The Nonparametric Box-Cox Model for High-Dimensional Regression Analysis

He Zhou and Hui Zou*

School of Statistics, University of Minnesota, United States

The mainstream theory for high-dimensional regression assumes that the underlying true model is a low-dimensional linear regression model. On the other hand, a standard technique in regression analysis, even in the traditional low-dimensional setting, is to employ the Box-Cox transformation for reducing anomalies such as non-additivity and heteroscedasticity in linear regression. In this paper, we propose a new high-dimensional regression method based on a nonparametric Box-Cox model with an unspecified monotone transformation function. Model fitting and computation become much more challenging than the usual penalized regression method, and a two-step method is proposed for the estimation of this model in high-dimensional settings. First, we propose a novel technique called composite probit regression (CPR) and use the folded concave penalized CPR for estimating the regression parameters. The strong oracle property of the estimator is established without knowing the nonparametric transformation function. Next, the nonparametric function is estimated by conducting univariate monotone regression. The computation is done efficiently by using a coordinate-majorization-descent algorithm. Extensive simulation studies show that the proposed method performs well in various settings. Our analysis of the supermarket data demonstrates the superior performance of the proposed method over the standard high-dimensional regression method.

*Corresponding author at: 313 Ford Hall, School of Statistics, 224 Church Street SE, Minneapolis, MN 55455, USA. E-mail address: zouxx019@umn.edu. Zou is supported in part by NSF grants 1915-842 and 2015-120.

Keywords: Box-Cox model, Composite estimation, High-dimensional regression, Nonparametric transformation, Strong oracle property.

1 Introduction

In a regression problem, let Y be the response variable and $(x_1, \ldots, x_p)^\intercal$ be the vector of covariates. When p diverges with the sample size or even exceeds the sample size, sparse estimation stands at the center of the stage of high-dimensional linear regression. The LASSO (Tibshirani, 1996) and the concave penalization (Fan and Li, 2001) are two mainstream penalization methods. The key idea is that sparse penalization encourages a sparse estimator and also reduces the estimation variability. A large number of papers have been devoted to the numerical and theoretical study of penalization approaches in sparse estimation. Readers are referred to Fan, Li, Zhang and Zou (2020) for a comprehensive treatment of this topic. The typical theoretical setup for modern high-dimensional regression methods begins with the following true model:

$$Y = \sum_{j \in \mathcal{A}} x_j \beta_j + \varepsilon \tag{1}$$

where $\mathcal{A}$ represents a small subset of important variables and the error ε is assumed to be independent identically distributed (i.i.d.) with a common variance. The theoretical model (1) may be insufficient in practice even in the low-dimensional case. The anomalies such as non-additivity and heteroscedasticity that violate model (1) arise in many situations. As a remedy, Box and Cox (1964) assumed that a (normal) linear model could be appropriate after a parametric power transformation has been applied to the response. Specifically, they modified the family of power transformations introduced by Tukey (1957) and proposed the following transformed linear model:

$$Y^{(\lambda)} = \boldsymbol{x}^\intercal \boldsymbol{\beta} + \varepsilon, \quad \varepsilon \sim N(0, \sigma^2), \tag{2}$$

where the transformation is from the *scaled power family*,

$$Y^{(\lambda)} = \begin{cases} (Y^\lambda - 1)/\lambda, & \text{if } \lambda \neq 0; \\ \log(Y), & \text{if } \lambda = 0; \end{cases}$$

with λ unknown. Box and Cox (1964) discussed the inferences about the transformation parameter λ and about the parameters $\boldsymbol{\beta}$ of the linear model by maximum likelihood and Bayesian analysis. The Box-Cox model was further developed by Draper and Cox (1969), Bickel and Doksum (1981), Carroll and Ruppert (1981), and Carroll (1982), among others. Since that

time, the Box-Cox transformation technique has enjoyed wide practical use and considerable success. Nowadays, it is a standard technique covered in most applied regression textbooks (Draper and Smith, 1998; Weisberg, 2005).

It is interesting and desirable to develop the high-dimensional version of the Box-Cox regression model for applications. In this paper, we consider the following non-parametric transformed linear model:

$$g(Y) = \boldsymbol{x}^\top \boldsymbol{\beta} + \varepsilon, \quad \varepsilon \sim N(0,1), \tag{3}$$

where $g(\cdot)$ is an unspecified monotone increasing function. Because a scaling parameter can be absorbed into the monotone function $g(\cdot)$, we assume that the error variance is one. Notice that the essential idea of the classical Box-Cox model (2) is to achieve additivity, normality and homoscedasticity via data transformation. Our model does the same but avoids using a parametric transformation. Thus, we name (3) the *nonparametric Box-Cox regression model*. In the high-dimensional case, $\boldsymbol{\beta}$ is assumed to be sparse and $\boldsymbol{x}^\top \boldsymbol{\beta} = \sum_{j \in \mathcal{A}} x_j \beta_j$, where $\mathcal{A}$ represents the unknown subset of important variables. We avoid the parametric Box-Cox transformation in our model for some good reasons. First, despite the popularity of Box-Cox transformation, often we are not certain about the correct form of the right transformation. Any pre-chosen parametric form can suffer from mis-specification for a given application. Therefore, it is more desirable to use a data-driven transformation. Second, previous theoretical studies on the Box-Cox model suggested that assuming a parametric transformation does not offer any theoretical advantage. In the usual low-dimensional settings, Bickel and Doksum (1981) pointed out that the cost of estimating λ in the Box-Cox model could be very high. We can only expect this issue becomes more severe in high dimensions.

It is worth pointing out that many researchers have considered more flexible forms of the Box-Cox model in the literature. Han (1987) considered a general transformation model: $Y = D \cdot F(\boldsymbol{x}^\top \boldsymbol{\beta}, u)$, where the composite transformation $D \cdot F$ is only specified that D is non-degenerate monotonic and F is strictly monotonic in each of its variables. For the low-dimensional case, Han (1987) proposed the maximum rank correlation (MRC) estimator for estimating the direction vector of $\boldsymbol{\beta}$, i.e., $\boldsymbol{\beta}/\|\boldsymbol{\beta}\|_2$. The MRC estimator is shown to be root-n consistent and has an asymptotic normal distribution (Sherman, 1993). The model considered in Chen (2002) assumes $g(Y) = \boldsymbol{x}^\top \boldsymbol{\beta} + \varepsilon$ but the distribution of ε is also unspecified. Chen (2002) proposed a rank based estimator for $\boldsymbol{\beta}$ under a normalizing assumption $\beta_1 = 1$. Our model assumes the error is $N(0,1)$ so that we can also directly estimate $\boldsymbol{\beta}$ and then make prediction about Y. Note that the

optimal prediction function for Y under the squared error loss is the conditional expectation of Y given $\boldsymbol{x}$ which depends on the distribution of $\boldsymbol{\varepsilon}$. More importantly, our model can be well estimated under the ultra-high dimension setting, while it remains an open question how to extend the methodology and theory for these more flexible models (Chen, 2002; Han, 1987; Sherman, 1993) to the ultra-high-dimensional setting. Only recently, the rate of convergence of MRC under diverging dimensions was established (Fan, Han, Li and Zhou, 2020) where p is still assumed to be much less than n. The theoretical property of MRC is still unknown when $p \gg n$. Moreover, the computation of MRC is very challenging when p is high, which would be a practical concern. For high-dimensional regression analysis, the non-parametric Box-Cox model is a viable choice and ready to be applied in applications.

We now elaborate more on our model-fitting procedure for the nonparametric Box-Cox model when p is large. Consider an idealized setting where the true $\boldsymbol{\beta}$ is given, then the estimation of g function becomes a univariate monontone regression problem which has been fully studied in the literature. Thus, the most challenging part of model-fitting, especially when p is large, is how to estimate $\boldsymbol{\beta}$ optimally and select the right subset of variables. The classical Box-Cox model is fitted via maximum likelihood in which the transformation parameter and $\boldsymbol{\beta}$ are jointly estimated. Bickel and Doksum (1981) analyzed the maximum likelihood approach and pointed out that the estimation of the regression parameter is unstable and the cost of not knowing λ could be enormous since the inference relies on the estimated transformation parameter. In the high-dimensional setting, this issue cannot be resolved by using a sparse penalization technique in the maximum likelihood approach, because this issue remains even we knew the true support of $\boldsymbol{\beta}$. This motivates us to consider a new approach to estimating $\boldsymbol{\beta}$.

Our solution is a new two-step method for the estimation of the non-parametric Box-Cox regression model (3). This method separates the estimation of the regression parameters and the estimation of the transformation function into two steps. Firstly, we focus on the regression parameter which is the most important and crucial part in high-dimensional regression. Note that we have to deal with a sparse estimation problem with a nonparametric $g(\cdot)$ function in the response. Our estimator of $\boldsymbol{\beta}$ is based on the fact that a probit regression model is obtained from (3) after turning the response variable into a dichotomous variable. Then, a sparse penalized probit regression estimator is derived for estimating $\boldsymbol{\beta}$. As multiple probit regression models can be made from (3), we further combine those estimators via a novel technique called the *composite probit regression (CPR)*. We prove the optimality of our estimator by establishing the strong

oracle property of the folded concave penalized CPR. Moreover, the proposed estimator can be computed efficiently by using the a coordinate-majorization-descent algorithm. Once $\boldsymbol{\beta}$ is well-estimated, the estimation of transformation function $g(\cdot)$ is a univariate regression problem which can be solved by conducting monotone regression on the working data $\{(\boldsymbol{x}_i^\intercal \hat{\boldsymbol{\beta}}, y_i)\}_{i=1}^n$ where $\hat{\boldsymbol{\beta}}$ is the estimated parameter. There are plenty of mature methods in the literature devoted to conducting monotone regression (Dette et al., 2006; Hall and Huang, 2001; Mammen, 1991; Ramsay, 1998). In this work, we apply the monotone smoother proposed by Dette et al. (2006) which has nice asymptotical properties and is implemented in a public R package.

The rest of the paper is organized as follows. In section 2, the model-fitting methodology is discussed with greater details. In section 3, we discuss the folded concave penalized composite probit regression, including its computational as well as theoretical properties. Simulation results are presented in section 4, and in section 5 we apply this new model to analyze the supermarket data. We conclude the paper with a discussion section. All proofs are relegated to an appendix.

2 The Two-Step Methodology

2.1 Estimation of β

Consider estimating the parameter $\boldsymbol{\beta}$ in our model (3) based on n i.i.d. observations $\{(\boldsymbol{x}_i, y_i)\}_{i=1}^n$. Given a user-chosen threshold y_0, we have

$$\mathbb{P}(Y \geq y_0|\boldsymbol{x}) = \mathbb{P}(g(Y) \geq g(y_0)|\boldsymbol{x}) = \mathbb{P}(\varepsilon \geq g(y_0) - \boldsymbol{x}^\intercal\boldsymbol{\beta}|\boldsymbol{x}) = \Phi(-g(y_0) + \boldsymbol{x}^\intercal\boldsymbol{\beta}),$$

where $\Phi(\cdot)$ is the cumulative density function (CDF) of standard normal distribution. This is a direct consequence of the non-parametric Box-Cox model. Thus, we create new response variables $\tilde{y}_i = I_{\{y_i \geq y_0\}}$ and then $\{(\tilde{y}_i, \boldsymbol{x}_i)\}_{i=1}^n$ follow a probit regression model with intercept $-g(y_0)$ and regression coefficient $\boldsymbol{\beta}$. We can consider multiple threshold levels and then obtain several probit regression models. Note that the choice of threshold values only affects the intercept term in the corresponding probit model but not the regression coefficient. In order to borrow strengths from those probit regression models, we propose to simultaneously consider multiple probit regression models and aggregate their results. Let $\{y_0^{(k)}, k = 1, \cdots, K\}$ be the sequence of threshold values, $\tilde{y}_{ki} = I_{\{y_i \geq y_0^{(k)}\}}$, and $\boldsymbol{b}_0 = (b_{10}, \cdots, b_{K0})^\intercal$. The penalized composite

probit regression estimator $\hat{\boldsymbol{\beta}}^{\text{pcpr}}$ is defined as the solution to the following optimization problem:

$$\min_{\boldsymbol{\beta},\boldsymbol{b}_0} \sum_{k=1}^{K} w_k \left\{ \frac{1}{n} \sum_{i=1}^{n} \left[-\tilde{y}_{ki} h(\boldsymbol{x}_i^\top \boldsymbol{\beta} - b_{k0}) - \log(1 - \Phi(\boldsymbol{x}_i^\top \boldsymbol{\beta} - b_{k0})) \right] \right\} + \sum_{j=1}^{p} p_\lambda(|\beta_j|), \tag{4}$$

where the first term is the weighted summation of K negative log-likelihood functions with composite weights w_k's and $h(\eta) = \log\left(\frac{\Phi(\eta)}{1-\Phi(\eta)}\right)$. The penalized composite probit regression shares the spirit of the penalized composite quantile regression (Zou and Yuan, 2008) in which the loss function is the weighted sum of loss functions for several quantile regression models.

The folded concave penalization (Fan and Li, 2001) is used in (4). The local linear approximation algorithm (Zou and Li, 2008) is applied to obtain an estimator with the strong oracle property. See details in section 3.3.

2.2 Estimation of $g(\cdot)$

If we define $Z = \boldsymbol{x}^\top \boldsymbol{\beta}$, then our model (3) becomes

$$Z = g(Y) + \varepsilon', \varepsilon' = -\varepsilon \sim N(0,1), \tag{5}$$

Once parameter $\boldsymbol{\beta}$ is estimated, the monotone function $g(\cdot)$ can be estimated by conducting monotone regression on the working data $\{(\boldsymbol{x}_i^\top \hat{\boldsymbol{\beta}}, y_i)\}_{i=1}^n$. Univariate monotone regression has been extensively studied in the literature (Dette et al., 2006; Hall and Huang, 2001; Mammen, 1991; Ramsay, 1998). To fix idea, we adopt the monotone smoother proposed by Dette et al. (2006) whichi is constructed in a three-step procedure.

