Published in final edited form as:

J Am Stat Assoc. 2016 ; 111(513): 355–376. doi:10.1080/01621459.2015.1008363.

Variable Selection with Prior Information for Generalized Linear Models via the Prior LASSO Method

Yuan Jiang, Yunxiao He, and Heping Zhang

Yuan Jiang is an assistant professor at Department of Statistics, Oregon State University, Corvallis, Oregon 97331-4606. Yunxiao He is an associate director at the Nielsen Company, 770 Broadway, New York, New York 10003-9595. Heping Zhang is a Susan Dwight Bliss Professor at Department of Biostatistics, Yale University School of Public Health, and a Professor at the Child Study Center, Yale University School of Medicine, New Haven, Connecticut 06520-8034. He is also a Chang-Jiang and 1000-plan scholar at Sun Yat-Sen University, Guangzhou, China

Yuan Jiang: yuan.jiang@stat.oregonstate.edu; Yunxiao He: yunxiaohe@gmail.com; Heping Zhang: heping.zhang@yale.edu

Abstract

LASSO is a popular statistical tool often used in conjunction with generalized linear models that can simultaneously select variables and estimate parameters. When there are many variables of interest, as in current biological and biomedical studies, the power of LASSO can be limited. Fortunately, so much biological and biomedical data have been collected and they may contain useful information about the importance of certain variables. This paper proposes an extension of LASSO, namely, prior LASSO (pLASSO), to incorporate that prior information into penalized generalized linear models. The goal is achieved by adding in the LASSO criterion function an additional measure of the discrepancy between the prior information and the model. For linear regression, the whole solution path of the pLASSO estimator can be found with a procedure similar to the Least Angle Regression (LARS). Asymptotic theories and simulation results show that pLASSO provides significant improvement over LASSO when the prior information is relatively accurate. When the prior information is less reliable, pLASSO shows great robustness to the misspecification. We illustrate the application of pLASSO using a real data set from a genome-wide association study.

Keywords

Asymptotic efficiency; Oracle inequalities; Solution path; Weak oracle property

1 INTRODUCTION

In many scientific areas, it is common that experiments with similar goals are performed simultaneously or repetitively to establish a replicable conclusion. A typical example arises from the genome-wide association studies (GWAS). GWAS have become a common practice in exploring the genetic factors underlying complex diseases with the rapid

advancement in high-throughput sequencing technology. It usually happens that multiple studies investigating the same question of interest (e.g., looking for the genetic variants associated with a particular human disease) have been carried out repetitively. This brings in the “prior” information for a current study. For example, previous studies provide us with some knowledge about possible effects of certain genes and/or single nucleotide polymorphisms (SNPs) on a disease, although the existing information may include both true and false positives.

Meanwhile, the massive data collected from current research are challenging the statistical analysis. Taking GWAS as an example again, it is well documented that the large number of genetic markers and the generally small size of their effects make it very unlikely for researchers to detect important genetic factors with a desired statistical significance in a single study. For example, hundreds of thousands of SNPs are typically genotyped, and yet their individual effect sizes are usually small. Therefore, it is imperative for us to utilize the existent biological and biomedical information. This prior information, if used appropriately, can alleviate the “curse of dimensionality” and hence improve the power of a study to discover important genetic factors.

Meta-analysis has actually provided effective means to combine different studies or their results to improve the statistical power, especially for genetic studies as well as biomedical studies broadly (Zeggini and Ioannidis, 2009 ; Cantor et al., 2010). Although meta-analysis has been proven successful in many studies (Zeggini et al., 2008), its applicability is restrictive for the present application. For example, the studies in a meta-analysis have to be similar in certain aspects such as the design of the experiments and the analytical methods. In addition, in genome-wide association studies, data are often collected using different genotyping platforms. As a consequence, the SNPs are different across studies, which makes it difficult to combine the results. While there exist methods to impute SNPs across platforms, the implication is not well understood between the results for genotyped and imputed SNPs. Most analytic methods treat the imputed SNPs naively as if they are genotyped SNPs.

We consider the situation where the data from the current study are analyzed using LASSO (Tibshirani, 1996) in conjunction with generalized linear models (McCullagh and Nelder, 1989). LASSO simultaneously selects a subset of explanatory variables and estimates their effects on the response. The goals of LASSO are achieved by balancing the model fitting and model complexity through a penalized likelihood approach. In general, LASSO considers all explanatory variables equally in terms of penalization. Suppose previous studies offered the plausible information about the importance of explanatory variables (e.g., genes or SNPs in GWAS). To incorporate the prior information, a seemingly natural method is to make use of the adaptive LASSO (Zou, 2006 ; Zhang and Lu, 2007), by assigning different weights of penalization to different variables to reflect various levels of prior evidence for significance. For example, variables with smaller p -values (or larger effect sizes) from previous studies should be penalized less such that they have a greater chance to be selected than the variables with larger p -values (or smaller effect sizes). However, this approach suffers from similar difficulties to those that we mentioned for meta-analysis

methods. Moreover, this approach will perform poorly when the prior information contains a lot of false positives.

In this paper we propose a new solution named prior LASSO (pLASSO) to incorporate prior information into a LASSO problem. To utilize pLASSO, we only need to assume that the prior information can make certain prediction of the response. Then, pLASSO imposes an extra term using these predictive response values on the criterion function of LASSO to incorporate the prior information. In addition to the balance between model fitting and complexity in LASSO, pLASSO further balances the trade-off between the data and the prior information. It is noteworthy that pLASSO is a very flexible framework that can handle different types of prior information, including information presented as a set of significant explanatory variables (genes/SNPs), the effect sizes of some explanatory variables (genes/SNPs), or even the response values (disease status) given a set of explanatory variables (gene/SNP conditions).

Computationally, we investigate the solution path of pLASSO where the path reflects the change of belief in the prior information. Compared to the piecewise linear paths in the Least Angle Regression (LARS) of Efron et al. (2004) and in the similar work by Rosset and Zhu (2007), our path is piecewise smooth but not linear. Similar to the relationship between LARS and LASSO, our solution path fully characterizes the local effect of incorporating the prior information into a LASSO problem.

