathbf{Y}^{\mathrm{T}}\mathbf{X}_{(s)}(\mathbf{X}_{(s)}^{\mathrm{T}}\mathbf{X}_{(s)})^{-1}\{\mathbf{W}_{(s)}^{\mathrm{T}}\mathbf{W}_{(s)} - n\widehat{\mathbf{\Sigma}}_{(s)} - \mathbf{X}_{(s)}^{\mathrm{T}}\mathbf{X}_{(s)}\}(\mathbf{X}_{(s)}^{\mathrm{T}}\mathbf{X}_{(s)})^{-1}\mathbf{X}_{(s)}^{\mathrm{T}}\mathbf{Y}\\
&+ \mathbf{Y}^{\mathrm{T}}\mathbf{X}_{(s)}\{(\mathbf{W}_{(s)}^{\mathrm{T}}\mathbf{W}_{(s)} - n\widehat{\mathbf{\Sigma}}_{(s)})^{-1} - (\mathbf{X}_{(s)}^{\mathrm{T}}\mathbf{X}_{(s)})^{-1}\}\{\mathbf{W}_{(s)}^{\mathrm{T}}\mathbf{W}_{(s)} - n\widehat{\mathbf{\Sigma}}_{(s)} - \mathbf{X}_{(s)}^{\mathrm{T}}\mathbf{X}_{(s)}\}\\
&\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad \times (\mathbf{X}_{(s)}^{\mathrm{T}}\mathbf{X}_{(s)})^{-1}\mathbf{X}_{(s)}^{\mathrm{T}}\mathbf{Y}\\
&- \mathbf{Y}^{\mathrm{T}}\mathbf{X}_{(s)}(\mathbf{X}_{(s)}^{\mathrm{T}}\mathbf{X}_{(s)})^{-1}\mathbf{X}_{(s)}^{\mathrm{T}}\mathbf{Y} - 2\mathbf{Y}^{\mathrm{T}}\mathbf{X}_{(s)}(\mathbf{X}_{(s)}^{\mathrm{T}}\mathbf{X}_{(s)})^{-1}\mathbf{U}_{(s)}^{\mathrm{T}}\mathbf{Y} + O_p(p^2)\\
=&\mathbf{Y}^{\mathrm{T}}\mathbf{Y} + \mathbf{Y}^{\mathrm{T}}\mathbf{X}_{(s)}(\mathbf{X}_{(s)}^{\mathrm{T}}\mathbf{X}_{(s)})^{-1}\{\mathbf{W}_{(s)}^{\mathrm{T}}\mathbf{W}_{(s)} - n\widehat{\mathbf{\Sigma}}_{(s)} - \mathbf{X}_{(s)}^{\mathrm{T}}\mathbf{X}_{(s)}\}(\mathbf{X}_{(s)}^{\mathrm{T}}\mathbf{X}_{(s)})^{-1}\mathbf{X}_{(s)}^{\mathrm{T}}\mathbf{Y}\\
&+ O_p(n)O_p(p/n^{3/2})O_p\{(n)^{1/2}p\}O_p(n^{-1})O_p(n)\\
&- \mathbf{Y}^{\mathrm{T}}\mathbf{X}_{(s)}(\mathbf{X}_{(s)}^{\mathrm{T}}\mathbf{X}_{(s)})^{-1}\mathbf{X}_{(s)}^{\mathrm{T}}\mathbf{Y} - 2\mathbf{Y}^{\mathrm{T}}\mathbf{X}_{(s)}(\mathbf{X}_{(s)}^{\mathrm{T}}\mathbf{X}_{(s)})^{-1}\mathbf{U}_{(s)}^{\mathrm{T}}\mathbf{Y} + O_p(p^2)\\
=&\mathbf{Y}^{\mathrm{T}}\mathbf{Y} + \mathbf{Y}^{\mathrm{T}}\mathbf{X}_{(s)}(\mathbf{X}_{(s)}^{\mathrm{T}}\mathbf{X}_{(s)})^{-1}(\mathbf{X}_{(s)}^{\mathrm{T}}\mathbf{U}_{(s)} + \mathbf{U}_{(s)}^{\mathrm{T}}\mathbf{X}_{(s)} + \mathbf{U}_{(s)}^{\mathrm{T}}\mathbf{U}_{(s)} - n\widehat{\mathbf{\Sigma}}_{(s)})(\mathbf{X}_{(s)}^{\mathrm{T}}\mathbf{X}_{(s)})^{-1}\mathbf{X}_{(s)}^{\mathrm{T}}\mathbf{Y}\\