Denote $z_i = \boldsymbol{x}_i^\top \hat{\boldsymbol{\beta}}$, for $i = 1, \cdots, n$. It starts with an unconstrained estimate of the regression function, say the classical Nadaraya-Watson estimate $\hat{g}(\cdot)$:

$$\hat{g}(z) = \frac{\sum_{i=1}^n K_r((z_i - z)/h_r) y_i}{\sum_{i=1}^n K_r((z_i - z)/h_r)}. \tag{6}$$

In a second step, a density estimate of the observations $(\hat{g}^{-1})'(U_i)$ is calculated, which is integrated to obtain an estimate of the inverse of the regression function:

$$\hat{g}_I^{-1}(t) := \frac{1}{N h_d} \int_{-\infty}^{t} \sum_{i=1}^{N} K_d\left(\frac{\hat{g}(i/N) - u}{h_d}\right) du, \tag{7}$$

where K_r and K_d denote symmetric kernels with compact support and finite second moment, and h_r, h_d are the corresponding bandwidths converging to 0 with increasing sample size n. The estimate $\hat{g}_I^{-1}$ is isotonic when the kernel K_d is positive. So an isotonic estimate of the regression function $\hat{g}_I$ is simply obtained by reflection of the function $\hat{g}_I^{-1}$ in the line $y = x$. The

asymptotic normality of this monotone smoother is also established in Dette et al. (2006). The R package `monreg` is used for the implementation of this method (`https://CRAN.R-project.org/package=monreg`).

3 Penalized Composite Probit Regression

Our major theoretical contribution is to show that the regression parameter in the non-parametric Box-Cox model can be optimally estimated by the penalized composite probit regression estimator defined in (4) in section 2.1. In this section, we give a full account of the penalized CPR. We first reformulate the probit regression problem as a large margin classifier. Then we develop an efficient algorithm for solving the penalized CPR estimator. Finally, we show the strong oracle property of the estimator.

3.1 Probit regression as a large margin classifier

In a binary classification problem, we are given n pairs of training data $\{(\boldsymbol{x}_i, \check{y}_i)\}_{i=1}^n$ where $\boldsymbol{x}_i \in \mathbb{R}^p$ are predictors with the first being identity predictor, and $\check{y}_i \in \{-1, 1\}$ denotes class labels. A large margin classifier uses a margin-based loss function $L(Y, f(\boldsymbol{x})) = L(Y \cdot f(\boldsymbol{x}))$. By using $\{-1, 1\}$ to code the class label as opposed to $\{0, 1\}$ we can reformulate the logistic regression model as a large margin classifier with the loss function being $L(t) = \log(1 + e^{-t})$. We show that the same can be done for the probit regression which is typically introduced as a maximum likelihood estimator.

Note that the probit regression model assumes that $\mathbb{P}(\check{Y}_i = 1|\boldsymbol{x}_i) = \Phi(\boldsymbol{x}_i^\top \boldsymbol{\beta})$ and $\mathbb{P}(\check{Y}_i = -1|\boldsymbol{x}_i) = \Phi(-(\boldsymbol{x}_i^\top \boldsymbol{\beta}))$, where $\Phi(\cdot)$ is the CDF of standard norm distribution. Equivalently, we can use a unified formula

$$\mathbb{P}(\check{Y}_i = \check{y}_i|\boldsymbol{x}_i) = \Phi(\check{y}_i(\boldsymbol{x}_i^\top \boldsymbol{\beta})). \tag{8}$$

The negative log-likelihood function (scaled by n) of the probit regression model then becomes

$$\ell_n(\boldsymbol{\beta}) := \frac{1}{n} \sum_{i=1}^n L(t_i), \tag{9}$$

where $L(t) = -\log(\Phi(t))$ is the loss function induced by the probit model, named as *probit regression loss function*, and $t_i = \check{y}_i(\boldsymbol{x}_i^\top \boldsymbol{\beta})$ is the margin of the i-th pair of data.

The graph of probit regression loss function $L(t) = -\log(\Phi(t))$ is illustrated in the left panel of Figure 1. Its second derivative is shown in the right panel. It can be seen that the probit loss

function is strongly convex and its second derivative is bounded below 1. These two properties of the probit loss function are formally established in the following lemma.

Lemma 1. *The probit regression loss function* $L(t) = -\log(\Phi(t))$ *has the following property:*

$$L''(t) = \frac{t\varphi(t)}{\Phi(t)} + \left(\frac{\varphi(t)}{\Phi(t)}\right)^2 \in (0, 1).$$

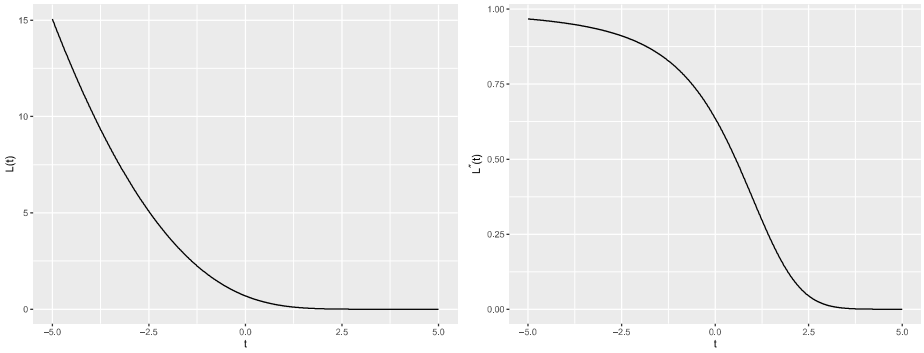

Figure 1: *First panel: Plot of probit regression loss function* $L(t) = -\log(\Phi(t))$. *Second panel: Plot of second derivative of probit loss funcion* $L''(t)$.

By using the probit loss function, we can rewrite the penalized composite probit regression estimator $\hat{\boldsymbol{\beta}}^{\text{pcpr}}$ as the solution to the following equivalent optimization problem:

$$\min_{\boldsymbol{\beta}, \boldsymbol{b}_0} \underbrace{\sum_{k=1}^{K} w_k \left\{ \frac{1}{n} \sum_{i=1}^{n} L(\check{y}_{ki}(\boldsymbol{x}_i^\top \boldsymbol{\beta} - b_{k0})) \right\}}_{M_n(\boldsymbol{\beta}, \boldsymbol{b}_0)} + \sum_{j=1}^{p} p_\lambda(|\beta_j|) =: F(\boldsymbol{\beta}, \boldsymbol{b}_0) \tag{10}$$

where $\check{y}_{ki} = -1$ if $\tilde{y}_{ki} = 0$ and $\check{y}_{ki} = 1$ if $\tilde{y}_{ki} = 1$. The reformulation of the objective function makes it easier for us to derive the following computing algorithm.

3.2 Computing algorithm: LLA-CMD

To solve the folded concave penalized composite probit regression problem (10), we consider combining the local linear approximation (LLA) algorithm (Fan et al., 2014; Zou and Li, 2008) and the coordinate-majorization-descent (CMD) algorithm (Yang and Zou, 2013). Details of the computing algorithm are discussed below.

Outer loop: local linear approximation The local linear approximation (LLA) algorithm (Zou and Li, 2008) takes advantage of the special folded concave structure and utilizes the

majorization-minimization (MM) principle to turn a concave regularization problem into a sequence of weighted ℓ_1 penalized problems. Let $(\tilde{\boldsymbol{\beta}}, \tilde{\boldsymbol{b}}_0)$ be the current estimate. The folded concave penalty could be majorized by a local linear approximation function:

$$\sum_j p_\lambda(|\beta_j|) \le \sum_j p_\lambda(|\tilde{\beta}_j|) + p'_\lambda(|\tilde{\beta}_j|)(|\beta_j| - |\tilde{\beta}_j|), \tag{11}$$

which is the best convex majorization of the concave penalty function (Theorem 2 of Zou and Li 2008). Then the objective function of (10) could be majorized by a weighted ℓ_1 penalized problem:

$$M_n(\boldsymbol{\beta}, \boldsymbol{b}_0) + \sum_j p_\lambda(|\tilde{\beta}_j|) + p'_\lambda(|\tilde{\beta}_j|)(|\beta_j| - |\tilde{\beta}_j|). \tag{12}$$

The details of the LLA algorithm are summarized in Algorithm 1.

Inner loop: coordinate-majorization-descent For our weighted ℓ_1 penalized composite probit regression problem (19) within each LLA iteration, we may also apply the coordinate descent algorithm (Friedman et al., 2010) which has been successfully used in solving some high-dimensional models. In the case of probit regression, we need to pay attention to computer overflow errors that may occur during the computation of the CDF $\Phi(\cdot)$. Standard algorithms like Newton-Raphson are very sensitive to large values of the linear predictor (Demidenko, 2001).

We prefer to use a numerically stable and efficient algorithm to solve (19). Due to the good property of probit regression loss given in Lemma 1, that is, the second derivative of the probit regression loss function can be bounded by 1, we can fix the computer overflow error issue by using the coordinate-majorization-descent algorithm (Yang and Zou, 2013) which only uses the gradient information of the composite probit loss function.

Let $(\tilde{\boldsymbol{\beta}}, \tilde{\boldsymbol{b}}_0)$ be the current estimate. Define the current margin $r_{ki} = \check{y}_{ki}(\boldsymbol{x}_i^\top \tilde{\boldsymbol{\beta}} - \tilde{b}_{k0})$ for $k = 1, \cdots, K$, $i = 1, \cdots, n$ and current ℓ_1 weights $\tilde{\omega}_j = p'_\lambda(|\tilde{\beta}_j|)$ for $j = 1, \cdots, p$. To update the s-th coordinate of $\boldsymbol{\beta}$, define the F function:

$$F(\beta_s | \tilde{\boldsymbol{\beta}}, \tilde{\boldsymbol{b}}_0) = \sum_{k=1}^K w_k \left\{ \frac{1}{n} \sum_{i=1}^n L(\check{y}_{ki} x_{is}(\beta_s - \tilde{\beta}_s) + r_{ki}) \right\} + \tilde{\omega}_s |\beta_s|. \tag{13}$$

By Lemme 1 and $\check{y}_{ki}^2 = 1$, this F function can be majorized by a penalized quadratic function defined as

$$Q(\beta_s | \tilde{\beta}, \tilde{\boldsymbol{b}}_0) =: \sum_{k=1}^K w_k \cdot \frac{1}{n} \sum_{i=1}^n \left[L(r_{ki}) + L'(r_{ki}) \check{y}_{ki} x_{is}(\beta_s - \tilde{\beta}_s) + \frac{1}{2} x_{is}^2 (\beta_s - \tilde{\beta}_s)^2 \right] + \tilde{\omega}_s |\beta_s|. \tag{14}$$

We can easily solve the minimizer of the majorization function by a simple soft thresholding rule:

$$\tilde{\beta}_s^{\text{new}} = \mathcal{S}\left(\tilde{\beta}_s + \frac{-\sum_{k=1}^K w_k \sum_{i=1}^n L'(r_{ki})\check{y}_{ki}x_{is}}{\sum_{i=1}^n x_{is}^2}, \frac{\tilde{\omega}_s}{\frac{1}{n}\sum_{i=1}^n x_{is}^2}\right), \quad \text{if } \tilde{\omega}_s \neq 0; \tag{15}$$

where $\mathcal{S}(z,t) = (|z| - t)_+ \text{sgn}(z)$ and

$$\tilde{\beta}_s^{\text{new}} = \tilde{\beta}_s + \frac{-\sum_{k=1}^K w_k \sum_{i=1}^n L'(r_{ki})\check{y}_{ki}x_{is}}{\sum_{i=1}^n x_{is}^2}, \quad \text{if } \tilde{\omega}_s = 0. \tag{16}$$

We then set $\tilde{\beta}_s = \tilde{\beta}_s^{\text{new}}$ as the new estimate.

We use the same trick to update the k-th intercept b_{k0}. Similar to (14), we consider minimizing the quadratic majorization:

$$Q(b_{k0}|\tilde{\boldsymbol{\beta}}, \tilde{\boldsymbol{b}}_0) =: \frac{1}{n}\sum_{i=1}^n \left\{L(r_{ki}) + L'(r_{ki})(-\check{y}_{ki})(b_{k0} - \tilde{b}_{k0}) + \frac{1}{2}(b_{k0} - \tilde{b}_{k0})^2\right\}, \tag{17}$$

which has a minimizer

$$\tilde{b}_{k0}^{\text{new}} = \tilde{b}_{k0} + \frac{1}{n}\sum_{i=1}^n L'(r_{ki})\check{y}_{ki}. \tag{18}$$

To sum up, the CMD algorithm for solving the weighted ℓ_1-penalized CPR is given in Algorithm 2. We have implemented both Algorithms 1 and 2 in an R package `copor` which is available from the authors upon request.

Algorithm 1: The LLA Algorithm

1: Initialize $\hat{\boldsymbol{\beta}}^{(0)} = \hat{\boldsymbol{\beta}}^{\text{initial}}$ and compute the adaptive weight

$$\hat{\boldsymbol{\omega}}^{(0)} = \left(\hat{\omega}_1^{(0)}, \cdots, \hat{\omega}_p^{(0)}\right)^\intercal = \left(p'_\lambda(|\hat{\beta}_1^{(0)}|), \cdots, p'_\lambda(|\hat{\beta}_p^{(0)}|)\right)^\intercal$$

2: For $m = 1, 2, \cdots$, repeat the LLA iteration till convergence

(2.a) Obtain $(\hat{\boldsymbol{\beta}}^{(m)}, \hat{\boldsymbol{b}}_0^{(m)})$ by solving the following optimization problem

$$(\hat{\boldsymbol{\beta}}^{(m)}, \hat{\boldsymbol{b}}_0^{(m)}) = \arg\min_{\boldsymbol{\beta}, \boldsymbol{b}_0} M_n(\boldsymbol{\beta}, \boldsymbol{b}_0) + \sum_i \hat{\omega}_j^{(m-1)} \cdot |\beta_j|, \tag{19}$$

(2.b) Update the adaptive weight vector $\hat{\boldsymbol{\omega}}^{(m)}$ with $\hat{\omega}_j^{(m)} = p'_\lambda(|\hat{\beta}_j^{(m)}|)$.