Theoretically, we compare the asymptotic properties of the LASSO and pLASSO estimators in three aspects: the asymptotic distribution in the low-dimensional setting, the weak oracle property in the high-dimensional setting, and the oracle inequalities for excess risk and estimation error in the high-dimensional setting. All three comparisons show that pLASSO results in a more efficient estimator than LASSO when the prior information is reasonably reliable. They also illustrate the trade-off between the variance reduction by incorporating the correct prior information and the additive bias due to the prior error.

Finally, numerically, the simulation results verify that pLASSO provides significant improvement over LASSO when the prior information is accurate. We will also see that when the prior information is less reliable, pLASSO shows great robustness to the misspecification. To illustrate the applicability of pLASSO to real data, we re-analyze a genetic data set on bipolar disorder using both LASSO and pLASSO and compare their results.

2 PRIOR LASSO

2.1 LASSO

Assume that we have a collection of independent observations $\{(\mathbf{X}_i, Y_i)\}_{i=1}^n$, where Y_i 's are independently observed response values given the p -dimensional predictor vector $\mathbf{X}_i$. Let $\mathbf{Y} = (Y_1, \dots, Y_n)^T$ and $\mathbf{X}$ be the design matrix with $\mathbf{X}_i$ as its i -th row. Assume that, with a canonical link, the conditional distribution of Y_i given $\mathbf{X}_i$ belongs to the canonical exponential family with the following density function (ignoring a multiplier that depends on Y_i only):

$$f(Y_i|\theta_i) \propto \exp[Y_i\theta_i - b(\theta_i)], \quad (1)$$

where $\theta_i = \mathbf{Z}_i^T \boldsymbol{\beta}$ with $\mathbf{Z}_i = (1, \mathbf{X}_i^T)^T$. As is common in generalized linear models, $b(\theta)$ is assumed to be twice continuously differentiable with positive $b''(\theta)$. The loss function (the negative log-likelihood function) of a generalized linear model is given by

$$L(\boldsymbol{\beta}; \mathbb{X}, \mathbf{Y}) = -\frac{1}{n} \sum_{i=1}^n [Y_i \theta_i - b(\theta_i)] = -\frac{1}{n} \sum_{i=1}^n [Y_i \mathbf{Z}_i^T \boldsymbol{\beta} - b(\mathbf{Z}_i^T \boldsymbol{\beta})], \quad (2)$$

where $\boldsymbol{\beta} = (\beta_0, \beta_1, \dots, \beta_p)^T$.

Traditionally, the maximum likelihood estimator (MLE) of $\boldsymbol{\beta}$, denoted by $\hat{\boldsymbol{\beta}}$, is obtained by minimizing $L(\boldsymbol{\beta}; \mathbb{X}, \mathbf{Y})$. To select a subset of predictors, LASSO adds an L_1 -norm penalty term of $(\beta_1, \dots, \beta_p)^T$ to $L(\boldsymbol{\beta}; \mathbb{X}, \mathbf{Y})$. Thus, the LASSO estimator $\hat{\boldsymbol{\beta}}^\lambda$ is calculated by minimizing the following L_1 -norm penalized loss function:

$$L_\lambda(\boldsymbol{\beta}; \mathbb{X}, \mathbf{Y}) = L(\boldsymbol{\beta}; \mathbb{X}, \mathbf{Y}) + \lambda \sum_{j=1}^p |\beta_j|, \quad (3)$$

where λ is a tuning parameter.

The following well-known observation about a generalized linear model provides a good intuition on how LASSO works. Note that the term $-[Y_i \mathbf{Z}_i^T \boldsymbol{\beta} - b(\mathbf{Z}_i^T \boldsymbol{\beta})]$ in (2) is proportional to the Kullback-Leibler (KL) distance between the distribution in (1) and the following distribution:

$$f(Y_i|\theta_{i1}) \propto \exp[Y_i \theta_{i1} - b(\theta_{i1})], \text{ where } \theta_{i1} \text{ is determined by } b'(\theta_{i1}) = Y_i. \quad (4)$$

Model (4) is the so-called saturated model which often serves as the reference model for evaluating the performance of a fitted model (Collett, 2003). In other words, a generalized linear model is a model reduction tool to approximate the saturated model. The idea of LASSO is to balance the trade-off between the approximation accuracy and the model complexity.

2.2 Prior LASSO

As discussed in the Introduction, we focus on the prior information about possible effects of certain predictors on the response. It can be a set of significant predictors, their effect sizes, or the predictive values given a set of predictors. Mathematically, these can be presented as some existing knowledge about the support of $\boldsymbol{\beta}$, the values of $\boldsymbol{\beta}$, or the predicted values of the response, etc. For the time being, let us assume that a preprocessing step can summarize

the prior information into the predicted responses denoted by $\hat{\mathbf{Y}}^P = (\hat{Y}_1^P, \dots, \hat{Y}_n^P)^T$.

Depending on the nature of the prior information, different methods may be used in this step, and we shall revisit this issue below.

To incorporate the prior information, we need to weigh between the prior information and the presently observed data. Then, in the same spirit of the loss function $L(\beta; \mathbb{X}, \mathbf{Y})$, it is natural to add the KL distance between the distribution in (1) and the following distribution in $L_\lambda(\beta; \mathbb{X}, \mathbf{Y})$:

$$f(Y_i|\theta_{i2}) \propto \exp[Y_i\theta_{i2}-b(\theta_{i2})], \text{ where } \theta_{i2} \text{ is determined by } b'(\theta_{i2})=\hat{Y}_i^P. \quad (5)$$

Consequently, the pLASSO estimator $\hat{\beta}^{\lambda,\eta}$ is defined to be the minimizer of the following criterion function:

$$L_{\lambda,\eta}(\beta; \mathbb{X}, \mathbf{Y}, \hat{\mathbf{Y}}^P) = L(\beta; \mathbb{X}, \mathbf{Y}) + \eta L(\beta; \mathbb{X}, \hat{\mathbf{Y}}^P) + \lambda \sum_{j=1}^p |\beta_j|, \quad (6)$$

where η is another tuning parameter playing in a similar role to λ . The parameter η balances the relative importance of the data themselves and the prior information. In the extreme case of $\eta = 0$, pLASSO is reduced to LASSO. If $\eta = \infty$, pLASSO will solely rely on the prior information to fit the model. Furthermore, the tuning parameter η has another appealing interpretation: it controls the variance of β in its prior distribution from a Bayesian viewpoint (see Section 2.3.1).