&- \mathbf{Y}^{\mathrm{T}}\mathbf{X}_{(s)}(\mathbf{X}_{(s)}^{\mathrm{T}}\mathbf{X}_{(s)})^{-1}\mathbf{X}_{(s)}^{\mathrm{T}}\mathbf{Y} - 2\mathbf{Y}^{\mathrm{T}}\mathbf{X}_{(s)}(\mathbf{X}_{(s)}^{\mathrm{T}}\mathbf{X}_{(s)})^{-1}\mathbf{U}_{(s)}^{\mathrm{T}}\mathbf{Y} + O_p(p^2)\\
=&\mathbf{Y}^{\mathrm{T}}\mathbf{Y} - \mathbf{Y}^{\mathrm{T}}\mathbf{X}_{(s)}(\mathbf{X}_{(s)}^{\mathrm{T}}\mathbf{X}_{(s)})^{-1}\mathbf{X}_{(s)}^{\mathrm{T}}\mathbf{Y} + 2\mathbf{Y}^{\mathrm{T}}\mathbf{X}_{(s)}(\mathbf{X}_{(s)}^{\mathrm{T}}\mathbf{X}_{(s)})^{-1}\mathbf{X}_{(s)}^{\mathrm{T}}\mathbf{U}_{(s)}(\mathbf{X}_{(s)}^{\mathrm{T}}\mathbf{X}_{(s)})^{-1}\mathbf{X}_{(s)}^{\mathrm{T}}\mathbf{Y}\\
&+ \boldsymbol{\beta}_{(s)}^{\mathrm{T}}(\mathbf{U}_{(s)}^{\mathrm{T}}\mathbf{U}_{(s)} - n\widehat{\mathbf{\Sigma}}_{(s)})\boldsymbol{\beta}_{(s)} - 2\mathbf{Y}^{\mathrm{T}}\mathbf{X}_{(s)}(\mathbf{X}_{(s)}^{\mathrm{T}}\mathbf{X}_{(s)})^{-1}\mathbf{U}_{(s)}^{\mathrm{T}}\mathbf{Y} + O_p(p^2)\\
=&\|\{\mathbf{I}_n - \breve{\mathbf{P}}_{(s)}\}\mathbf{Y}\|^2 + 2\boldsymbol{\epsilon}^{\mathrm{T}}\mathbf{X}_{(s)}(\mathbf{X}_{(s)}^{\mathrm{T}}\mathbf{X}_{(s)})^{-1}\mathbf{X}_{(s)}^{\mathrm{T}}\mathbf{U}_{(s)}\boldsymbol{\beta}_{(s)} + 2\boldsymbol{\mu}^{\mathrm{T}}\mathbf{X}_{(s)}(\mathbf{X}_{(s)}^{\mathrm{T}}\mathbf{X}_{(s)})^{-1}\mathbf{X}_{(s)}^{\mathrm{T}}\mathbf{U}_{(s)}\boldsymbol{\beta}_{(s)}\\
&+ \boldsymbol{\beta}_{(s)}^{\mathrm{T}}(\mathbf{U}_{(s)}^{\mathrm{T}}\mathbf{U}_{(s)} - n\widehat{\mathbf{\Sigma}}_{(s)})\boldsymbol{\beta}_{(s)} - 2\boldsymbol{\beta}_{(s)}^{\mathrm{T}}\mathbf{U}_{(s)}^{\mathrm{T}}\boldsymbol{\mu} - 2\boldsymbol{\beta}_{(s)}\mathbf{U}_{(s)}^{\mathrm{T}}\boldsymbol{\epsilon} + O_p(p^2)\\
=&\|\{\mathbf{I}_n - \breve{\mathbf{P}}_{(s)}\}\mathbf{Y}\|^2 + \boldsymbol{\beta}_{(s)}^{\mathrm{T}}(\mathbf{U}_{(s)}^{\mathrm{T}}\mathbf{U}_{(s)} - n\widehat{\mathbf{\Sigma}}_{(s)})\boldsymbol{\beta}_{(s)} - 2\boldsymbol{\beta}_{(s)}^{\mathrm{T}}\mathbf{U}_{(s)}^{\mathrm{T}}\boldsymbol{\epsilon} + O_p(p^2)\\
=&\|\boldsymbol{\epsilon}\|^2 - \boldsymbol{\epsilon}^{\mathrm{T}}\mathbf{P}_{(s)}\boldsymbol{\epsilon} + \boldsymbol{\beta}_{(s)}^{\mathrm{T}}(\mathbf{U}_{(s)}^{\mathrm{T}}\mathbf{U}_{(s)} - n\widehat{\mathbf{\Sigma}}_{(s)})\boldsymbol{\beta}_{(s)} - 2\boldsymbol{\beta}_{(s)}^{\mathrm{T}}\mathbf{U}_{(s)}^{\mathrm{T}}\boldsymbol{\epsilon} + O_p(p^2)\\
=&\|\boldsymbol{\epsilon}\|^2 + O_p(n^{1/2}p). \qquad (29)
\end{aligned}
$$