3.3 Strong oracle property

Throughout this paper, we follow the definition given by Fan et al. (2014) and assume that the penalty $p_\lambda(|t|)$ is a general folded concave penalty function defined on $t \in (-\infty, \infty)$ satisfying

Algorithm 2: The CMD algorithm for solving (19).

1: Input the weight vector $\tilde{\boldsymbol{\omega}}$.

2: Initialize $(\tilde{\boldsymbol{\beta}}, \tilde{\boldsymbol{b}}_0)$.

3: Iterate 3(a)-3(b) until convergence:

(3(a) Cyclic coordinate descent for coefficients: for $s = 1, 2, \cdots, p$,

(3.a.1) Compute the current margin $r_{ki} = \check{y}_{ki}(\boldsymbol{x}_i^\top \tilde{\boldsymbol{\beta}} - \tilde{b}_{k0})$.

(3.a.2) Compute

$$\tilde{\beta}_s^{\text{new}} = \begin{cases} \mathcal{S}\left(\tilde{\beta}_s + \frac{-\sum_{k=1}^K w_k \sum_{i=1}^n L'(r_{ki})\check{y}_{ki}x_{is}}{\sum_{i=1}^n x_{is}^2}, \frac{\tilde{\omega}_s}{\frac{1}{n}\sum_{i=1}^n x_{is}^2}\right), & \text{if } \tilde{\omega}_s \neq 0; \\ \tilde{\beta}_s^{\text{new}} = \tilde{\beta}_s + \frac{-\sum_{k=1}^K w_k \sum_{i=1}^n L'(r_{ki})\check{y}_{ki}x_{is}}{\sum_{i=1}^n x_{is}^2}, & \text{if } \tilde{\omega}_s = 0. \end{cases}$$

(3.a.3) Set $\tilde{\beta}_s = \tilde{\beta}_s^{\text{new}}$.

(3(b) Cyclic coordinate descent for intercepts: for $k = 1, 2, \cdots, K$,

(3.b.1) Recompute the current margin $r_{ki} = \check{y}_{ki}(\boldsymbol{x}_i^\top \tilde{\boldsymbol{\beta}} - \tilde{b}_{k0})$.

(3.a.2) Compute

$$\tilde{b}_{k0}^{\text{new}} = \tilde{b}_{k0} + \frac{1}{n}\sum_{i=1}^n L'(r_{ki})\check{y}_{ki}.$$

(3.a.3) Set $\tilde{b}_{k0} = \tilde{b}_{k0}^{\text{new}}$.

(i) $p_\lambda(|t|)$ is increasing and concave in $t \in [0, \infty)$ with $p_\lambda(0) = 0$;

(ii) $p_\lambda(|t|)$ is differentiable in $t \in (0, \infty)$ with $p'_\lambda(0) := p'_\lambda(0_+) \geq a_1\lambda$;

(iii) $p'_\lambda(t) \geq a_1\lambda$ for $t \in (0, a_2\lambda]$;

(iv) $p'_\lambda(t) = 0$ for $t \in [a\lambda, \infty)$ with the pre-specified constant $a > a_2$;

where a_1 and a_2 are two fixed positive constants. The above definition extends previous works on the SCAD and the MCP (Fan and Li, 2001; Fan and Lv, 2011; Zhang, 2010). The derivative of the SCAD penalty is defined as

$$p'_\lambda(t) = \lambda I_{\{t \leq \lambda\}} + \frac{(a\lambda - t)_+}{a - 1} I_{\{t > \lambda\}}, \quad \text{for some } a > 2, \tag{20}$$

with $a_1 = a_2 = 1$, and the derivative of the MCP is $p'_\lambda(t) = (\lambda - t/a)_+$, for some $a > 1$, with $a_1 = 1 - a^{-1}$ and $a_2 = 1$.

For ease of analysis of strong oracle property, we now consider the composite probit regression model based on observations $\{(\boldsymbol{x}_i, \tilde{y}_{1i}, \cdots, \tilde{y}_{Ki})\}_{i=1}^n$, where $\tilde{y}_{ki} \in \{0, 1\}$, and define

$$\mathcal{B} = \begin{pmatrix} \boldsymbol{b}_0 \\ \boldsymbol{\beta} \end{pmatrix} \in \mathbb{R}^{K+p}, \quad \boldsymbol{x}_i^k = \begin{pmatrix} -e_k \\ x_i \end{pmatrix}, \quad \boldsymbol{X}^k = \begin{pmatrix} (\boldsymbol{x}_1^k)^\intercal \\ \vdots \\ (\boldsymbol{x}_n^k)^\intercal \end{pmatrix}, \quad \tilde{\boldsymbol{y}}^k = \begin{pmatrix} \tilde{y}_{1k} \\ \vdots \\ \tilde{y}_{nk} \end{pmatrix},$$

where $e_k \in \mathbb{R}^K$ with the k-th element 1 and all others 0. Then the negative composite probit regression likelihood function can be written as

$$\tilde{M}_n(\mathcal{B}) = \sum_{k=1}^K w_k \left\{ \frac{1}{n} \sum_{i=1}^n \left[-\tilde{y}_{ik} h((\boldsymbol{x}_i^k)^\intercal \mathcal{B}) - \log(1 - \Phi((\boldsymbol{x}_i^k)^\intercal \mathcal{B})) \right] \right\}, \tag{21}$$

where $h(\eta) = \log\left(\frac{\Phi(\eta)}{1-\Phi(\eta)}\right)$.

In high-dimensional data analysis, the dimension p of the parameter $\boldsymbol{\beta}$ is assumed to be larger than the sample size n. Its support set is defined as $\mathcal{A} = \{j : \beta_j \neq 0\}$ with cardinality s assumed to be much smaller than p: $s \ll p$. The *oracle estimator* is then defined as the minimizer that knows in advance the true support set:

$$\hat{\mathcal{B}}^{\text{oracle}} = (\hat{\mathcal{B}}_{\mathcal{A}}^{\text{oracle}}, \boldsymbol{0}) = \arg \min_{\mathcal{B}: \mathcal{B}_{\mathcal{A}^c} = 0} \tilde{M}_n(\mathcal{B}). \tag{22}$$

The oracle estimator is unique due to the strong convexity of composite probit regression. The oracle estimator is not a genuine estimator, but it can be used as a theoretic benchmark for other estimators. An estimator is said to have the *strong oracle property* if the estimator equals the oracle estimator with overwhelming probability (Fan and Lv, 2011). Theorem 1 and 2 in Fan et al. (2014) state that as long as the problem is localizable and regular, we can find the oracle estimator by using the one-step local linear approximation and the LLA algorithm actually converges after two iterations. These conditions can be established for the composite probit regression model.

Define $H^k(\mathcal{B}) = \text{diag}\left\{\frac{\varphi(\eta_i^k)}{\Phi(\eta_i^k)(1-\Phi(\eta_i^k))}, i = 1, \cdots, n\right\}$, $p^k(\mathcal{B}) = \left(\Phi(\eta_i^k), \cdots, \Phi(\eta_i^k)\right)^\intercal$, and $\Sigma^k(\mathcal{B}) = \text{diag}\left\{\tilde{y}_{ik} L''(\eta_i^k) + (1 - \tilde{y}_{ik}) L''(-\eta_i^k), i = 1, \cdots, n\right\}$, where $\eta_i^k = (\boldsymbol{x}_i^k)^\intercal \mathcal{B}$. We also define four useful quantities: $Q_1 = \max_k \max_{j \in \mathcal{A}} \lambda_{\max}\left(\frac{1}{n}(\boldsymbol{X}_{\mathcal{A}}^k)^\intercal \text{diag}\{|\boldsymbol{x}_{(j)}^k|\} \boldsymbol{X}_{\mathcal{A}}^k\right)$, where $\text{diag}\{|\boldsymbol{x}_{(j)}^k|\}$ is a diagonal matrix with elements $\{|x_{ij}^k|\}_{i=1}^n$,

$$Q_2 = \left\| \left(\sum_{k=1}^K w_k \frac{1}{n} (\boldsymbol{X}_{\mathcal{A}}^k)^\intercal \Sigma^k(\mathcal{B}^*) \boldsymbol{X}_{\mathcal{A}}^k \right)^{-1} \right\|_{\ell_\infty},$$

where $\mathcal{B}^*$ is the true parameter,

$$Q_3 = \left\| \left[\sum_{k=1}^{K} w_k (\boldsymbol{X}^k_{\mathcal{A}^c})^\intercal \Sigma^k(\mathcal{B}^*) \boldsymbol{X}^k_{\mathcal{A}} \right] \cdot \left[\sum_{k=1}^{K} w_k (\boldsymbol{X}^k_{\mathcal{A}})^\intercal \Sigma^k(\mathcal{B}^*) \boldsymbol{X}^k_{\mathcal{A}} \right]^{-1} \right\|_{\ell_\infty},$$

and $Q_4 = \max_{i,k} \left\{ \frac{\varphi((\boldsymbol{x}_i^k)^\intercal \mathcal{B}^*)}{\Phi((\boldsymbol{x}_i^k)^\intercal \mathcal{B}^*)(1-\Phi((\boldsymbol{x}_i^k)^\intercal \mathcal{B}^*))} \right\}$, where $\mathcal{B}^*$ is the true parameter.

Theorem 1. *Suppose the minimal signal strength satisfies* (A_0) : $\|\mathcal{B}^*_{\mathcal{A}}\|_{\min} > (a+1)\lambda$. *Given a folded concavs penalty* $p_\lambda(\cdot)$ *satisfying (i)-(iv), let* $a_0 = \min\{1, a_2\}$. *Under the event*

$$\mathcal{E} = \{\|\hat{\mathcal{B}}^{initial} - \mathcal{B}^*\|_{\max} \le a_0\lambda\} \cap \{\|\nabla_{\mathcal{A}^c} \tilde{M}_n(\hat{\mathcal{B}}^{oracle})\|_{\max} < a_1\lambda\} \cap \{\|\hat{\mathcal{B}}^{oracle}_{\mathcal{A}}\|_{\min} > a\lambda\},$$

the LLA algorithm initialized by $\hat{\mathcal{B}}^{initial}$ *converges to* $\hat{\mathcal{B}}^{oracle}$ *after two iterations. Therefore, under the composite probit regression model assumption, with probability at least* $1-\delta_0-\delta_1^{cpr}-\delta_2^{cpr}$, *the LLA algorithm initialized by* $\hat{\mathcal{B}}^{initial}$ *converges to* $\hat{\mathcal{B}}^{oracle}$ *after two iterations, where* $\delta_0 = \mathbb{P}(\|\hat{\mathcal{B}}^{initial} - \mathcal{B}^*\|_{\max} > a_0\lambda)$,

$$\delta_1^{cpr} = 2s \exp\left(-\frac{2n}{Q_4^2 M} \min\left\{ \frac{a_1^2\lambda^2}{4(2Q_3+1)^2}, \frac{1}{16\tau^2 Q_1^2 Q_2^4 s^2} \right\} \right) + 2(K+p-s)\exp\left(-\frac{2n}{Q_4^2 M} \frac{a_1^2\lambda^2}{4} \right),$$

with $M = \max_{j\in\mathcal{A}} \left\{ \frac{1}{n} \sum_{i=1}^n \left(\sum_{k=1}^K w_k |x_{ij}^k| \right)^2 \right\}$, $\tau = \max_{\eta\in\mathbb{R}} |L'''(\eta)| (\approx 0.3)$ *and*

$$\delta_2^{cpr} = 2s \exp\left(-\frac{2n}{Q_4^2 M} \min\left\{ \frac{1}{16\tau^2 Q_1^2 Q_2^4 s^2}, \frac{1}{4Q_2^2} \left(\|\mathcal{B}^*_{\mathcal{A}}\|_{\min} - a\lambda \right)^2 \right\} \right).$$

Under fairly weak assumptions, δ_1^{cpr} and δ_2^{cpr} go to zero very quickly. The remaining challenge is to bound $\delta_0 = \mathbb{P}\left(\|\hat{\mathcal{B}}^{\text{initial}} - \mathcal{B}^*\|_{\max} > a_0\lambda \right)$. We consider using the ℓ_1-penalized M-estimator as the initial estimator, that is,

$$\hat{\mathcal{B}}^{\text{lasso}} = \arg\min_{\mathcal{B}} \tilde{M}_n(\mathcal{B}) + \lambda_{\text{lasso}} \|\boldsymbol{\beta}\|_{\ell_1}. \tag{23}$$

Theorem 2. *Let* $m = \max_k \max_{(i,j)} |x_{i,j}^k|$ *and define the simple general invertibility factor (GIF):*

$$(C): \qquad \kappa_{cpr} = \min_{u\neq 0:\, \|u_{\mathcal{A}^c}(\boldsymbol{\beta})\|_{\ell_1} \le 3\|u_{\mathcal{A}}(\boldsymbol{\beta})\|_{\ell_1} + \|u(\boldsymbol{b}_0)\|_{\ell_1}} \frac{u^\intercal \nabla^2 \tilde{M}_n(\mathcal{B}^*) u}{\|u\|_{\ell_1}\|u\|_{\ell_\infty}} \in (0,\infty), \tag{24}$$

if the Lasso parameter λ_{lasso} *satifies*

$$\lambda_{lasso} \le \frac{\sqrt{(\frac{1}{2}\|\mathcal{B}^*\|_{\ell_1})^2 + m^{-2}} - \frac{1}{2}\|\mathcal{B}^*\|_{\ell_1}}{5\kappa_{cpr}}, \tag{25}$$

then with probability at least $1 - 2p\exp\left(-\frac{n}{2Q_4^2 M} \lambda_{lasso}^2 \right)$, *we have*

$$\|\hat{\mathcal{B}}^{lasso} - \mathcal{B}^*\|_{\max} \le 5\kappa_{cpr}^{-1} \lambda_{lasso}. \tag{26}$$

The condition (C) on the simple GIF is the generalization of the restricted eiganvalue condition (Bickel et al., 2009). In light of Theorem 2, we can obtain the following corollary that bounds δ_0, δ_1 and δ_2 simultaneously.