We can simplify $L_{\lambda,\eta}(\beta; \mathbb{X}, \mathbf{Y}, \hat{\mathbf{Y}}^P)$ as follows:

$$\begin{aligned} L_{\lambda,\eta}(\beta; \mathbb{X}, \mathbf{Y}, \hat{\mathbf{Y}}^P) &= -\frac{1}{n} \sum_{i=1}^n [(Y_i + \eta \hat{Y}_i^P) \mathbf{Z}_i^T \beta - (1 + \eta) b(\mathbf{Z}_i^T \beta)] + \lambda \sum_{j=1}^p |\beta_j| \\ &\propto -\frac{1}{n} \sum_{i=1}^n \left[\frac{Y_i + \eta \hat{Y}_i^P}{1 + \eta} \mathbf{Z}_i^T \beta - b(\mathbf{Z}_i^T \beta) \right] + \frac{\lambda}{1 + \eta} \sum_{j=1}^p |\beta_j| \\ &= L_{\lambda/(1+\eta)}(\beta; \mathbb{X}, \tilde{\mathbf{Y}}), \end{aligned} \quad (7)$$

where $\tilde{\mathbf{Y}} = (\mathbf{Y} + \eta \hat{\mathbf{Y}}^P)/(1 + \eta)$ is the adjusted response values using the prior information. The pLASSO criterion function is in exactly the same form as the LASSO criterion function in (3). Essentially any LASSO fitting algorithm can be used to solve pLASSO.

As discussed above, a preprocessing step is needed to summarize the prior information into predictions $\hat{\mathbf{Y}}^P$. For generalized linear models, we can obtain the predicted values of the response by plugging in a prior estimator $\hat{\beta}$ of β and then calculate $\hat{Y}_i^P = b'(\mathbf{Z}_i^T \hat{\beta})$, i.e., the expectation of Y_i given $\mathbf{X}_i$ and $\hat{\beta}$.

A common type of prior information is about the support of β . A subset of predictors, denoted by $\mathcal{S}^P$, have been identified to be associated with the response by previous studies. The following simple version of weighted LASSO can be employed to force the predictors $\mathcal{S}^P$ to stay in the model,

$$L_{\kappa, \mathcal{S}^P}(\beta; \mathbb{X}, \mathbf{Y}) = -\frac{1}{n} \sum_{i=1}^n [Y_i \mathbf{Z}_i^T \beta - b(\mathbf{Z}_i^T \beta)] + \kappa \sum_{j=1}^p |\beta_j| I(X_j \notin \mathcal{S}^P). \quad (8)$$

The resultant estimator from (8) can be used as the prior estimator $\hat{\beta}$. In fact, minimizing (8) is to compute the predictions $\hat{Y}^p$ given that we fully trust the previously identified signals in S^p . Certainly, we also consider the variables not in S^p but impose a L_1 penalty on them. By doing so, both $\hat{\beta}$ and $\hat{Y}^p$ depend directly on the reliability of the prior information. Therefore, the pLASSO criterion function in (6) further balances the trade-off between the data and the prior information summarized in $\hat{Y}^p$ from (8).

2.3 Special Cases

In this subsection, we consider the application of pLASSO to linear regression and logistic regression—two simplest, yet most frequently used generalized linear models.

2.3.1 Linear regression—Assume both $\mathbf{Y}$ and $\mathbf{X}$ are centralized, so we can set $\beta_0 = 0$ and let $\beta = (\beta_1, \dots, \beta_p)^T$ without ambiguity. Setting $b(\theta) = \theta^2/2$ in (2) leads to the loss function for linear regression

$$L(\beta; \mathbf{X}, \mathbf{Y}) = \frac{1}{2n} (\mathbf{Y} - \mathbf{X}\beta)^T (\mathbf{Y} - \mathbf{X}\beta). \quad (9)$$

Then the pLASSO criterion function for linear regression can be written as

$$L_{\lambda, \eta}(\beta; \mathbf{X}, \mathbf{Y}, \hat{\mathbf{Y}}^p) = L(\beta; \mathbf{X}, \mathbf{Y}) + \eta L(\beta; \mathbf{X}, \hat{\mathbf{Y}}^p) + \lambda \sum_{j=1}^p |\beta_j| \quad (10)$$

$$\propto \frac{1}{2n} (\tilde{\mathbf{Y}} - \mathbf{X}\beta)^T (\tilde{\mathbf{Y}} - \mathbf{X}\beta) + \frac{\lambda}{1+\eta} \sum_{j=1}^p |\beta_j|, \quad (11)$$

where L is defined in (9). It is noteworthy that $\{\tilde{Y}_i: i = 1, \dots, n\}$ in $\tilde{\mathbf{Y}}$ are usually not statistically independent, as $\{\hat{Y}_i^p: i = 1, \dots, n\}$ are in general not statistically independent.

It is also interesting to interpret the above criterion function from the Bayesian viewpoint. The middle term in (10) can be considered from a prior distribution of the parameters β as $p(\beta) \propto \exp\{-\eta(\hat{\mathbf{Y}}^p - \mathbf{X}\beta)^T(\hat{\mathbf{Y}}^p - \mathbf{X}\beta)/(2n)\}$, provided that $\hat{\mathbf{Y}}^p$ is known. Assume that $\hat{\mathbf{Y}}^p$ is obtained from an estimator $\hat{\beta}$ of β by $\hat{\mathbf{Y}}^p = \mathbf{X}\hat{\beta}$. Then $p(\beta) \propto \exp\{-(\beta - \hat{\beta})^T(\eta \mathbf{X}^T \mathbf{X})(\beta - \hat{\beta})/(2n)\}$ is the density of a normal distribution with mean $\hat{\beta}$ and a variance inversely proportional to η . Hence, the weight of the prior information increases as η increases. We shall revisit this observation in Section 3 when we investigate the pLASSO solution path with respect to η .