Note that Condition (C.1) implies

$$\max_{s\in\mathcal{S}_n^I}\frac{n^{1/2}p^2}{\breve{L}_{(s)}} = o_p(1). \qquad (30)$$

From (24), (27), (28) and (30), we have

$$C_\lambda(s) - \|\boldsymbol{\epsilon}\|^2 = \breve{L}_{(s)} + o_p(\breve{L}_{(s)}) + \lambda_n\widehat{\sigma}^2\widetilde{df}_{(s)} \text{ uniformly for } \mathcal{S}_n^I. \qquad (31)$$

By using (27) and Conditions (C.4) and (C.2) we have

$$\max_{s\in\mathcal{S}_n^I}\frac{\lambda_n\widehat{\sigma}^2\widetilde{df}_{(s)}}{\breve{L}_{(s)}} = o_p(1),$$

and thus

$$C_\lambda(s) - \|\boldsymbol{\epsilon}\|^2 = \breve{L}_{(s)} + o_p(\breve{L}_{(s)}) \text{ uniformly for } \mathcal{S}_n^I. \tag{32}$$

Condition (C.1) and (24) imply that $\widetilde{L}_{(s)}/\breve{L}_{(s)} \to 1$ in probability, which, together with (32), indicate that the criterion is asymptotic optimal if $\mathcal{S}_n^C$ is empty.

If $\mathcal{S}_n^C$ is not empty, then from (27) and (29) we have

$$C_\lambda(s) - \|\boldsymbol{\epsilon}\|^2 = O_p(n^{1/2}p) + \lambda_n \widehat{\sigma}^2 \widetilde{df}_{(s)} \text{ uniformly for } \mathcal{S}_n^C. \tag{33}$$

If $\lambda_n/(n^{1/2}p) \to \infty$, then from (27) and Conditions (C.4) and (C.6), (33) can be written as

$$C_\lambda(s) - \|\boldsymbol{\epsilon}\|^2 = \lambda_n \widehat{\sigma}^2 \{p_{(s)} + o_P(1)\} \text{ uniformly for } \mathcal{S}_n^C. \tag{34}$$

So the criterion picks the smallest model among correct models. Furthermore, by using (33), (34) and Condition (C.2) we have

$$\frac{\max_{\mathcal{S}^C}\{C_\lambda(s) - \|\boldsymbol{\epsilon}\|^2\}}{\min_{\mathcal{S}^I}\{C_\lambda(s) - \|\boldsymbol{\epsilon}\|^2\}} = o_p(1). \tag{35}$$

Thus the probability of selecting the correct model with the smallest dimension goes to one.