Corollary 1. *Under assumptions* (A_0) : $\|\mathcal{B}^*_{\mathcal{A}}\|_{min} > (a+1)\lambda$ *and* (C), *if* $\lambda \geq \frac{5\lambda_{lasso}}{a_0\kappa_{cpr}}$, *the LLA algorithm initialized by* $\hat{\mathcal{B}}^{lasso}$ *converges to* $\hat{\mathcal{B}}^{oracle}$ *after two iterations with probability at least* $1 - 2p\exp\left(-\frac{n}{2Q_4^2 M}\lambda^2_{lasso}\right) - \delta_1^{cpr} - \delta_2^{cpr}$.

Corollary 1 suggests that sometimes it is good to use zero to initialize the LLA algorithm. If $\hat{\mathcal{B}}^{\text{initial}} = \mathbf{0}$, the first LLA iteration gives a ℓ_1-penalized composite probit regression estimator with $\lambda_{\text{lasso}} = p'_\lambda(0)$. For both SCAD and MCP, $p'_\lambda(0) = \lambda$. If $\lambda_{\text{lasso}} = \lambda$ satisfies requirements in Corollary 1, then after two more LLA iterations, the LLA algorithm converges to the oracle estimator with the high probability. To be more specific, we have the following corollary.

Corollary 2. *Consider the SCAD/MAP penalized probit regression. Under assumption* (A_0) *and* (C), *if* $a_0\kappa_{cpr} \geq 5$ *holds, the LLA algorithm initialized by zero converges to the oracle estimator after three iterations with probability at least* $1 - 2p\exp\left(-\frac{n}{2Q_4^2 M}\lambda^2\right) - \delta_1^{cpr} - \delta_2^{cpr}$.

The folded concave penalized CPR is rate optimal adaptive to the unknown transformation function. This can be seen based on the following argument. If we knew the true g function, then we could define a double oracle estimator by considering the least squares on $(g(y_i), \boldsymbol{x}_i)_{i=1}^n$. This double oracle estimator is more efficient than the oracle estimator in (22), but the two have the same rate of convergence. By the strong oracle property, the folded concave penalized CPR is rate optimal.

4 Numerical Studies

In this section we use simulations to examine the performance of the proposed methodology under various settings. To fix idea, we use the equal weights in composite probit regression.

The settings of the simulation studies are given below. We consider the monotone transformation function: $g_1^{-1}(t) = \frac{1}{2}(2t-1)^{1/3} + \frac{1}{2}$ and $g_2^{-1}(t) = \frac{\exp((t-2)/3)}{1+0.5((t-2)/3)^2}$. The true regression parameter $\boldsymbol{\beta}^*$ is set to be $\boldsymbol{\beta}^*_1 = (3, 1.5, 0, 0, 2, \mathbf{0}_{p-5})/2$ with support set $\mathcal{A}_1 = \{1, 2, 5\}$, and $\boldsymbol{\beta}^*_2 = (0.85, 0.85, 0.85, 0.85, 0.85, \mathbf{0}_{p-5})/\sqrt{2}$ with support set $\mathcal{A}_2 = \{1, 2, 3, 4, 5\}$. The covariates are generated from $\boldsymbol{x} \sim N_p(\mathbf{0}, \Sigma)$ with covariance matrix having the first-order autoregressive structure $\Sigma = (\rho^{|i-j|})_{p\times p}$ or the compound symmetry structure $\Sigma = (\rho^{I_{\{i\neq j\}}})_{p\times p}$ with parameter

$\rho = 0.5$ or 0.8. The number of observations and predictors (n, p) are set to be $(200, 1000)$ and $(400, 2000)$. For each combination of the above options, 100 replications are done.

We include the usual penalized least squares without regarding to the transformation function (abbreviated as LS), an oracle-assisted estimator that fits a penalized least squares on $\{(\boldsymbol{x}_i, g(y_i)\}_{i=1}^n$ (abbreviated as OA):

$$\hat{\boldsymbol{\beta}}^{\mathrm{oa}} = \arg\min_{\boldsymbol{\beta}} \frac{1}{2}\sum_{i=1}^{n}(g(y_i) - \boldsymbol{x}_i^\top \boldsymbol{\beta})^2 + \sum_{j=1}^{p} p_\lambda(|\beta_j|), \tag{27}$$

and an double-oracle-assisted estimator that fits a least squares on $\{(\boldsymbol{x}_i)_{\mathcal{A}}, g(y_i)\}_{i=1}^n$ (abbreviated as DOA):

$$\hat{\boldsymbol{\beta}}^{\mathrm{doa}} = \arg\min_{\boldsymbol{\beta},\, \boldsymbol{\beta}_{\mathcal{A}^c}=0} \frac{1}{2}\sum_{i=1}^{n}(g(y_i) - \boldsymbol{x}_i^\top \boldsymbol{\beta})^2. \tag{28}$$

For the composite probit regression, we use the 25%, 50% and 75% empirical percentiles of the responses to dichotomize the response variable and combine the three probit regression models by using the equal weights (abbreviated as NBC). We use the SCAD penalty with $a = 3.7$ as the folded concave penalization for sparse estimation of $\boldsymbol{\beta}$. The penalization parameter λ is tuned from 5-fold cross-validation. When applying the LLA-CMD algorithm for obtaining $\hat{\boldsymbol{\beta}}^{\mathrm{pcpr}}$, we compute the three-step LLA solution initialized by $\mathbf{0}$.

The bandwidths h_r and h_d for estimating the monotone function $g(\cdot)$ in the second step are chosen as $h_r = (\hat{\sigma}^2/n)^{1/5}$, $h_d = h_r^3$, where

$$\hat{\sigma}^2 = \frac{1}{2(n-1)}\sum_{i=1}^{n-1}\left(z_{[i+1]} - z_{[i]}\right)^2,$$

here $z_{[1]}, \cdots, z_{[n]}$ denote the observations ordered with respect to their corresponding y-values, as suggested by Dette et al. (2006).

The optimal prediction function for Y under the squared error loss is $\mathbb{E}[Y|\boldsymbol{x}]$, which equals $\mathbb{E}_{\varepsilon\sim N(0,1)}[g^{-1}(\boldsymbol{x}^\top\boldsymbol{\beta}^* + \varepsilon)]$ under the nonparametric Box-Cox model. Our prediction of Y is obtained by Monte Carlo method:

$$\hat{\mathbb{E}}[Y|\boldsymbol{x}] := \frac{1}{N_{\mathrm{mc}}}\sum_{s=1}^{N_{\mathrm{mc}}} \hat{g}^{-1}(\boldsymbol{x}^\top\hat{\boldsymbol{\beta}} + \varepsilon_s), \tag{29}$$

where $\{\varepsilon_s\}_{s=1}^{N_{\mathrm{mc}}}$ is a Monte Carlo sample from $N(0,1)$, N_{mc} is set to be sufficiently large. The M.S.E between y_i's and $\hat{\mathbb{E}}[y_i|\boldsymbol{x}_i]$'s on an independent validation set is used to measure the prediction performance of our model. For the oracle-assisted and double-oracle-assisted estimator, the prediction is given by $\mathbb{E}_{\tilde{\varepsilon}\sim N(0,1)}[g^{-1}(\boldsymbol{x}^\top\hat{\boldsymbol{\beta}}+\tilde{\varepsilon})]$ where $\tilde{\varepsilon}$ is independent of the training set, thus

independent of $\hat{\boldsymbol{\beta}}$. The prediction errors of oracle-assisted and double-oracle-assisted estimators can be served as the benchmark for our method.

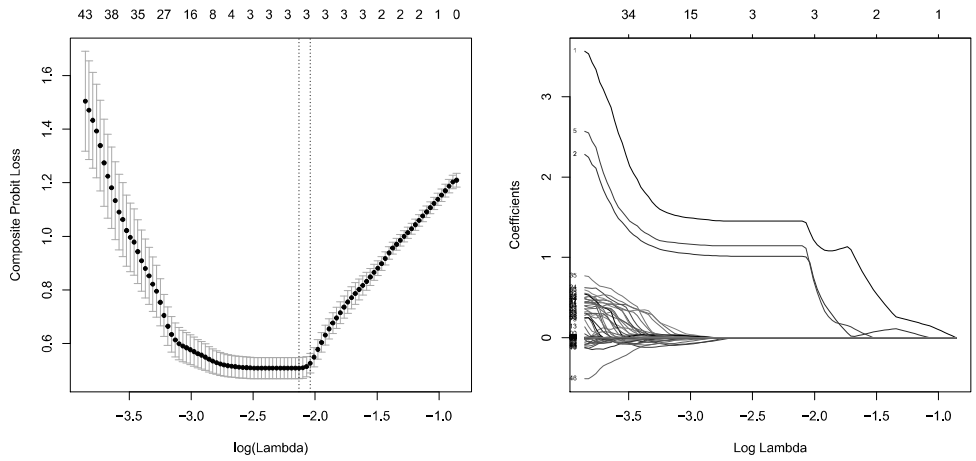

Figure 2: *Left panel: the cross-validation curve for chosing λ. Right panel: Solution path of each coefficient w.r.t. λ. Setting: $g_1(\cdot)$, $\boldsymbol{\beta}_1^*$, $\Sigma = (0.5^{|i-j|})_{p\times p}$, $(n, p) = (200, 1000)$.*

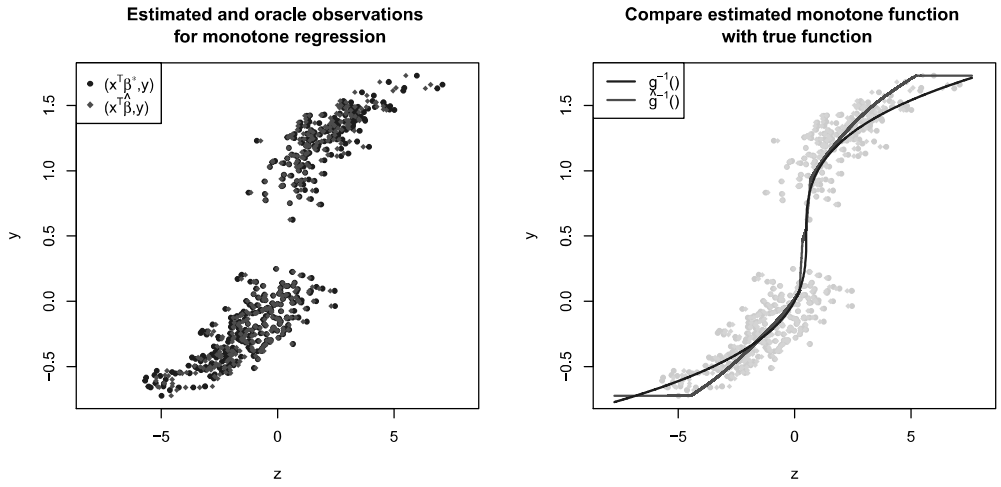

Figure 3: *Left panel: the oracle observations $\{(\boldsymbol{x}_i^\top \boldsymbol{\beta}^*, y_i)\}$ (blue points) and the estimated observations $\{(\boldsymbol{x}_i^\top \hat{\boldsymbol{\beta}}, y_i)\}$ (red points). Right panel: the estimated monotone function $\hat{g}^{-1}(\cdot)$ (red curve) is compared with the true function $g^{-1}(\cdot)$ (blue curve). Setting: $g_1(\cdot)$, $\boldsymbol{\beta}_1^*$, $\Sigma = (0.5^{|i-j|})_{p\times p}$, $(n, p) = (400, 2000)$.*

Table 1: *Comparing the performance of estimating the regression parameter vectors of penalized least squares (LS), nonparametric Box-Cox with composite probit (NBC), oracle-assisted estimator (OA, defined in (27)) & double-oracle-assisted estimator (DOA, defined in (28)) under different settings. Estimation accuracy is measured by the ℓ_2-loss. Each metric is averaged over 100 replications with its standard error shown in the parenthesis.*