2.3.2 Logistic regression—Logistic regression fits each observation $(\mathbf{X}_i, Y_i)$ by a Bernoulli distribution with probability $p(\mathbf{X}_i) = P(Y_i = 1 | \mathbf{X}_i)$, where

$$p(\mathbf{X}_i) = \frac{\exp(\mathbf{Z}_i^T \boldsymbol{\beta})}{1 + \exp(\mathbf{Z}_i^T \boldsymbol{\beta})}. \quad (12)$$

The corresponding loss function is

$$L(\boldsymbol{\beta}; \mathbb{X}, \mathbf{Y}) = -\frac{1}{n} \sum_{i=1}^n [Y_i \log\{p(\mathbf{X}_i)\} + (1 - Y_i) \log\{1 - p(\mathbf{X}_i)\}], \quad (13)$$

which is obtained by setting $b(\theta) = \log[1 + \exp(\theta)]$ in (2). The prediction $\hat{Y}_i^{\text{P}}$ can be computed as $\hat{Y}_i^{\text{P}} = b'(\mathbf{Z}_i^T \hat{\boldsymbol{\beta}}^{\text{P}}) = \exp(\mathbf{Z}_i^T \hat{\boldsymbol{\beta}}^{\text{P}}) / [1 + \exp(\mathbf{Z}_i^T \hat{\boldsymbol{\beta}}^{\text{P}})]$. Following (7), the pLASSO criterion function for logistic regression is:

$$\begin{aligned} L_{\lambda, \eta}(\boldsymbol{\beta}; \mathbb{X}, \mathbf{Y}, \hat{\mathbf{Y}}^{\text{P}}) &= L(\boldsymbol{\beta}; \mathbb{X}, \mathbf{Y}) + \eta L(\boldsymbol{\beta}; \mathbb{X}, \hat{\mathbf{Y}}^{\text{P}}) + \lambda \sum_{j=1}^p |\beta_j| \\ &\propto -\frac{1}{n} \sum_{i=1}^n [\tilde{Y}_i \log\{p(\mathbf{X}_i)\} + (1 - \tilde{Y}_i) \log\{1 - p(\mathbf{X}_i)\}] + \frac{\lambda}{1 + \eta} \sum_{j=1}^p |\beta_j|, \end{aligned} \quad (14)$$

where L is defined in (13) and $\tilde{\mathbf{Y}} = (\tilde{Y}_1, \dots, \tilde{Y}_n)^T$.

Note that each term $Y_i \log\{p(\mathbf{X}_i)\} + (1 - Y_i) \log\{1 - p(\mathbf{X}_i)\}$ in $L(\boldsymbol{\beta}; \mathbb{X}, \mathbf{Y})$ is the KL distance between two Bernoulli distributions—Bernoulli(Y_i) and Bernoulli($p(\mathbf{X}_i)$). Each term $\hat{Y}_i^{\text{P}} \log\{p(\mathbf{X}_i)\} + (1 - \hat{Y}_i^{\text{P}}) \log\{1 - p(\mathbf{X}_i)\}$ in $L(\boldsymbol{\beta}; \mathbb{X}, \hat{\mathbf{Y}}^{\text{P}})$ is the KL distance between Bernoulli($\hat{Y}_i^{\text{P}}$) and Bernoulli($p(\mathbf{X}_i)$). Alternatively, we can view the pLASSO criterion function as a LASSO criterion function with observations $\{(\mathbf{X}_i, \tilde{Y}_i)\}_{i=1}^n$. Here, $\tilde{Y}_i \in [0, 1]$ can be regarded as the probability of $Y_i = 1$ obtained after adjusting the distribution Bernoulli(Y_i) using the prior information.

We note that the same KL distance has been used in Schapire et al. (2005) to incorporate prior information into the boosting classification algorithm. It is not surprising that it also works for logistic regression given the close connection between logistic regression and boosting as shown in Friedman et al. (2000). However, the goal and development of our method are obviously different from those in Schapire et al. (2005).

3 SOLUTION PATH

For linear regression, the geometric interpretation of the LASSO estimator provides a useful insight into the mechanism of LASSO, leading to the development of the LARS algorithm (Efron et al., 2004). The pLASSO solution would enjoy the same property of the LASSO solution regarding the change of λ for a given η . In this section, we show that the pLASSO estimator of (10) has a similar solution path with respect to the change of η for a given λ . This result allows us to fully characterize the local effect of incorporating the prior information into a LASSO problem.

Let $\mathbf{G}^{\lambda, \eta} = n^{-1}(1 + \eta) \mathbf{X}^T \mathbf{X} \hat{\boldsymbol{\beta}}^{\lambda, \eta} - n^{-1} \mathbf{X}^T (\mathbf{Y} + \eta \hat{\mathbf{Y}}^{\text{P}})$ be the gradient evaluated at $\boldsymbol{\beta} = \hat{\boldsymbol{\beta}}^{\lambda, \eta}$ of the pLASSO criterion function in (10) excluding the L_1 penalty. Similar to Efron et al. (2004),

$\hat{\beta}^{\lambda,\eta}$ and $\mathbf{G}^{\lambda,\eta}$ are continuous of both λ and η . Let $\mathcal{A}^{\lambda,\eta}$ be the current active set, i.e., $\mathcal{A}^{\lambda,\eta} = \{j: \hat{\beta}_j^{\lambda,\eta} \neq 0\}$. The points of discontinuity of $\mathcal{A}^{\lambda,\eta}$ are denoted by $\{\eta_1, \eta_2, \eta_3, \dots\}$ where $0 < \eta_1 < \dots < \eta_k < \eta_{k+1} < \dots < \infty$. Therefore, within the interval of $\eta \in (\eta_k, \eta_{k+1})$, $\mathcal{A}_k = \mathcal{A}^{\lambda,\eta}$ is a constant set. Finally, define $\mathbb{X}_{\mathcal{A}_k} = (\dots, \mathbf{x}_j, \dots)_{j \in \mathcal{A}_k}$ to be the submatrix of $\mathbb{X}$ consisting of the columns in $\mathcal{A}_k$, and similarly $\beta_{\mathcal{A}_k} = (\dots, \beta_j, \dots)_{j \in \mathcal{A}_k}^T$ to be the subvector of β consisting of the elements in $\mathcal{A}_k$.