References

Brinton, L. A., Daling, J. R., Liff, J. M., Schoenberg, J. B., Malone, K. E., Stanford, J. L., Coates, R. J., Gammon, M. D., Hanson, L., and Hoover, R. N. (1995), "Oral contraceptives and breast cancer risk among younger women," *Journal of the National Cancer Institute*, 87, 827–835.

Carroll, R. J., Delaigle, A., and Hall, P. (2009), "Nonparametric prediction in measurement error models," *Journal of the American Statistical Association*, 104.

Carroll, R. J., Ruppert, D., Stefanski, L. A., and Crainiceanu, C. M. (2006), *Measurement Error in Nonlinear Models: a Modern Perspective*, Chapman and Hall/CRC.

Efron, B. (2004), "The estimation of prediction error: Covariance penalties and crossvalidation (with discussion)," *Journal of the American Statistical Association*, 99, 619–642.

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.

Flynn, C., Hurvich, C., and Simonoff, J. (2013), "Efficiency for regularization parameter selection in penalized likelihood estimation of misspecified models," *Journal of the American Statistical Association*, 108, 1031–1043.

Hastie, T. J. and Tibshirani, R. J. (1990), *Generalized Additive Models*, vol. 43, CRC Press.

Li, K. C. (1987), "Asymptotic optimality for C_p, C_l, cross-validation and generalized cross-validation: Discrete index set," *Annals of Statistics*, 15, 958–975.

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., Wu, H. L., and Zou, G. H. (2008), "A note on conditional AIC for linear mixed-effects models," *Biometrika*, 95, 773–778.

Ma, Y. and Li, R. (2010), "Variable selection in measurement error models," *Bernoulli*, 16, 274–300.

Magnus, J. R. and Neudecker, H. (1979), "The commutation matrix: Some properties and applications," *Annals of Statistics*, 7, 381–394.

— (2007), *Matrix Differential Calculus with Applications in Statistics and Econometrics, 3nd Edition*, John Wiley & Sons.

Mukherjee, A., Chen, K., Wang, N., and Zhu, J. (2015), "On the degrees of freedom of reduced-rank estimators in multivariate regression," *Biometrika*, 102, 457–477.

Nusser, S., Carriquiry, A., Dodd, K., and Fuller, W. (1996), "A semiparametric transformation approach to estimating usual daily intake distributions," *Journal of the American Statistical Association*, 91, 1440–1449.

Potischman, N., Carroll, R. J., Iturria, S., Mittl, B., Curtin, J., Thompson, F., and Brinton, L. (1999), "Comparison of the 60- and 100-item NCI Block Questionnaires with validation data," *Nutrition and Cancer*, 34, 70–85.

Schott, J. R. (2005), *Matrix Analysis for Statistics*, Wiley.

Shao, J. (1997), "An asymptotic theory for linear model selection," *Statistica Sinica*, 7, 221–242.

Spiegelman, D., Carroll, R. J., and Kipnis, V. (2001), "Efficient regression calibration for logistic regression in main study/internal validation study designs with an imperfect reference instrument," *Statistics in Medicine*, 20, 139–160.

Stein, C. M. (1981), "Estimation of the mean of a multivariate normal distribution," *Annals of Statistics*, 153, 1135–1151.

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

Wang, H., Zou, G., and Wan, A. T. K. (2012), "Model averaging for varying-coefficient partially linear measurement error models," *Electronic Journal of Statistics*, 6, 1017–1039.

Yi, G., Ma, Y., Spiegelman, D., and Carroll, R. J. (2015), "Functional and structural methods with mixed measurement error and misclassification in covariates," *Journal of the American Statistical Association*, 110, 681–696.

Zhang, Y., Li, R., and Tsai, C.-L. (2010), "Regularization parameter selections via generalized information criterion," *Journal of the American Statistical Association*, 105, 312–323.