	$(g_1, \beta_1^*, n = 200, p = 1000)$				$(g_2, \beta_1^*, n = 200, p = 1000)$			
Method	AR(0.5)	AR(0.8)	CS(0.5)	CS(0.8)	AR(0.5)	AR(0.8)	CS(0.5)	CS(0.8)
LS	1.410	1.450	1.470	1.550	1.580	1.580	1.600	1.630
	(0.023)	(0.029)	(0.037)	(0.089)	(0.021)	(0.030)	(0.029)	0.069
NBC	0.238	0.337	0.304	0.896	0.238	0.337	0.304	0.896
	(0.127)	(0.233)	(0.194)	(0.349)	(0.127)	(0.233)	(0.194)	(0.349)
OA	0.177	0.278	0.181	0.314	0.177	0.278	0.181	0.314
	(0.072)	(0.242)	(0.097)	(0.246)	(0.072)	(0.242)	(0.097)	(0.246)
DOA	0.129	0.181	0.136	0.203	0.129	0.181	0.136	0.203
	(0.053)	(0.096)	(0.065)	(0.099)	(0.053)	(0.096)	(0.065)	(0.099)
	$(g_1, \beta_1^*, n = 400, p = 2000)$				$(g_2, \beta_1^*, n = 400, p = 2000)$			
	AR(0.5)	AR(0.8)	CS(0.5)	CS(0.8)	AR(0.5)	AR(0.8)	CS(0.5)	CS(0.8)
LS	1.410	1.450	1.450	1.510	1.580	1.590	1.590	1.600
	(0.016)	(0.023)	(0.020)	(0.046)	(0.013)	(0.020)	(0.018)	(0.024)
NBC	0.166	0.222	0.190	0.488	0.166	0.222	0.190	0.488
	(0.075)	(0.116)	(0.100)	(0.288)	(0.075)	(0.116)	(0.100)	(0.288)
OA	0.109	0.141	0.113	0.175	0.109	0.141	0.113	0.175
	(0.053)	(0.069)	(0.058)	(0.112)	(0.053)	(0.069)	(0.058)	(0.112)
DOA	0.096	0.131	0.093	0.161	0.096	0.131	0.093	0.161
	(0.048)	(0.068)	(0.041)	(0.083)	(0.048)	(0.068)	(0.041)	(0.083)
	$(g_1, \beta_2^*, n = 200, p = 1000)$				$(g_2, \beta_2^*, n = 200, p = 1000)$			
	AR(0.5)	AR(0.8)	CS(0.5)	CS(0.8)	AR(0.5)	AR(0.8)	CS(0.5)	CS(0.8)
LS	1.000	1.070	1.090	1.250	1.100	1.130	1.140	1.260
	(0.046)	(0.037)	(0.081)	(0.074)	(0.021)	(0.024)	(0.051)	(0.064)
NBC	0.340	0.736	0.705	1.540	0.340	0.736	0.705	1.540
	(0.177)	(.297)	(0.277)	(0.291)	(0.177)	(0.297)	(0.277)	(.291)
OA	0.337	0.639	0.384	0.928	0.337	0.639	0.384	0.928
	(0.119)	(0.356)	(0.146)	(0.301)	(0.119)	(0.356)	(0.146)	(0.301)
DOA	0.191	0.285	0.204	0.305	0.191	0.285	0.204	0.305
	(0.070)	(0.120)	(0.065)	(0.095)	(0.070)	(0.120)	(0.065)	(0.095)
	$(g_1, \beta_2^*, n = 400, p = 2000)$				$(g_2, \beta_2^*, n = 400, p = 2000)$			
	AR(0.5)	AR(0.8)	CS(0.5)	CS(0.8)	AR(0.5)	AR(0.8)	CS(0.5)	CS(0.8)
LS	0.957	1.030	0.997	1.170	1.080	1.100	1.100	1.190
	(0.022)	(0.033)	(0.039)	(0.091)	(0.013)	(0.017)	(0.023)	(0.067)
NBC	0.193	0.348	0.257	0.990	0.193	0.348	0.257	0.990
	(0.078)	(0.180)	(0.137)	(0.270)	(0.078)	(0.180)	(0.137)	(0.270)
OA	0.163	0.335	0.171	0.371	0.163	0.335	0.171	0.371
	(0.063)	(0.234)	(0.061)	(0.238)	(0.063)	(0.234)	(0.061)	(0.238)
DOA	0.129	0.212	0.142	0.207	0.129	0.212	0.142	0.207
	(0.049)	(0.087)	(0.045)	(0.078)	(0.049)	(0.087)	(0.045)	(0.078)

Table 2: *Comparing the prediction performance of penalized least squares (LS), nonparametric Box-Cox with composite probit (NBC), oracle-assisted estimator (OA)& double-oracle-assisted estimator (DOA) under different settings. The prediction accuracy is measured by the M.S.E. between the response and the prediction on an independent validation set. Each metric is averaged M.S.E. over 100 replications with its standard error shown in the parenthesis.*

	$(g_1, \beta_1^*, n = 200, p = 1000)$				$(g_2, \beta_1^*, n = 200, p = 1000)$			
Method	AR(0.5)	AR(0.8)	CS(0.5)	CS(0.8)	AR(0.5)	AR(0.8)	CS(0.5)	CS(0.8)
LS	.168	.176	.173	.190	.070	.094	.080	.098
	(.010)	(.012)	(.015)	(.017)	(.005)	(.007)	(.007)	(.009)
NBC	.122	.107	.121	.128	.050	.058	.056	.066
	(.005)	(.007)	(.006)	(.012)	(.006)	(.007)	(.016)	(.013)
OA	.119	.105	.118	.112	.041	.048	.046	.047
	(.002)	(.007)	(.002)	(.005)	(.001)	(.002)	(.001)	(.002)
DOA	.118	.102	.117	.110	.040	.047	.045	.046
	(.002)	(.002)	(.001)	(.001)	(.000)	(.001)	(.001)	(.000)
	$(g_1, \beta_1^*, n = 400, p = 2000)$				$(g_2, \beta_1^*, n = 400, p = 2000)$			
	AR(0.5)	AR(0.8)	CS(0.5)	CS(0.8)	AR(0.5)	AR(0.8)	CS(0.5)	CS(0.8)
LS	.168	.171	.163	.167	.074	.088	.080	.094
	(.007)	(.008)	(.005)	(.008)	(.003)	(.004)	(.003)	(.005)
NBC	.126	.112	.116	.109	.056	.054	.056	.056
	(.003)	(.002)	(.002)	(.006)	(.003)	(.004)	(.004)	(.005)
OA	.122	.111	.114	.102	.045	.047	.048	.044
	(.001)	(.001)	(.001)	(.001)	(.001)	(.001)	(.000)	(.001)
DOA	.122	.110	.114	.102	.045	.047	.048	.044
	(.001)	(.001)	(.000)	(.001)	(.000)	(.000)	(.000)	(.001)
	$(g_1, \beta_2^*, n = 200, p = 1000)$				$(g_2, \beta_2^*, n = 200, p = 1000)$			
	AR(0.5)	AR(0.8)	CS(0.5)	CS(0.8)	AR(0.5)	AR(0.8)	CS(0.5)	CS(0.8)
LS	.215	.206	.202	.209	.075	.102	.083	.095
	(.017)	(.016)	(.016)	(.012)	(.007)	(.010)	(.010)	(.006)
NBC	.164	.137	.137	.165	.063	.063	.069	.079
	(.012)	(.008)	(.009)	(.021)	(.006)	(.006)	(.013)	(.014)
OA	.161	.135	.130	.136	.053	.056	.050	.055
	(.006)	(.007)	(.008)	(.010)	(.002)	(.004)	(.003)	(.005)
DOA	.154	.126	.124	.122	.050	.051	.047	.048
	(.002)	(.002)	(.002)	(.002)	(.001)	(.001)	(.001)	(.001)
	$(g_1, \beta_2^*, n = 400, p = 2000)$				$(g_2, \beta_2^*, n = 400, p = 2000)$			
	AR(0.5)	AR(0.8)	CS(0.5)	CS(0.8)	AR(0.5)	AR(0.8)	CS(0.5)	CS(0.8)
LS	.170	.183	.175	.177	.066	.081	.075	.091
	(.010)	(.010)	(.007)	(.008)	(.004)	(.005)	(.004)	(.006)
NBC	.128	.122	.135	.129	.054	.058	.057	.068
	(.003)	(.004)	(.005)	(.010)	(.006)	(.006)	(.004)	(.006)
OA	.125	.121	.131	.113	.045	.049	.047	.048
	(.002)	(.004)	(.001)	(.003)	(.001)	(.002)	(.001)	(.002)
DOA	.124	.118	.130	.111	.044	.048	.047	.047
	(.001)	(.001)	(.001)	(.001)	(.001)	(.001)	(.000)	(.000)

Figure 2 gives the illustration of estimating $\boldsymbol{\beta}$ using penalized composite probit regression under one setting within one replication. The left panel shows the cross-validation curve for choosing penalization parameter λ. The right panel shows the solution paths of each coefficient in terms of λ. Figure 3 gives the illustration of estimating the monotone function $g(\cdot)$ within the same replication. The left panel shows the estimated observations $\{(\boldsymbol{x}_i^\top \hat{\boldsymbol{\beta}}, y_i)\}$ for monotone regression and the true observations $\{(\boldsymbol{x}_i^\top \boldsymbol{\beta}^*, y_i)\}$. The right panel shows the estimated monotone function and the true function. The quantitative results are summarized in Table 1 and Table 2. In Table 1, we compare the accuracy of estimating the regression parameter of different methods by presenting the average ℓ_2-loss $\|\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}^*\|_{\ell_2}$ under different g functions and covariance structures. In Table 2, we compare the prediction performance of different methods by presenting the M.S.E. between response and prediction on an independent validation set. We can see that LS performs the worst as expected, because it has an intrinsic bias. Our estimator is slightly worse than the OA and the DOA and is much better LS. The prediction performance of our method could be really close to the prediction performance of the OA under many circumstances.

5 Analysis of the Supermarket Data

In our empirical study, we use the supermarket data in Lan et al. (2016). This data set contains a total of $n = 464$ daily records. For each record, the response is the number of customers and the predictors are the sales volumes of $p = 6398$ products. Both the response and predictors are standardized so that they have zero mean and unit variance. The purpose of this study is to identify the products that attract the most customers and to evaluate the impact of those products on the customers.

We apply the SCAD-penalized least squares (LS), which is the current standard practice for high-dimensional regression, and our non-parametric Box-Cox (NBC) method to this dataset. For a fair comparison, we randomly split the data set into two subsets: one for training (232 observations) and one for testing (232 observations). We used the Spearman rank correlation (Li et al., 2012) to do a variable screening to reduce the dimension to 1000, that is, we picked the top 1000 predictors with the largest Spearman rank correlation with the response. There are many other variable screening methods in the literature. We chose to use the rank correlation screening because it is coherent with the nonparametric Box-Cox model. Five-fold cross-validation was use to select the penalization parameter. It is worth pointing out that we did the variable screening

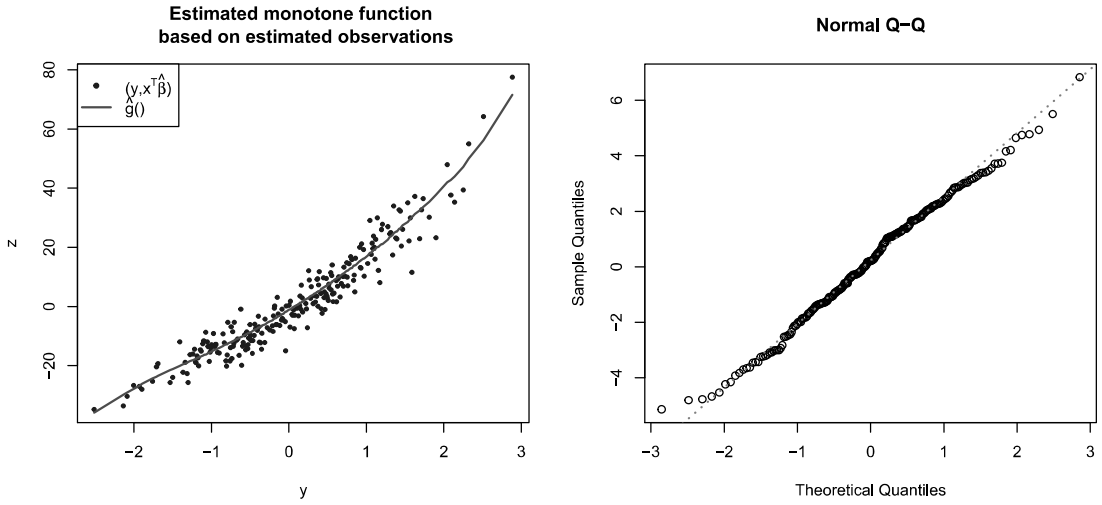

Figure 4: *Supermarket data. Left panel: The estimated observations* $\{(y_i, \boldsymbol{x}_i^\top \hat{\boldsymbol{\beta}})\}$ *(blue points) and the estimated monotone function* $\hat{g}(\cdot)$ *(red curve). Right panel: normal Q-Q plot of the residuals* $\{\hat{g}(y_i) - \boldsymbol{x}_i^\top \hat{\boldsymbol{\beta}}\}$.

within cross-validation to avoid any bias, that is, we used four of the five folds to screen variables and then fit the model before computing the prediction M.S.E. on the fifth fold. This process is repeated for each of the five folds. After cross-validation, we again used Spearman rank correlation screening to pick the top 1000 covariates on the whole training set and fit the sparse regression models using the chosen penalization parameters. Then, we computed the prediction M.S.E. between the response and the predicted value on the test set The whole procedure was repeated 20 times.

The prediction errors of LS and NBC have mean 0.1227 with standard error 0.0125 and mean 0.1029 with standard error 0.0117, respectively. Compared with the least-squares based regression model, the nonparamatrix Box-Cox model reduces the prediction error by about 16%. Left panel of Figure 4 shows the estimated monotone function in our NBC method under one replication in which the nonlinear trend is very clear. Right panel of Figure 4 shows the normal Q-Q plot by which we can see that the normal assumption in our nonparametric Box-Cox model is sufficient.

6 Discussion

In this article, we have studied the use of a nonparametric Box-Cox regression model as an alternative to the typically linear regression model for high-dimensional regression. The nonparametric Box-Cox regression model is less prone to model mis-specification than the typical linear regression model but also presents greater challenges in model-fitting. We have shown that we can separate the estimation of regression parameter (which is responsible for variable selection) and the estimation of the nonparametric monotone transformation function. A rate-optimal estimator of the regression parameter has been proposed based on a novel penalized composite probit regression estimator. We have also developed an efficient and stable algorithm to compute the proposed estimator. Our numerical experiments have confirmed the promising performance of the new method.

There are several worthy discussion points. We have shown the results of using three probit regression models in composite probit regression. We have also tried a single probit regression model and it generally performs worse than the composite one. For sake of space, those results are not shown in the paper. If we combine more probit regression models, the results may get further improved, although the computation time is longer. To fix idea, we have used the equal weights in composite probit regression in the numerical studies. Though the results are encouraging, it is interesting to study the optimal weights for a given problem. A possible approach is to check the asymptotic efficiency of the oracle estimator given the composite weights and then optimize the efficiency with respect to the weights. This approach was studied in Bradic et al. (2011) for penalized composite quasi-likelihood estimators. We leave it in a future paper.

Appendix: technical Proofs

Proof of Lemma 1.

Proof. The second derivative of the probit loss function is

$$L''(t) = \frac{t\varphi(t)}{\Phi(t)} + \left(\frac{\varphi(t)}{\Phi(t)}\right)^2. \tag{30}$$

To show that $L''(t) > 0$, it is equivalent to show that $f(t) := t\Phi(t) + \varphi(t) > 0$. We notice that $f(t)$ is a strictly increasing function because $f'(t) = \Phi(t) > 0$. Then it remains to show that

$f(t) \to 0$ as $t \to -\infty$. Since $0 \geq t\Phi(t) \geq \int_{-\infty}^{t} s\varphi(s)ds \to 0$ as $t \to -\infty$ by dominated convergence theorem, then $\lim_{t\to-\infty} t\Phi(t) = 0$. Therefore, $\lim_{t\to-\infty}(t\Phi(t) + \varphi(t)) = 0$.