The necessary and sufficient conditions for $\hat{\beta}^{\lambda,\eta}$ to be a pLASSO estimator are that

$$\begin{cases} \text{(C1): } G_j^{\lambda,\eta} = -\lambda \text{sign}(\hat{\beta}_j^{\lambda,\eta}), & j \in \mathcal{A}^{\lambda,\eta}, \\ \text{(C2): } |G_j^{\lambda,\eta}| \leq \lambda, & j \notin \mathcal{A}^{\lambda,\eta}. \end{cases} \quad (15)$$

From condition (C1) in (15), we have that

$$n^{-1}(1+\eta)\mathbb{X}_{\mathcal{A}_k}^T \mathbb{X}_{\mathcal{A}_k} \hat{\beta}_{\mathcal{A}_k}^{\lambda,\eta} - n^{-1}\mathbb{X}_{\mathcal{A}_k}^T (\mathbf{Y} + \eta \hat{\mathbf{Y}}^P) = -\lambda \text{sign}(\hat{\beta}_{\mathcal{A}_k}^{\lambda,\eta}), \eta \in (\eta_k, \eta_{k+1}). \quad (16)$$

Choose any η and η' such that $\eta_k < \eta' < \eta < \eta_{k+1}$. It follows that $\text{sign}(\hat{\beta}_{\mathcal{A}_k}^{\lambda,\eta}) = \text{sign}(\hat{\beta}_{\mathcal{A}_k}^{\lambda,\eta'})$ from the continuity of $\hat{\beta}^{\lambda,\eta}$ and the fact that $\mathcal{A}^{\lambda,\eta}$ does not change for $\eta \in (\eta_k, \eta_{k+1})$. Plugging η and η' into (16) and then subtracting, we can solve the linear equation and obtain

$$\hat{\beta}_{\mathcal{A}_k}^{\lambda,\eta} = \frac{1+\eta'}{1+\eta} \hat{\beta}_{\mathcal{A}_k}^{\lambda,\eta'} + \frac{\eta-\eta'}{1+\eta} (\mathbb{X}_{\mathcal{A}_k}^T \mathbb{X}_{\mathcal{A}_k})^{-1} \mathbb{X}_{\mathcal{A}_k}^T \hat{\mathbf{Y}}^P.$$

Let us define $\hat{\beta}^{\lambda,\eta}$ as the vector that satisfies $\hat{\beta}_{\mathcal{A}_k}^{\lambda,\eta} = (\mathbb{X}_{\mathcal{A}_k}^T \mathbb{X}_{\mathcal{A}_k})^{-1} \mathbb{X}_{\mathcal{A}_k}^T \hat{\mathbf{Y}}^P$ and $\hat{\beta}_{\mathcal{A}_k^c}^{\lambda,\eta} = 0$. In fact, $\hat{\beta}^{\lambda,\eta}$ is the restricted least squares estimator with the response $\hat{\mathbf{Y}}^P$ and the design matrix restricted to the columns $\mathcal{A}_k$. By the continuity of $\hat{\beta}^{\lambda,\eta}$, setting $\eta' \rightarrow \eta_k$ leads to

$$\hat{\beta}_{\mathcal{A}_k}^{\lambda,\eta} = \frac{1+\eta_k}{1+\eta} \hat{\beta}_{\mathcal{A}_k}^{\lambda,\eta_k} + \frac{\eta-\eta_k}{1+\eta} \hat{\beta}_{\mathcal{A}_k}^{\lambda,\eta_k,P}, \eta \in (\eta_k, \eta_{k+1}). \quad (17)$$

Also note that $\hat{\beta}_{\mathcal{A}_k^c}^{\lambda,\eta_k} = \hat{\beta}_{\mathcal{A}_k^c}^{\lambda,\eta_k,P} = 0$. We have proved the following theorem.

Theorem 1—For $\eta \in (\eta_k, \eta_{k+1})$, the pLASSO solution of (10) satisfies that

$$\hat{\beta}^{\lambda,\eta} = \frac{1+\eta_k}{1+\eta} \hat{\beta}^{\lambda,\eta_k} + \frac{\eta-\eta_k}{1+\eta} \hat{\beta}^{\lambda,\eta_k,P}.$$

Theorem 1 implies that the pLASSO solution follows a piecewise smooth path with respect to the change of η . A similar solution path can be found in the proof of Lemma 4 in Meinshausen and Yu (2009). Their solution path was obtained by moving the response variable along the direction to the noise, which modifies only the linear term of β in the loss function. In our case, the change to the loss function with respect to η is slightly more complicated than their situation because both linear and quadratic terms of β are involved. As a consequence, the path of pLASSO in Theorem 1 is smooth, but not linear as in Meinshausen and Yu (2009).

To understand the pLASSO solution path better, we need to derive how the active set evolves with respect to the change of η . In light of the conditions in (15), for any given λ and η , the elements of an estimator $\hat{\beta}^{\lambda,\eta}$ can be partitioned into the following three non-overlapping groups,

$$\mathcal{J}_1^{\lambda,\eta} = \{j: \hat{\beta}_j^{\lambda,\eta} \neq 0, G_j^{\lambda,\eta} = -\lambda \text{sign}(\hat{\beta}_j^{\lambda,\eta})\}, \quad (18)$$

$$\mathcal{J}_2^{\lambda,\eta} = \{j: \hat{\beta}_j^{\lambda,\eta} = 0, |G_j^{\lambda,\eta}| = \lambda\}, \quad (19)$$

$$\mathcal{J}_3^{\lambda,\eta} = \{j: \hat{\beta}_j^{\lambda,\eta} = 0, |G_j^{\lambda,\eta}| < \lambda\}. \quad (20)$$

Obviously, $\mathcal{J}_1^{\lambda,\eta} = \mathcal{A}^{\lambda,\eta}$. The following theorem fully characterizes the evolvement of the active set with the change of η .