Zou, H., Hastie, T., and Tibshirani, R. (2007), “On the degrees of freedom of the lasso,” *Annals of Statistics*, 35, 2173–2192.

Table 1: Simulation results of Example I in Section 3.2: relative loss. Methods compared are UBERIC with $\lambda_n = 2$ and $\lambda_n = \log(n)pn^{1/2}$, AIC$_0$, BIC$_0$, AIC$_1$, BIC$_1$, SCAD-BIC, and SCAD-GCV. The best results are in bold face.

				$\lambda_n =$					SCAD-	SCAD-
n	σ	τ	$\lambda_n = 2$	$\log(n)pn^{1/2}$	AIC$_0$	BIC$_0$	AIC$_1$	BIC$_1$	BIC	GCV
25	0.5	0.85	**1.390**	4.094	5.668	5.612	1.753	1.674	5.218	5.108
		0.95	**1.412**	1.450	1.709	1.661	1.510	1.431	1.820	1.707
	1	0.85	**1.511**	5.321	4.676	4.598	1.743	1.666	5.542	5.319
		0.95	1.465	1.774	1.782	1.705	1.530	**1.424**	1.720	1.619
50	0.5	0.85	**1.350**	2.899	2.021	1.966	1.866	1.791	3.130	2.787
		0.95	**1.356**	1.069	1.564	1.513	1.510	1.391	1.788	1.378
	1	0.85	**1.496**	4.971	2.236	2.155	1.839	1.731	2.711	2.549
		0.95	1.446	**1.068**	1.693	1.600	1.562	1.418	2.049	1.406
100	0.5	0.85	**1.261**	1.512	1.471	1.441	1.952	1.879	2.084	1.851
		0.95	1.192	**1.015**	1.323	1.285	1.466	1.362	1.703	1.299
	1	0.85	**1.348**	3.417	1.602	1.538	1.864	1.787	2.163	1.847
		0.95	1.235	**1.008**	1.378	1.322	1.455	1.348	2.193	1.645
200	0.5	0.85	1.161	**1.006**	1.249	1.233	1.978	1.925	2.129	1.926
		0.95	1.111	**1.006**	1.180	1.163	1.425	1.357	1.596	1.554
	1	0.85	**1.205**	1.553	1.308	1.269	1.930	1.866	2.198	1.937
		0.95	1.139	**1.003**	1.215	1.174	1.427	1.345	2.176	1.692

Table 2: Simulation results of Example II in Section 3.3: frequency in selecting the smallest correct model. Methods compared are UBERIC with $\lambda_n = 2$ and $\lambda_n = \log(n)pn^{1/2}$, AIC$_0$, BIC$_0$, AIC$_1$, BIC$_1$, SCAD-BIC, and SCAD-GCV. The best results are in bold face.

				$\lambda_n =$					SCAD-	SCAD-
n	σ	τ	$\lambda_n = 2$	$\log(n)pn^{1/2}$	AIC$_0$	BIC$_0$	AIC$_1$	BIC$_1$	BIC	GCV
25	0.5	0.85	**0.772**	0.576	0.646	0.832	0.216	0.330	0.628	0.620
		0.95	0.756	**0.976**	0.710	0.862	0.272	0.362	0.754	0.716
	1	0.85	0.758	0.468	0.670	**0.818**	0.268	0.354	0.614	0.606
		0.95	0.744	**0.960**	0.704	0.862	0.290	0.404	0.778	0.736
50	0.5	0.85	0.646	**0.832**	0.492	0.794	0.080	0.190	0.782	0.754
		0.95	0.608	**0.974**	0.552	0.834	0.104	0.206	0.758	0.662
	1	0.85	0.606	0.696	0.480	**0.802**	0.092	0.194	0.774	0.754
		0.95	0.560	**0.990**	0.542	0.842	0.132	0.270	0.742	0.654
100	0.5	0.85	0.584	**0.962**	0.416	0.772	0.042	0.154	0.920	0.872
		0.95	0.590	**0.982**	0.508	0.852	0.086	0.206	0.870	0.788
	1	0.85	0.602	0.872	0.434	0.786	0.082	0.246	**0.906**	0.858
		0.95	0.614	**0.998**	0.540	0.880	0.134	0.320	0.860	0.794
200	0.5	0.85	0.548	**0.984**	0.274	0.718	0.046	0.166	0.974	0.920
		0.95	0.598	**0.996**	0.450	0.874	0.066	0.210	0.958	0.916
	1	0.85	0.544	**0.982**	0.304	0.752	0.088	0.292	0.960	0.918
		0.95	0.584	**1.000**	0.484	0.878	0.146	0.370	0.946	0.888