To show that $L''(t) < 1$, it is equivalent to show that $G(t) = \Phi^2(t) - \varphi^2(t) - t\varphi(t)\Phi(t) > 0$. Since $\lim_{t\to-\infty} G(t) = 0$, then it suffices to show that $G'(t) = \varphi(t)\left[t\varphi(t) + \Phi(t)(1+t^2)\right] > 0$, or equivalently, $\tilde{G}(t) := t\varphi(t) + \Phi(t)(1+t^2) > 0$. This can be derived from the facts that $\lim_{t\to-\infty} \tilde{G}(t) = 0$ and $\tilde{G}'(t) = 2(\varphi(t) + t\Phi(t)) = 2f(t) > 0$. □

Proof of Theorem 1.

Proof. By Theorem 1 of Fan et al. (2014), we have that under the event

$$\mathcal{E}_1 = \{\|\hat{\mathcal{B}}^{\text{initial}} - \mathcal{B}^*\|_{\max} \leq a_0\lambda\} \cap \{\|\nabla_{\mathcal{A}^c}\tilde{M}_n(\hat{\mathcal{B}}^{\text{oracle}})\|_{\max} < a_1\lambda\},$$

the LLA algorithm initialized by $\hat{\mathcal{B}}^{\text{initial}}$ finds $\hat{\mathcal{B}}^{\text{oracle}}$ after one iteration. And by Theorem 2 of Fan et al. (2014), we have under the event

$$\mathcal{E}_2 = \{\|\nabla_{\mathcal{A}^c}\tilde{M}_n(\hat{\mathcal{B}}^{\text{oracle}})\|_{\max} < a_1\lambda\} \cap \{\|\hat{\mathcal{B}}^{\text{oracle}}_{\mathcal{A}}\|_{\min} > a\lambda\},$$

if $\hat{\mathcal{B}}^{\text{oracle}}$ is obtained, the LLA algorithm will find $\hat{\mathcal{B}}^{\text{oracle}}$ again in the next iteration, that is, it converges to $\hat{\mathcal{B}}^{\text{oracle}}$ in the next iteration and is a fixed point. Therefore, under the event $\mathcal{E} = \{\|\hat{\mathcal{B}}^{\text{initial}} - \mathcal{B}^*\|_{\max} \leq a_0\lambda\} \cap \{\|\nabla_{\mathcal{A}^c}\tilde{M}_n(\hat{\mathcal{B}}^{\text{oracle}})\|_{\max} < a_1\lambda\} \cap \{\|\hat{\mathcal{B}}^{\text{oracle}}_{\mathcal{A}}\|_{\min} > a\lambda\}$, the LLA algorithm initialized by $\hat{\mathcal{B}}^{\text{initial}}$ converges to $\hat{\mathcal{B}}^{\text{oracle}}$ after two iterations. It remains to bound $\delta_1 = \mathbb{P}\left(\|\nabla_{\mathcal{A}^c}\tilde{M}_n(\hat{\mathcal{B}}^{\text{oracle}})\|_{\max} \geq a_1\lambda\right)$ and $\delta_2 = \mathbb{P}\left(\|\hat{\mathcal{B}}^{\text{oracle}}_{\mathcal{A}}\|_{\min} \leq a\lambda\right)$.

By the definition (22) of the oracle estimator of this probit regression model, the first-order optimality condition for $\hat{\mathcal{B}}^{\text{oracle}}$ is

$$\sum_{k=1}^{K} w_k \left(\boldsymbol{X}^k_{\mathcal{A}}\right)^\top H^k(\hat{\mathcal{B}}^{\text{oracle}}) \left(-\tilde{y}^k + p^k(\hat{\mathcal{B}}^{\text{oracle}})\right) = 0. \tag{31}$$

We now use this to bound $\delta_2 = \mathbb{P}\left(\|\hat{\mathcal{B}}^{\text{oracle}}_{\mathcal{A}}\|_{\min} \leq a\lambda\right)$.

Let

$$r = 2Q_2 \cdot \left\| \sum_{k=1}^{K} w_k \frac{1}{n} (\boldsymbol{X}^k_{\mathcal{A}})^\top H^k(\mathcal{B}^*) \left(\tilde{y}^k - p^k(\mathcal{B}^*)\right) \right\|_{\max},$$

and $\mathbb{B}(r) = \left\{\boldsymbol{\Delta} \in \mathbb{R}^{K+p}, \|\boldsymbol{\Delta}_{\mathcal{A}}\|_{\max} \leq r, \boldsymbol{\Delta}_{\mathcal{A}^c} = 0\right\}$. Define a map $F : \mathbb{B}(r) \subset \mathbb{R}^{K+p} \to \mathbb{R}^{K+p}$, $F(\boldsymbol{\Delta}) = (F_{\mathcal{A}}(\boldsymbol{\Delta}_{\mathcal{A}})^\top, \mathbf{0}^\top)^\top$ with

$$F_{\mathcal{A}}(\boldsymbol{\Delta}_{\mathcal{A}}) := \left[\sum_{k=1}^{K} w_k (\boldsymbol{X}^k_{\mathcal{A}})^\top \Sigma^k(\mathcal{B}^*) \boldsymbol{X}^k_{\mathcal{A}})\right]^{-1} \cdot \left[\sum_{k=1}^{K} w_k (\boldsymbol{X}^k_{\mathcal{A}})^\top H^k(\mathcal{B}^* + \boldsymbol{\Delta}) \left(\tilde{y}^k - p^k(\mathcal{B}^* + \boldsymbol{\Delta})\right)\right] + \boldsymbol{\Delta}_{\mathcal{A}}.$$

Our aim is to show

$$F(\mathbb{B}(r)) \subset \mathbb{B}(r), \tag{32}$$

when

$$\left\| \sum_{k=1}^{K} w_k \frac{1}{n} (\boldsymbol{X}_{\mathcal{A}}^k)^\intercal H^k(\mathcal{B}^*) \left(\tilde{y}^k - p^k(\mathcal{B}^*) \right) \right\|_{\max} \leq \frac{1}{4\tau Q_1 Q_2^2 s}. \tag{33}$$

If (32) holds, by the Brouwer's fixed-point theorem, there always exists a fixed point $\hat{\boldsymbol{\Delta}} \in \mathbb{B}(r)$ such that $F(\hat{\boldsymbol{\Delta}}) = \hat{\boldsymbol{\Delta}}$. It immediately follows that $\sum_{k=1}^{K} w_k (\boldsymbol{X}_{\mathcal{A}}^k)^\intercal H^k(\mathcal{B}^* + \hat{\boldsymbol{\Delta}}) \left(\tilde{y}^k - p^k(\mathcal{B}^* + \hat{\boldsymbol{\Delta}}) \right) = \mathbf{0}$ and $\hat{\boldsymbol{\Delta}}_{A^c} = \mathbf{0}$, which implies that $\mathcal{B}^* + \hat{\boldsymbol{\Delta}} = \hat{\mathcal{B}}^{\text{oracle}}$ by uniqueness of the solution to (31). Thus

$$\|\hat{\mathcal{B}}^{\text{oracle}} - \mathcal{B}^*\|_{\max} = \|\hat{\boldsymbol{\Delta}}\|_{\max} \leq r. \tag{34}$$

If further

$$\left\| \sum_{k=1}^{K} w_k \frac{1}{n} (\boldsymbol{X}_{\mathcal{A}}^k)^\intercal H^k(\mathcal{B}^*) \left(\tilde{y}^k - p^k(\mathcal{B}^*) \right) \right\|_{\max} \leq \frac{1}{2Q_2} \left(\|\mathcal{B}_{\mathcal{A}}^*\|_{\min} - a\lambda \right), \tag{35}$$

we have $r \leq \left(\|\mathcal{B}_{\mathcal{A}}^*\|_{\min} - a\lambda \right)$, and then $\|\hat{\mathcal{B}}_{\mathcal{A}}^{\text{oracle}}\|_{\min} \geq a\lambda$. Therefore, we have δ_2 can be upper bounded by the following probability:

$$\mathbb{P}\left(\left\| \sum_{k=1}^{K} w_k \frac{1}{n} (\boldsymbol{X}_{\mathcal{A}}^k)^\intercal H^k(\mathcal{B}^*) \left(\tilde{y}^k - p^k(\mathcal{B}^*) \right) \right\|_{\max} > \min \left\{ \frac{1}{4\tau Q_1 Q_2^2 s}, \frac{1}{2Q_2} (\|\mathcal{B}_{\mathcal{A}}^*\|_{\min} - a\lambda) \right\} \right).$$

By the Union bound and Hoeffding's bound, we have

$$\delta_2 \leq 2s \exp\left(-\frac{2n}{Q_4^2 M} \min \left\{ \frac{1}{16\tau^2 Q_1^2 Q_2^4 s^2}, \frac{1}{4Q_2^2} \left(\|\mathcal{B}_{\mathcal{A}}^*\|_{\min} - a\lambda \right)^2 \right\} \right) =: \delta_2^{\text{cpr}}. \tag{36}$$

We now derive (32). Using Taylor expansion around $\mathbf{0}$, $\forall \boldsymbol{\Delta} \in \mathbb{B}(r)$, for $j \in \mathcal{A}$,

$$\begin{aligned}
&\sum_{k=1}^{K} w_k (\boldsymbol{x}_{(j)}^k)^\intercal H^k(\mathcal{B}^* + \boldsymbol{\Delta}) \left(-\tilde{y}^k + p^k(\mathcal{B}^* + \boldsymbol{\Delta}) \right) \\
=&\sum_{k=1}^{K} w_k (\boldsymbol{x}_{(j)}^k)^\intercal H^k(\mathcal{B}^*) \left(-\tilde{y}^k + p^k(\mathcal{B}^*) \right) + \sum_{k=1}^{K} w_k (\boldsymbol{x}_{(j)}^k)^\intercal \Sigma^k(\mathcal{B}^*) \boldsymbol{X}^k \boldsymbol{\Delta} + R_j(\tilde{\boldsymbol{\Delta}}_{(j)}),
\end{aligned}$$

where $R_j(\tilde{\boldsymbol{\Delta}}_{(j)}) = \sum_{k=1}^{K} w_k (\boldsymbol{x}_{(j)}^k)^\intercal \left(\Sigma^k(\mathcal{B}^* + \tilde{\boldsymbol{\Delta}}_{(j)}) - \Sigma^k(\mathcal{B}^*) \right) \boldsymbol{X}^k \boldsymbol{\Delta}$ with $\tilde{\boldsymbol{\Delta}}_{(j)}$ on the line segment between $\mathbf{0}$ and $\boldsymbol{\Delta}$. Since $\boldsymbol{\Delta}_{\mathcal{A}^c} = \mathbf{0}$, we have $\boldsymbol{X}^k \boldsymbol{\Delta} = \boldsymbol{X}_{\mathcal{A}}^k \boldsymbol{\Delta}_{\mathcal{A}}$. By the mean value theorem, we have

$$\max_{j \in \mathcal{A}} |R_j(\tilde{\boldsymbol{\Delta}}_{(j)})| \leq \sum_{k=1}^{K} w_k \max_{j \in \mathcal{A}} \boldsymbol{\Delta}_{\mathcal{A}}^\intercal (\boldsymbol{X}_{\mathcal{A}}^k)^\intercal \operatorname{diag} \left\{ |\boldsymbol{x}_{(j)}^k| \circ |(\Sigma^k)'(\bar{\mathcal{B}}_{(j)}^k)| \right\} \boldsymbol{X}_{\mathcal{A}}^k \boldsymbol{\Delta}_{\mathcal{A}}$$

for $\bar{\mathcal{B}}_{(j)}^k$ being on the line segment joining $\mathcal{B}^*$ and $\mathcal{B}^* + \boldsymbol{\Delta}$. Using the fact that $|(\Sigma^k)'(\bar{\mathcal{B}}_{(j)}^k)| \leq \tau$, we have

$$\max_{j \in \mathcal{A}} |R_j(\tilde{\boldsymbol{\Delta}}_{(j)})| \leq \sum_{k=1}^{K} w_k \tau n Q_1 \|\boldsymbol{\Delta}_{\mathcal{A}}\|_{\ell_2}^2 \leq \tau n Q_1 s r^2. \tag{37}$$

Notice that $F_{\mathcal{A}}(\mathbf{\Delta}_{\mathcal{A}})$ can be written as

$$\left[\sum_{k=1}^{K} w_k (\boldsymbol{X}_{\mathcal{A}}^k)^\intercal \Sigma^k(\mathcal{B}^*) \boldsymbol{X}_{\mathcal{A}}^k\right]^{-1} \cdot \left[\sum_{k=1}^{K} w_k (\boldsymbol{X}_{\mathcal{A}}^k)^\intercal H^k(\mathcal{B}^* + \mathbf{\Delta}) \left(\tilde{y}^k - p^k(\mathcal{B}^* + \mathbf{\Delta})\right)\right] + \mathbf{\Delta}_{\mathcal{A}}$$
$$= \left[\sum_{k=1}^{K} w_k (\boldsymbol{X}_{\mathcal{A}}^k)^\intercal \Sigma^k(\mathcal{B}^*) \boldsymbol{X}_{\mathcal{A}}^k\right]^{-1} \cdot \left[\sum_{k=1}^{K} w_k (\boldsymbol{X}_{\mathcal{A}}^k)^\intercal H^k(\mathcal{B}^*) \left(\tilde{y} - p^k(\mathcal{B}^*)\right) - \mathbf{R}_{\mathcal{A}}(\tilde{\mathbf{\Delta}})\right],$$

where $\mathbf{R}_{\mathcal{A}}(\tilde{\mathbf{\Delta}}) = \left(R_j(\tilde{\mathbf{\Delta}}_{(j)}), j \in \mathcal{A}\right)^\intercal$. Then using the triangle inequality to obtain

$$\|F_{\mathcal{A}}(\mathbf{\Delta}_{\mathcal{A}})\|_{\max} \le Q_2 \cdot \left(\left\|\sum_{k=1}^{K} w_k \frac{1}{n} (\boldsymbol{X}_{\mathcal{A}}^k)^\intercal H^k(\mathcal{B}^*) \left(\tilde{y}^k - p^k(\mathcal{B}^*)\right)\right\|_{\max} + \left\|\frac{1}{n}\mathbf{R}_{\mathcal{A}}(\tilde{\mathbf{\Delta}})\right\|_{\max}\right).$$

By (37) and the definition of r, we have $\|F_{\mathcal{A}}(\mathbf{\Delta}_{\mathcal{A}})\|_{\max} \le \frac{r}{2} + Q_1 Q_2 \tau s r^2 \le r$ when the event (33) holds. This established the desired contraction (32).