Theorem 2—Assume the “one at a time” condition in Efron et al. (2004). Given $\lambda > 0$ and the current active set $\mathcal{A}_k$,

- a. $\eta_{k+1} = \eta$ and $\mathcal{A}_{k+1} = \mathcal{A}_k \setminus \{j\}$ if and only if j enters $\mathcal{J}_2^{\lambda,\eta}$ from $\mathcal{J}_1^{\lambda,\eta}$ at η ;
- b. $\eta_{k+1} = \eta$ and $\mathcal{A}_{k+1} = \mathcal{A}_k \cup \{j\}$ if and only if j enters $\mathcal{J}_2^{\lambda,\eta}$ from $\mathcal{J}_3^{\lambda,\eta}$ at η .

The active set $\mathcal{A}^{\lambda,\eta}$ can increase or decrease as η increases. “One at a time” means that the increases and decreases never involve more than a single index j . The proof of Theorem 2 can be found in the Appendix. Theorems 1 and 2 present the whole picture of the solution path of the pLASSO estimator. This piecewise smooth path is different from the piecewise linear path as in LARS or the similar work by Rosset and Zhu (2007). It is also different from the differential equation-based solution path in Wu (2011). This unique solution path deepens our understanding of how the prior information is incorporated into the original LASSO problem locally.

Figure 1 shows an example of what the pLASSO solution path looks like. It is noteworthy that this solution path is limited to the linear regression setting. Also importantly, it is known that LARS, which is again a solution path algorithm, is not as efficient as some other algorithms such as the coordinate descent algorithm (Wu and Lange, 2008; Friedman et al., 2010). Hence, our presentation of the pLASSO solution path is purely to gain insight into

how and why pLASSO works. In simulation and real data analysis, however, we use other algorithms that are more convenient.

4 THEORETICAL PROPERTIES

This section mainly deals with the theoretical properties of pLASSO. In Section 4.1, we compare the asymptotic distributions of the LASSO and pLASSO estimators when p is fixed. In Section 4.2, we study the weak oracle property of pLASSO when $p \rightarrow \infty$ and compare it with LASSO. In Section 4.3, we establish the oracle inequalities for excess risk and estimation error of pLASSO, and again compare them with existing results of LASSO.

4.1 Low-Dimensional Asymptotic Distribution

Assuming that the dimension p is fixed, we derive the asymptotic distribution of the pLASSO estimator and compare it with that of the LASSO estimator. The asymptotic distribution of the LASSO estimator is studied thoroughly by Knight and Fu (2000) in the linear regression setting. A similar result can be obtained in the framework of generalized linear models.

Without loss of generality, let $\beta_0 = (\beta_{1,0}^T, \beta_{2,0}^T)^T$ be the true values of the parameters β with nonzero parameters $\beta_{1,0} = (\beta_{0,0}, \beta_{1,0}, \dots, \beta_{s,0})^T$ and zero parameters $\beta_{2,0} = (\beta_{s+1,0}, \dots, \beta_{p,0})^T = \mathbf{0}$. Accordingly, write $\beta = (\beta_1^T, \beta_2^T)^T$, where $\beta_1 = (\beta_0, \beta_1, \dots, \beta_s)^T$ and $\beta_2 = (\beta_{s+1}, \dots, \beta_p)^T$. Define

$$\begin{aligned}\mu(\beta) &= [b'(\mathbf{Z}_1^T \beta), \dots, b'(\mathbf{Z}_n^T \beta)]^T, \\ \Sigma(\beta) &= \text{diag}[b''(\mathbf{Z}_1^T \beta), \dots, b''(\mathbf{Z}_n^T \beta)],\end{aligned}$$

and $\mathbf{F}(\beta) = \mathbf{Z}^T \Sigma(\beta) \mathbf{Z}$ with $\mathbf{Z} = (\mathbf{Z}_1, \dots, \mathbf{Z}_n)^T$. In the definition of $\mu(\beta)$ and $\Sigma(\beta)$, for linear regression (9), $b'(\theta) = \theta$ and $b''(\theta) \equiv 1$; for logistic regression (13), $b'(\theta) = \exp(\theta)/[1 + \exp(\theta)]$ and $b''(\theta) = \exp(\theta)/[1 + \exp(\theta)]^2$.

Theorem 3 provides the asymptotic distribution of the LASSO estimator $\hat{\beta}^\lambda$. This result is well known in the linear regression setting (Knight and Fu, 2000).

Theorem 3—Suppose $\sqrt{n}\lambda \rightarrow \lambda_0 \geq 0$, $\mathbf{F}(\beta_0)/n \rightarrow \Omega$ with a positive definite matrix Ω , and $\sup\{\|\mathbf{F}(\beta)/n - \mathbf{F}(\beta_0)/n\|_2: \sqrt{n}\|\beta - \beta_0\|_2 \leq \delta\} \rightarrow 0$ for any $\delta > 0$. Let $\mathbf{u} \sim N(0, \Omega)$, then the LASSO estimator $\hat{\beta}^\lambda$ follows that $\sqrt{n}(\hat{\beta}^\lambda - \beta_0) \rightarrow_d \arg \min(V)$, where

$$V(\phi) = Q(\phi) + P_{\lambda_0}(\phi). \quad (21)$$

In (21), Q and P_{λ_0} are defined as

$$Q(\phi) = \phi^T \Omega \phi / 2 - \mathbf{u}^T \phi, \\ P_{\lambda_0}(\phi) = \lambda_0 \sum_{j=1}^p [\phi_j \text{sign}(\beta_{j,0}) I(\beta_{j,0} \neq 0) + |\phi_j| I(\beta_{j,0} = 0)].$$

In parallel, we will show that a similar result to the above theorem exists for pLASSO. To proceed, we need to be slightly more specific about how the prior information is summarized. As in Section 2.2, we use a prior set $\mathcal{S}^p$ to indicate the set of those identified variables of potential importance. Then, a prior estimator $\hat{\beta}^p$ is obtained using (8), and the prior information is summarized into $\hat{Y}_i^p = b' (Z_i^T \hat{\beta}^p)$. Theorem 4 provides the asymptotic distribution of the pLASSO estimator $\hat{\beta}^{\lambda, \eta}$ when $n \rightarrow \infty$.