Table 3: Analysis of WISH data. Selected models by UBERIC with $\lambda_n = 2$ and $\lambda_n = \log(n)pn^{1/2}$, AIC$_0$, BIC$_0$, AIC$_1$, BIC$_1$, SCAD-BIC, and SCAD-GCV.

		$\lambda_n =$					SCAD-	SCAD-
Methods	$\lambda_n = 2$	$\log(n)pn^{1/2}$	AIC$_0$	BIC$_0$	AIC$_1$	BIC$_1$	BIC	GCV
Models	(2,5)	(2)	(1,2,4)	(2,4)	(2,3,5)	(2,3,5)	(3)	(3)

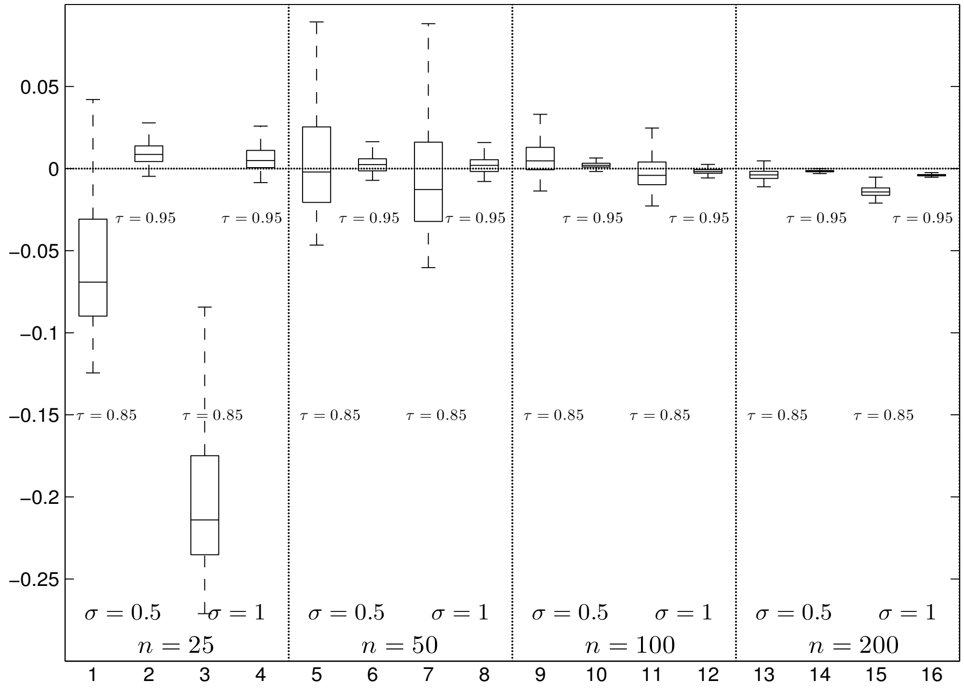

Figure 1: Numerical Example I in Section 3.2. Boxplots of 500 differences $\widehat{df}_{(s)} - df_{(s)}$ with $s = 2^{p-1}$.