Next, we prove the upper bound for $\delta_1 = \mathbb{P}(\|\nabla_{\mathcal{A}^c} \ell_n(\hat{\mathcal{B}}^{\text{oracle}})\|_\infty \ge a_1\lambda)$. By Taylor expansion, for all j,

$$\nabla_j \tilde{M}_n(\mathcal{B}^* + \hat{\mathbf{\Delta}}) = \nabla_j \tilde{M}_n(\mathcal{B}^*) + \frac{1}{n}\sum_{k=1}^{K} w_k (\boldsymbol{x}_{(j)}^k)^\intercal \Sigma^k(\mathcal{B}) \boldsymbol{X}^k \hat{\mathbf{\Delta}} + \frac{1}{n} R_j(\bar{\mathbf{\Delta}}_{(j)})$$

where $R_j(\bar{\mathbf{\Delta}}_{(j)}) = \sum_{k=1}^{K} w_k (\boldsymbol{x}_{(j)}^k)^\intercal \left(\Sigma^k(\mathcal{B}^* + \bar{\mathbf{\Delta}}_{(j)}) - \Sigma^k(\mathcal{B}^*)\right)\right) \boldsymbol{X}^k \hat{\mathbf{\Delta}}$ and $\bar{\mathbf{\Delta}}_{(j)}$ being on the line segment between $\mathbf{0}$ and $\hat{\mathbf{\Delta}}$. Denote

$$\mathbf{R}_{\mathcal{A}}(\bar{\mathbf{\Delta}}) = \left(R_j(\bar{\mathbf{\Delta}}_{(j)}), j \in \mathcal{A}\right)^\intercal, \ \mathbf{R}_{\mathcal{A}^c}(\bar{\mathbf{\Delta}}) = \left(R_j(\bar{\mathbf{\Delta}}_{(j)}), j \in \mathcal{A}^c\right)^\intercal.$$

Then we obtain

$$\nabla_{\mathcal{A}} \tilde{M}_n(\mathcal{B}^* + \hat{\mathbf{\Delta}}) = \nabla_{\mathcal{A}} \tilde{M}_n(\mathcal{B}^*) + \frac{1}{n}\sum_{k=1}^{K} w_k (\boldsymbol{X}_{\mathcal{A}}^k)^\intercal \Sigma^k(\mathcal{B}^*) \boldsymbol{X}_{\mathcal{A}}^k \hat{\mathbf{\Delta}}_{\mathcal{A}} + \frac{1}{n}\mathbf{R}_{\mathcal{A}}(\bar{\mathbf{\Delta}}),$$

and

$$\nabla_{\mathcal{A}^c} \tilde{M}_n(\mathcal{B}^* + \hat{\mathbf{\Delta}}) = \nabla_{\mathcal{A}^c} \tilde{M}_n(\mathcal{B}^*) + \frac{1}{n}\sum_{k=1}^{K} w_k (\boldsymbol{X}_{\mathcal{A}^c}^k)^\intercal \Sigma^k(\mathcal{B}^*) \boldsymbol{X}_{\mathcal{A}}^k \hat{\mathbf{\Delta}}_{\mathcal{A}} + \frac{1}{n}\mathbf{R}_{\mathcal{A}^c}(\bar{\mathbf{\Delta}}),$$

Since $\nabla_{\mathcal{A}} \tilde{M}_n(\mathcal{B}^* + \hat{\mathbf{\Delta}}) = \mathbf{0}$, we have

$$\begin{aligned}
&\nabla_{\mathcal{A}^c} \tilde{M}_n(\mathcal{B}^* + \hat{\mathbf{\Delta}}) \\
&= \left[\sum_{k=1}^{K} w_k (\boldsymbol{X}_{\mathcal{A}^c}^k)^\intercal \Sigma^k(\mathcal{B}^*) \boldsymbol{X}_{\mathcal{A}}^k\right] \cdot \left[\sum_{k=1}^{K} w_k (\boldsymbol{X}_{\mathcal{A}}^k)^\intercal \Sigma^k(\mathcal{B}^*) \boldsymbol{X}_{\mathcal{A}}^k\right]^{-1} \left(-\nabla_{\mathcal{A}} \tilde{M}_n(\mathcal{B}^*) - \frac{1}{n}\mathbf{R}_{\mathcal{A}}(\bar{\mathbf{\Delta}})\right) \\
&+ \nabla_{\mathcal{A}^c} \tilde{M}_n(\mathcal{B}^*) + \frac{1}{n}\mathbf{R}_{\mathcal{A}^c}(\bar{\mathbf{\Delta}}).
\end{aligned} \tag{38}$$

Recall that we have proved that (34) holds under the condition (33). Now under the condition (33) and the additional event

$$\left\{\left\|\nabla_{\mathcal{A}^c} \tilde{M}_n(\mathcal{B}^*)\right\|_{\max} < \frac{a_1\lambda}{2}\right\} \cap \left\{\left\|\nabla_{\mathcal{A}} \tilde{M}_n(\mathcal{B}^*)\right\|_{\max} < \frac{a_1\lambda}{4Q_3 + 2}\right\},$$

we can follow the same lines of proof as in (37) to show that

$$\|\mathbf{R}(\bar{\mathbf{\Delta}})\|_{\max} \le \tau n Q_1 \|\hat{\Delta}_{\mathcal{A}}\|_{\ell_2}^2 \le \tau n Q_1 s r^2.$$

Noticing that under condition (33),

$$\tau n Q_1 s r^2 = 4\tau n Q_1 Q_2^2 s \|\nabla_{\mathcal{A}} \tilde{M}_n(\mathcal{B}^*)\|_{\max}^2 \le n \|\nabla_{\mathcal{A}} \tilde{M}_n(\mathcal{B}^*)\|_{\max},$$

which implies that $\|\mathbf{R}(\bar{\mathbf{\Delta}})\|_{\max} \le n\|\nabla_{\mathcal{A}} \tilde{M}_n(\mathcal{B}^*)\|_{\max}$. Under the same event and by (38), we have

$$\begin{aligned}
&\|\nabla_{\mathcal{A}^c} \tilde{M}_n(\mathcal{B}^* + \hat{\mathbf{\Delta}})\|_{\max} \\
&\le Q_3 \cdot \left(\|\nabla_{\mathcal{A}} \tilde{M}_n(\mathcal{B}^*)\|_{\max} + \frac{1}{n} \left\| \mathbf{R}_{\mathcal{A}}(\bar{\mathbf{\Delta}}) \right\| \right) + \|\nabla_{\mathcal{A}^c} \tilde{M}_n(\mathcal{B}^*)\|_{\max} + \frac{1}{n} \left\| \mathbf{R}_{\mathcal{A}^c}(\bar{\mathbf{\Delta}}) \right\| \\
&\le (2Q_3 + 1) \cdot \|\nabla_{\mathcal{A}} \tilde{M}_n(\mathcal{B}^*)\|_{\max} + \|\nabla_{\mathcal{A}^c} \tilde{M}_n(\mathcal{B}^*)\|_{\max} < a_1 \lambda.
\end{aligned}$$

Thus,

$$\delta_1 \le \mathbb{P}\left(\|\nabla_{\mathcal{A}} \tilde{M}_n(\mathcal{B}^*)\|_{\max} > \min\left\{ \frac{a_1\lambda}{4Q_3+2}, \frac{1}{4\tau Q_1 Q_2^2 s} \right\} \right) + \mathbb{P}\left(\|\nabla_{\mathcal{A}^c} \tilde{M}_n(\mathcal{B}^*)\|_{\max} > \frac{a_1\lambda}{2} \right).$$

By the Union bound and Hoeffding's inequality, we have that $\delta_1 \le \delta_1^{\mathrm{cpr}}$, where

$$\delta_1^{\mathrm{cpr}} = 2s \exp\left(-\frac{2n}{Q_4^2 M} \min\left\{ \frac{a_1^2\lambda^2}{4(2Q_3+1)^2}, \frac{1}{16\tau^2 Q_1^2 Q_2^4 s^2} \right\} \right) + 2(K+p-s) \exp\left(-\frac{2n}{Q_4^2 M} \frac{a_1^2\lambda^2}{4} \right).$$

This completes the proof of Theorem 1. □

Proof of Theorem 2.

Proof. By definition of Lasso, it obviously holds that

$$\tilde{M}_n(\hat{\mathcal{B}}^{\mathrm{lasso}}) + \lambda_{\mathrm{lasso}} \|\hat{\boldsymbol{\beta}}^{\mathrm{lasso}}\|_{\ell_1} \le \tilde{M}_n(\mathcal{B}^*) + \lambda_{\mathrm{lasso}} \|\boldsymbol{\beta}^*\|_{\ell_1}$$

Use the convexity of ℓ_n, we obtain

$$\tilde{M}_n(\hat{\mathcal{B}}^{\mathrm{lasso}}) \ge \tilde{M}_n(\mathcal{B}^*) + \nabla \tilde{M}_n(\mathcal{B}^*)^\intercal (\hat{\mathcal{B}}^{\mathrm{lasso}} - \mathcal{B}^*).$$

Thus,

$$\nabla \tilde{M}_n(\mathcal{B}^*)^\intercal (\hat{\mathcal{B}}^{\mathrm{lasso}} - \mathcal{B}^*) \le \lambda_{\mathrm{lasso}} (\|\boldsymbol{\beta}^*\|_{\ell_1} - \|\hat{\boldsymbol{\beta}}^{\mathrm{lasso}}\|_{\ell_1}).$$

This entails that on the event

$$\left\{ \|\nabla \tilde{M}_n(\mathcal{B}^*)\|_{\max} \le \frac{1}{2}\lambda_{\mathrm{lasso}} \right\}, \tag{39}$$

we have

$$-\frac{1}{2}\lambda_{\text{lasso}}(\|\hat{\boldsymbol{\Delta}}(\boldsymbol{\beta})\|_{\ell_1}+\|\hat{\boldsymbol{\Delta}}(\mathbf{b}_0)\|_{\ell_1}) \le \lambda_{\text{lasso}}(\|\boldsymbol{\beta}^*_{\mathcal{A}}\|_{\ell_1}-\|\boldsymbol{\beta}_{\mathcal{A}}+\hat{\boldsymbol{\Delta}}_{\mathcal{A}}(\boldsymbol{\beta})\|_{\ell_1}-\|\hat{\boldsymbol{\Delta}}_{\mathcal{A}^c}(\boldsymbol{\beta})\|_{\ell_1}), \tag{40}$$

where $\hat{\boldsymbol{\Delta}}(\boldsymbol{\beta})=\hat{\boldsymbol{\beta}}^{\text{lasso}}-\boldsymbol{\beta}^*$, $\hat{\boldsymbol{\Delta}}(\mathbf{b}_0)=\hat{\mathbf{b}}_0^{\text{lasso}}-\mathbf{b}_0^*$, $\hat{\boldsymbol{\Delta}}=\hat{\mathcal{B}}^{\text{lasso}}-\hat{\mathcal{B}}^*$, which implies that

$$\|\hat{\boldsymbol{\Delta}}_{\mathcal{A}^c}(\boldsymbol{\beta})\|_{\ell_1} \le 3\|\hat{\boldsymbol{\Delta}}_{\mathcal{A}}(\boldsymbol{\beta})\|_{\ell_1}+\|\hat{\boldsymbol{\Delta}}(\mathbf{b}_0)\|_{\ell_1}. \tag{41}$$

In what follows, our aim is to derive the upper bound

$$\|\hat{\mathcal{B}}^{\text{lasso}}-\mathcal{B}^*\|_{\ell_\infty} \le 5\kappa_{\text{cpr}}^{-1}\lambda_{\text{lasso}} \tag{42}$$

under the event (39). Then along with the Hoeffding's inequality,

$$\mathbb{P}\left(\|\nabla\tilde{M}_n(\mathcal{B}^*)\|_{\max}\ge\frac{1}{2}\lambda_{\text{lasso}}\right) \le 2(K+p)\exp\left(-\frac{2n}{Q_4^2M}\left(\frac{1}{2}\lambda_{\text{lasso}}\right)^2\right), \tag{43}$$

one shows the desired probability bound.