Theorem 4—Suppose $\sqrt{n}\lambda \rightarrow \lambda_0 \geq 0$, $\sqrt{n}\kappa \rightarrow \kappa_0 \geq 0$, $\mathbf{F}(\beta_0)/n \rightarrow \Omega$ with a positive definite matrix Ω , and $\sup\{\|\mathbf{F}(\beta)/n - \mathbf{F}(\beta_0)/n\|_2 : \sqrt{n}\|\beta - \beta_0\|_2 \leq \delta\} \rightarrow 0$ for any $\delta > 0$. Then, the pLASSO estimator $\hat{\beta}^{\lambda, \eta}$ follows that $\sqrt{n}(\hat{\beta}^{\lambda, \eta} - \beta_0) \rightarrow_d \arg \min(V)$, where

$$V(\phi) = (1 + \eta_0) \phi^T \Omega \phi / 2 - [\arg \min(Q) + \eta_0 \arg \min(Q + P_{\kappa_0, \mathcal{S}^p})]^T \Omega \phi + P_{\lambda_0}(\phi) \quad (22)$$

when $\eta \rightarrow \eta_0 = 0$, and

$$V(\phi) = Q(\phi) + P_{\kappa_0, \mathcal{S}^p}(\phi) \quad (23)$$

when $\eta \rightarrow \infty$. In (22) and (23), Q and P_{λ_0} are defined as in Theorem 3, and $P_{\kappa_0, \mathcal{S}^p}$ is defined as

$$P_{\kappa_0, \mathcal{S}^p}(\phi) = \kappa_0 \sum_{j=1}^p [\phi_j \text{sign}(\beta_{j,0}) I(\beta_{j,0} \neq 0) + |\phi_j| I(\beta_{j,0} = 0)] I(X_j \notin \mathcal{S}^p).$$

The influence of incorporating the prior information can be revealed by comparing the asymptotic distributions of the LASSO and pLASSO estimators. In the extreme case when $\eta \rightarrow \infty$, the prior information dominates, and it is not surprising to observe that the asymptotic distribution of the pLASSO estimator is the same as the prior estimator as in (23). In the other extreme case when $\eta \rightarrow 0$, the contribution of the prior information vanishes eventually so that the asymptotic distribution of the pLASSO estimator simplifies to that of the LASSO estimator.

The non-trivial situation arises when $\eta \rightarrow \eta_0 > 0$. For simplicity, let's assume $\lambda_0 = 0$, which implies that the L_1 penalization has a negligible effect asymptotically. This is a fair assumption as both (21) and (22) have the same term $P_{\lambda_0}(\phi)$, which leads to

$$\sqrt{n}(\hat{\beta}^{\lambda} - \beta_0) \rightarrow_d \arg \min(Q) = \Omega^{-1} \mathbf{u}, \quad (24)$$

$$\sqrt{n}(\hat{\beta}^{\lambda, \eta} - \beta_0) \rightarrow_d \frac{1}{1 + \eta_0} [\arg \min(Q) + \eta_0 \arg \min(Q + P_{\kappa_0, \mathcal{S}^p})]. \quad (25)$$

Compared with the asymptotic form $\arg \min(Q)$ of the LASSO estimator, the pLASSO estimator yields a sum of $\arg \min(Q)$ and $\arg \min(Q + P_{\kappa_0, \mathcal{S}^p})$ weighted by the tuning parameter η_0 . Certainly, whether $\arg \min(Q + P_{\kappa_0, \mathcal{S}^p})$ is an improvement or a degradation of $\arg \min(Q)$ depends on the quality of the prior set $\mathcal{S}^p$ as well as the extent (reflected by the parameter κ_0) to which we penalize the variables in $(\mathcal{S}^p)^c$ when constructing $\hat{\mathbf{Y}}^p$.

Writing $\Omega = \begin{bmatrix} \Omega_{11} & \Omega_{12} \\ \Omega_{21} & \Omega_{22} \end{bmatrix}$ by splitting the subscripts $\{0, 1, \dots, p\}$ into $\{0, 1, \dots, s\}$ and $\{s+1, \dots, p\}$, we use the following two extreme cases for illustration. On the one hand, suppose that we happen to have the best prior set $\mathcal{S}^p = \{X_1, \dots, X_s\}$ and we believe they are the only nonzero parameters by setting κ_0 large enough. It leads to that

$\arg \min(Q + P_{\kappa_0, \mathcal{S}^p}) = [(\Omega_{11}^{-1} \mathbf{u}_1)^T, \mathbf{0}^T]^T$ with $\mathbf{u} = (\mathbf{u}_1^T, \mathbf{u}_2^T)^T \sim N(\mathbf{0}, \Omega)$. On the other hand, suppose that we have the worst prior set $\mathcal{S}^p = \{X_{s+1}, \dots, X_p\}$. This yields $\arg \min(Q + P_{\kappa_0, \mathcal{S}^p}) = \Omega^{-1}[\mathbf{u} - \kappa_0 \text{sign}(\mathbf{b})]$ with $\mathbf{b} = (0, \beta_{1,0}, \dots, \beta_{s,0}, 0, \dots, 0)^T$ and $\text{sign}(0) = 0$. These are the best and worst cases when we incorporate the prior information, while it is common to have an intermediate case in practice.