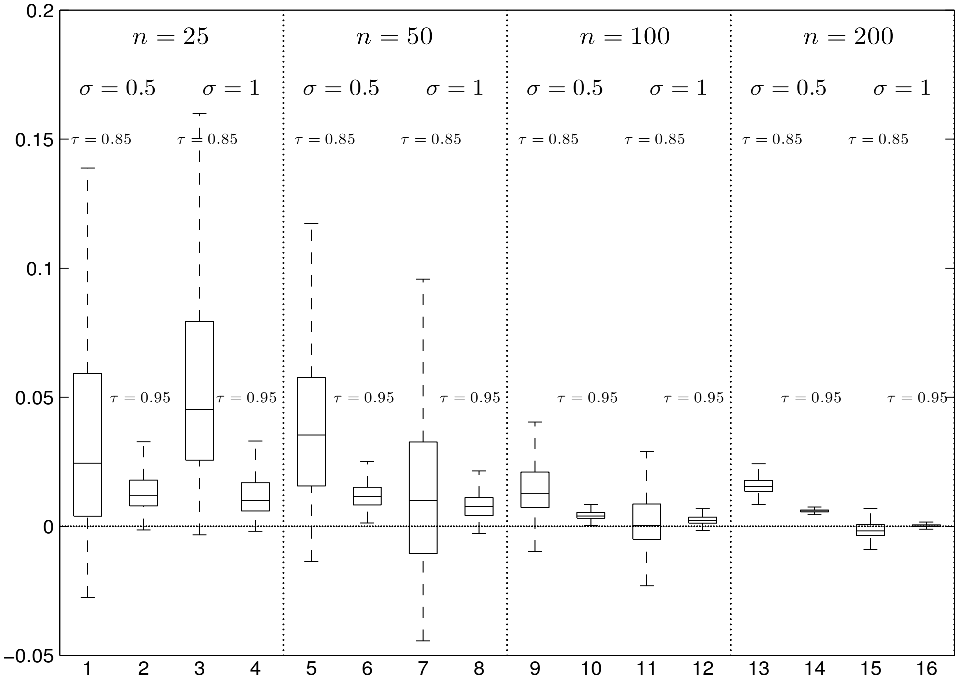

Figure 2: Numerical Example II in Section 3.3. Boxplots of 500 differences $\widehat{df}_{(s)} - df_{(s)}$ with $s = 2^{p-1}$.

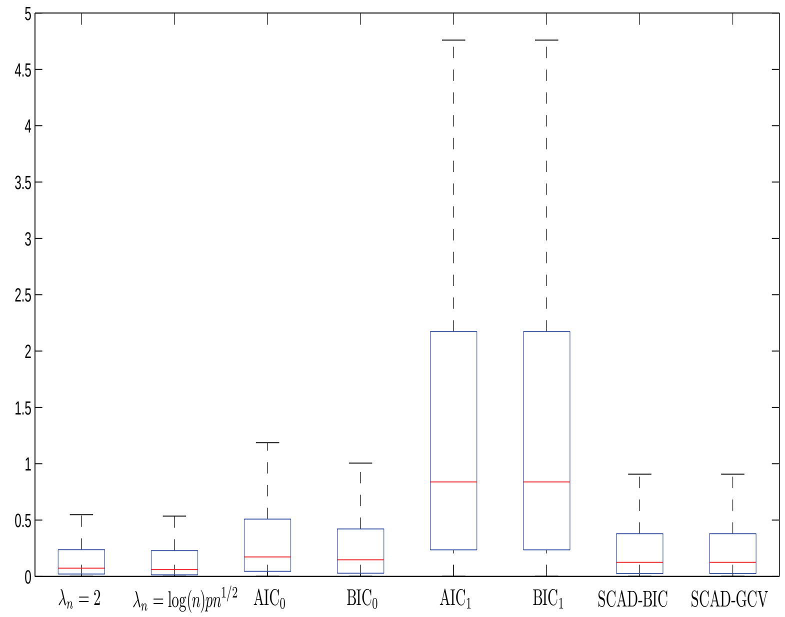

Figure 3: Analysis of WISH data. Boxplots of 192 squared prediction errors. Methods compared are UBERIC with $\lambda_n = 2$ and $\lambda_n = \log(n)pn^{1/2}$, AIC$_0$, BIC$_0$, AIC$_1$, BIC$_1$, SCAD-BIC, and SCAD-GCV.