Now we consider a map $F:\mathbb{R}^{K+p}\to\mathbb{R}$ satisfying

$$F(\boldsymbol{\Delta})=\tilde{M}_n(\mathcal{B}^*+\boldsymbol{\Delta})-\tilde{M}_n(\mathcal{B}^*)+\lambda_{\text{lasso}}(\|\boldsymbol{\beta}^*+\boldsymbol{\Delta}(\boldsymbol{\beta})\|_{\ell_1}-\|\boldsymbol{\beta}^*\|_{\ell_1}).$$

Then $\hat{\boldsymbol{\Delta}}=\arg\min_{\boldsymbol{\Delta}}F(\boldsymbol{\Delta})$. Since $F(\mathbf{0})=0$, $F(\hat{\boldsymbol{\Delta}})\le 0$, by Lemma 4 of Negahban et al. (2012), it suffices to show that $F(\boldsymbol{\Delta})>0$ for any $\boldsymbol{\Delta}\in\mathcal{D}$, where

$$\mathcal{D}=\left\{\boldsymbol{\Delta}\in\mathbb{R}^{K+p}:\|\boldsymbol{\Delta}_{\mathcal{A}^c}(\boldsymbol{\beta})\|_{\ell_1}\le 3\|\boldsymbol{\Delta}_{\mathcal{A}}(\boldsymbol{\beta})\|_{\ell_1}+\|\boldsymbol{\Delta}(\mathbf{b}_0)\|_{\ell_1}\text{ and }\|\boldsymbol{\Delta}\|_{\ell_\infty}=5\kappa_{\text{cpr}}^{-1}\lambda_{\text{lasso}}\right\}. \tag{44}$$

To this end, we first obtain a lower bound for $\|\boldsymbol{\beta}^*+\boldsymbol{\Delta}(\boldsymbol{\beta})\|_{\ell_1}-\|\boldsymbol{\beta}^*\|$:

$$\begin{aligned}\|\boldsymbol{\beta}^*+\boldsymbol{\Delta}(\boldsymbol{\beta})\|_{\ell_1}-\|\boldsymbol{\beta}^*\|_{\ell_1} &= \|\boldsymbol{\beta}^*_{\mathcal{A}}+\boldsymbol{\Delta}_{\mathcal{A}}(\boldsymbol{\beta})\|_{\ell_1}+\|\boldsymbol{\Delta}_{\mathcal{A}^c}(\boldsymbol{\beta})\|_{\ell_1}-\|\boldsymbol{\beta}^*_{\mathcal{A}}\|_{\ell_1}\\ &\ge \|\boldsymbol{\Delta}_{\mathcal{A}^c}(\boldsymbol{\beta})\|_{\ell_1}-\|\boldsymbol{\Delta}_{\mathcal{A}}(\boldsymbol{\beta})\|_{\ell_1}\end{aligned} \tag{45}$$

Next, we derive a lower bound for $\tilde{M}_n(\mathcal{B}^*+\boldsymbol{\Delta})-\tilde{M}_n(\mathcal{B}^*)$. To simplify notation, we define $G(u)=\tilde{M}_n(\mathcal{B}^*+u\boldsymbol{\Delta})$, $u\in[0,1]$. Then

$$\begin{aligned}G''(u)&=\sum_{k=1}^K w_k\frac{1}{n}\sum_{i=1}^n\Sigma_{ii}^k(\mathcal{B}^*+u\boldsymbol{\Delta})\left((\boldsymbol{x}_i^k)^\top\boldsymbol{\Delta}\right)^2,\\ G'''(u)&=\sum_{k=1}^K w_k\frac{1}{n}\sum_{i=1}^n(\Sigma_{ii}^k)'(\mathcal{B}^*+u\boldsymbol{\Delta})\left((\boldsymbol{x}_i^k)^\top\boldsymbol{\Delta}\right)^3.\end{aligned}$$

For all $i\in\{1,\cdots,n\}$,

$$0\le|(\Sigma_{ii}^k)'(\mathcal{B})|\le\max\left(|(\boldsymbol{x}_i^k)^\top\mathcal{B}|,1\right)\cdot\Sigma_{ii}^k(\mathcal{B}).$$

Then we have

$$|G'''(u)| \leq \sum_{k=1}^{K} w_k \frac{1}{n} \sum_{i=1}^{n} \max\left(|(\boldsymbol{x}_i^k)^\intercal(\mathcal{B}^* + u\boldsymbol{\Delta})|, 1\right) \cdot \Sigma_{ii}^k(\mathcal{B}^* + u\boldsymbol{\Delta}) \left|(\boldsymbol{x}_i^k)^\intercal \boldsymbol{\Delta}\right|^3.$$
$$\leq \max\left\{m^2\|\boldsymbol{\Delta}\|_{\ell_1} \cdot (\|\mathcal{B}^*\|_{\ell_1} + \|\boldsymbol{\Delta}\|_{\ell_1}), m\|\boldsymbol{\Delta}\|_{\ell_1}\right\} G''(u).$$

Let $z_0 = \max\left\{m^2\|\boldsymbol{\Delta}\|_{\ell_1} \cdot (\|\mathcal{B}^*\|_{\ell_1} + \|\boldsymbol{\Delta}\|_{\ell_1}), m\|\boldsymbol{\Delta}\|_{\ell_1}\right\}$, we have

$$|G'''(u)| \leq z_0 G''(u), \quad \forall u \in [0,1].$$

Then by integral, we obtain

$$G(1) - G(0) - G'(0) \geq G''(0)\xi(z_0), \tag{46}$$

where $\xi(z) = z^{-2}(\exp(-z) + z - 1)$. It is guaranteed that $z_0 \leq 1$ as long as

$$m^2\|\boldsymbol{\Delta}\|_{\ell_1} \cdot (\|\mathcal{B}^*\|_{\ell_1} + \|\boldsymbol{\Delta}\|_{\ell_1}) \leq 1 \quad \text{and} \quad m\|\boldsymbol{\Delta}\|_{\ell_1} \leq 1$$

or equivalently,

$$\|\boldsymbol{\Delta}\|_{\ell_1} \leq \sqrt{\left(\frac{1}{2}\|\mathcal{B}^*\|_{\ell_1}\right)^2 + \frac{1}{m^2}} - \frac{1}{2}\|\mathcal{B}^*\|_{\ell_1}.$$

Since $\|\boldsymbol{\Delta}\|_{\ell_\infty} = 5\kappa_{\text{cpr}}^{-1}\lambda_{\text{lasso}}$ and $\|\boldsymbol{\Delta}\|_{\ell_\infty} \leq \|\boldsymbol{\Delta}\|_{\ell_1}$, we have

$$\lambda_{\text{lasso}} \leq \frac{\sqrt{\left(\frac{1}{2}\|\mathcal{B}^*\|_{\ell_1}\right)^2 + \frac{1}{m^2}} - \frac{1}{2}\|\mathcal{B}^*\|_{\ell_1}}{5\kappa_{\text{cpr}}^{-1}}. \tag{47}$$

By simple calculation, it can be shown that $h(z)$ is a decreasing function in $z > 0$. Given that $z \leq 1$ holds by assumption (47) on λ_{lasso}, we have $\xi(z) \geq \xi(1) > 1/3$. Thus, we re-write (46) as

$$\begin{aligned}\tilde{M}_n(\mathcal{B}^* + \boldsymbol{\Delta}) - \tilde{M}_n(\mathcal{B}^*) &> \nabla\tilde{M}_n(\mathcal{B}^*)^\intercal\boldsymbol{\Delta} + \frac{1}{3}\boldsymbol{\Delta}^\intercal\nabla^2\tilde{M}_n(\mathcal{B}^*)\boldsymbol{\Delta} \\ &\geq -\frac{1}{2}\lambda_{\text{lasso}}\|\boldsymbol{\Delta}\|_{\ell_1} + \frac{1}{3}\kappa_{\text{cpr}}\|\boldsymbol{\Delta}\|_{\ell_1} \cdot \|\boldsymbol{\Delta}\|_{\ell_\infty} \end{aligned} \tag{48}$$

which holds under event (39) and by definition of general invertability factor κ_{cpr}. Now under the same event, we combine (45) and (48) to obtain

$$\begin{aligned} F(\boldsymbol{\Delta}) &\geq \frac{1}{3}\kappa_{\text{cpr}}\|\boldsymbol{\Delta}\|_{\ell_1} \cdot \|\boldsymbol{\Delta}\|_{\ell_\infty} - \frac{1}{2}\lambda_{\text{lasso}}\|\boldsymbol{\Delta}\|_{\ell_1} + \lambda_{\text{lasso}}\left(\|\boldsymbol{\Delta}_{\mathcal{A}^c}(\boldsymbol{\beta})\|_{\ell_1} - \|\boldsymbol{\Delta}_{\mathcal{A}}(\boldsymbol{\beta})\|_{\ell_1}\right) \\ &= \frac{5}{3}\lambda_{\text{lasso}}\|\boldsymbol{\Delta}\|_{\ell_1} - \frac{3}{2}\lambda_{\text{lasso}}\|\boldsymbol{\Delta}_{\mathcal{A}}(\boldsymbol{\beta})\|_{\ell_1} - \frac{1}{2}\lambda_{\text{lasso}}\|\boldsymbol{\Delta}(\mathbf{b}_0)\|_{\ell_1} + \frac{1}{2}\lambda_{\text{lasso}}\|\boldsymbol{\Delta}_{\mathcal{A}^c}(\boldsymbol{\beta})\|_{\ell_1} \\ &= \lambda_{\text{lasso}}\left(\frac{1}{6}\|\boldsymbol{\Delta}_{\mathcal{A}}(\boldsymbol{\beta})\|_{\ell_1} + \frac{7}{6}\|\boldsymbol{\Delta}(\mathbf{b}_0)\|_{\ell_1} + \frac{13}{6}\lambda_{\text{lasso}}\|\boldsymbol{\Delta}_{\mathcal{A}^c}(\boldsymbol{\beta})\|_{\ell_1}\right) \\ &> 0. \end{aligned}$$

This completes the proof of Theorem 2.

□

References

Bickel, P. J. and Doksum, K. A. 1981, 'An analysis of transformations revisited', *Journal of the American Statistical Association* **76**(374), 296–311.

Bickel, P. J., Ritov, Y. and Tsybakov, A. B. 2009, 'Simultaneous analysis of Lasso and Dantzig selector', *The Annals of Statistics* **37**(4), 1705–1732.

Box, G. E. and Cox, D. R. 1964, 'An analysis of transformations', *Journal of the Royal Statistical Society: Series B* **26**(2), 211–243.

Bradic, J., Fan, Y. and Wang, W. 2011, 'Penalized composite quasi-likelihood for ultrahigh dimensional variable selection', *Journal of Royal Statistics Society, Series B* **73**(3), 325–349.

Carroll, R. J. 1982, 'Prediction and power transformations when the choice of power is restricted to a finite set', *Journal of the American Statistical Association* **77**(380), 908–915.

Carroll, R. J. and Ruppert, D. 1981, 'On prediction and the power transformation family', *Biometrika* **68**(3), 609–615.

Chen, S. 2002, 'Rank estimation of transformation models', *Econometrica* **70**(4), 1683–1697.

Demidenko, E. 2001, 'Computational aspects of probit model', *Mathematical Communications* **6**(2), 233–247.

Dette, H., Neumeyer, N. and Pilz, K. F. 2006, 'A simple nonparametric estimator of a strictly monotone regression function', *Bernoulli* **12**(3), 469–490.

Draper, N. R. and Cox, D. R. 1969, 'On distributions and their transformation to normality', *Journal of the Royal Statistical Society: Series B* **31**(3), 472–476.

Draper, N. and Smith, H. 1998, *Applied regression analysis*, Vol. 326, John Wiley & Sons.

Fan, J. and Li, R. 2001, 'Variable selection via nonconcave penalized likelihood and its oracle properties', *Journal of the American Statistical Association* **96**(456), 1348–1360.

Fan, J., Li, R., Zhang, C.-H. and Zou, H. 2020, *Statistical Foundations of Data Science*, Chapman and Hall/CRC.

Fan, J. and Lv, J. 2011, 'Nonconcave penalized likelihood with np-dimensionality', *IEEE Transactions on Information Theory* **57**(8), 5467–5484.

Fan, J., Xue, L. and Zou, H. 2014, 'Strong oracle optimality of folded concave penalized estimation', *The Annals of Statistics* **42**(3), 819–849.

Fan, Y., Han, F., Li, W. and Zhou, X.-H. 2020, 'On rank estimators in increasing dimensions', *Journal of Econometrics* **214**(2), 379–412.

Friedman, J., Hastie, T. and Tibshirani, R. 2010, 'Regularization paths for generalized linear models via coordinate descent', *Journal of statistical software* **33**(1), 1–22.

Hall, P. and Huang, L.-S. 2001, 'Nonparametric kernel regression subject to monotonicity constraints', *The Annals of Statistics* **29**(3), 624–647.

Han, A. K. 1987, 'Non-parametric analysis of a generalized regression model: the maximum rank correlation estimator', *Journal of Econometrics* **35**(2-3), 303–316.

Lan, W., Zhong, P.-S., Li, R., Wang, H. and Tsai, C.-L. 2016, 'Testing a single regression coefficient in high dimensional linear models', *Journal of Econometrics* **195**(1), 154–168.

Li, G., Peng, H., Zhang, J. and Zhu, L. 2012, 'Robust rank correlation based screening', *The Annals of Statistics* **40**(3), 1846–1877.

Mammen, E. 1991, 'Estimating a Smooth Monotone Regression Function', *The Annals of Statistics* **19**(2), 724–740.

Negahban, S. N., Ravikumar, P., Wainwright, M. J. and Yu, B. 2012, 'A unified framework for high-dimensional analysis of m-estimators with decomposable regularizers', *Statistical Science* **27**(4), 538–557.

Ramsay, J. O. 1998, 'Estimating smooth monotone functions', *Journal of the Royal Statistical Society: Series B* **60**(2), 365–375.

Sherman, R. P. 1993, 'The limiting distribution of the maximum rank correlation estimator', *Econometrica* **61**(1), 123–137.

Tibshirani, R. 1996, 'Regression shrinkage and selection via the lasso', *Journal of the Royal Statistical Society: Series B* **58**(1), 267–288.

Tukey, J. W. 1957, 'On the comparative anatomy of transformations', *The Annals of Mathematical Statistics* **28**(3), 602–632.

Weisberg, S. 2005, *Applied linear regression*, Vol. 528, John Wiley & Sons.

Yang, Y. and Zou, H. 2013, 'An efficient algorithm for computing the hhsvm and its generalizations', *Journal of Computational and Graphical Statistics* **22**(2), 396–415.

Zhang, C.-H. 2010, 'Nearly unbiased variable selection under minimax concave penalty', *The Annals of Statistics* **38**(2), 894–942.

Zou, H. and Li, R. 2008, 'One-step sparse estimates in nonconcave penalized likelihood models', *The Annals of Statistics* **36**(4), 1509–1533.

Zou, H. and Yuan, M. 2008, 'Composite quantile regression and the oracle model selection theory', *The Annals of Statistics* **36**(3), 1108–1126.