In the above two cases, the asymptotic distribution in (25) is respectively simplified into

$$\frac{1}{1 + \eta_0} \left[\Omega^{-1} \mathbf{u} + \eta_0 \begin{pmatrix} \Omega_{11}^{-1} \mathbf{u}_1 \\ \mathbf{0} \end{pmatrix} \right] \sim N \left[\mathbf{0}, \frac{1}{(1 + \eta_0)^2} \Omega^{-1} \{ \Omega + (2\eta_0 + \eta_0^2) \tilde{\Omega} \} \Omega^{-1} \right], \quad (26)$$

$$\frac{1}{1 + \eta_0} [\Omega^{-1} \mathbf{u} + \eta_0 \Omega^{-1} \{ \mathbf{u} - \kappa_0 \text{sign}(\mathbf{b}) \}] \sim N \left[-\frac{\eta_0 \kappa_0}{1 + \eta_0} \Omega^{-1} \text{sign}(\mathbf{b}), \Omega^{-1} \right], \quad (27)$$

with $\tilde{\Omega} = \begin{bmatrix} \Omega_{11} & \Omega_{12} \\ \Omega_{21} & \Omega_{21} \Omega_{11}^{-1} \Omega_{12} \end{bmatrix}$. In the best case (26), $\Omega - \tilde{\Omega}$ is non-negative definite because $\Omega_{22} - \Omega_{21} \Omega_{11}^{-1} \Omega_{12}$ is non-negative definite. Thus, the pLASSO estimator has a smaller asymptotic variance than the LASSO estimator by comparing (26) and (24). Both estimators are asymptotically unbiased. In the worst case (27), the LASSO and pLASSO estimators have the same asymptotic variance Ω^{-1} but the pLASSO estimator becomes asymptotically biased. This result clearly illustrates the trade-off between variance reduction and increased bias depending on the quality of the prior information. It is also apparent from the asymptotic distributions in (26)–(27) that the theoretical choice of η depends on the quality of the prior information.

4.2 High-Dimensional Weak Oracle Property

Statistical properties of LASSO for high-dimensional data ($p \gg n$) have been studied by many works, including Zhao and Yu (2006) for variable selection consistency, Meinshausen

and Yu (2009) for estimation consistency, and Fan and Lv (2011) for weak oracle property. We focus on the weak oracle property of pLASSO in this subsection.

To provide a coherent comparison with the existing results, we adopt the same notation as in Fan and Lv (2011). Therefore, we do not include the intercept β_0 and set $\boldsymbol{\beta} = (\beta_1, \dots, \beta_p)^T$ as in Fan and Lv (2011). Slightly different from Section 4.1, $\boldsymbol{\beta}_1 = (\beta_1, \dots, \beta_s)^T$ and $\boldsymbol{\beta}_2 = (\beta_{s+1}, \dots, \beta_p)^T$. This setting results in the following LASSO and pLASSO criterion functions:

$$L_\lambda(\boldsymbol{\beta}; \mathbf{X}, \mathbf{Y}) = -\frac{1}{n} \sum_{i=1}^n [Y_i \mathbf{X}_i^T \boldsymbol{\beta} - b(\mathbf{X}_i^T \boldsymbol{\beta})] + \lambda \|\boldsymbol{\beta}\|_1, \quad (28)$$

$$L_{\lambda, \eta}(\boldsymbol{\beta}; \mathbf{X}, \mathbf{Y}, \hat{\mathbf{Y}}^p) = -\frac{1}{n} \sum_{i=1}^n [Y_i \mathbf{X}_i^T \boldsymbol{\beta} - b(\mathbf{X}_i^T \boldsymbol{\beta})] - \frac{\eta}{n} \sum_{i=1}^n [\hat{Y}_i^p \mathbf{X}_i^T \boldsymbol{\beta} - b(\mathbf{X}_i^T \boldsymbol{\beta})] + \lambda \|\boldsymbol{\beta}\|_1. \quad (29)$$

We review in the following the weak oracle property of LASSO in Fan and Lv (2011). Let $\mathbf{X}_1$ and $\mathbf{X}_2$ respectively be the submatrices of the design matrix $\mathbf{X} = (\mathbf{x}_1, \dots, \mathbf{x}_p)$ formed by columns in the nonzero positions of $\boldsymbol{\beta}_0$ and the complement. Without loss of generality we assume that each covariate $\mathbf{x}_j$ has been standardized so that $\|\mathbf{x}_j\|_2 = \sqrt{n}$. Let $d_n = 2^{-1} \min\{|\beta_{j0}| : \beta_{j0} \neq 0\}$. Similar to Section 4.1, define

$$\boldsymbol{\mu}(\boldsymbol{\beta}) = [b'(\mathbf{X}_1^T \boldsymbol{\beta}), \dots, b'(\mathbf{X}_n^T \boldsymbol{\beta})]^T, \\ \boldsymbol{\Sigma}(\boldsymbol{\beta}) = \text{diag}[b''(\mathbf{X}_1^T \boldsymbol{\beta}), \dots, b''(\mathbf{X}_n^T \boldsymbol{\beta})].$$

Condition A: The design matrix $\mathbf{X}$ satisfies that

$$\begin{aligned} \|\mathbf{X}_1^T \boldsymbol{\Sigma}(\boldsymbol{\beta}_0) \mathbf{X}_1\|^{-1} &= O(b_s n^{-1}), \\ \|\mathbf{X}_2^T \boldsymbol{\Sigma}(\boldsymbol{\beta}_0) \mathbf{X}_1\| \|\mathbf{X}_1^T \boldsymbol{\Sigma}(\boldsymbol{\beta}_0) \mathbf{X}_1\|^{-1} &\leq C \in (0, 1), \\ \max_{\boldsymbol{\beta}_1 \in \mathcal{A}_0} \max_{1 \leq j \leq p} \lambda_{\max} [\mathbf{X}_1^T \text{diag}\{|\mathbf{x}_j| \circ |\boldsymbol{\mu}''((\boldsymbol{\beta}_1^T, \mathbf{0}^T)^T)|\} \mathbf{X}_1] &= O(n), \end{aligned}$$

where the L_∞ norm of a matrix is the maximum of the L_1 norm of each row, b_s is a diverging sequence of positive numbers depending on s , $\mathbf{X} = \{\boldsymbol{\beta}_1 \in \mathbb{R}^s : \|\boldsymbol{\beta}_1 - \boldsymbol{\beta}_{1,0}\|_\infty \leq d_n\}$, and $\circ$ denotes the Hadamard product.

Condition B: Assume that $d_n = n^{-\gamma} \log n$ and $b_s = o\{\min(n^{1/2-\gamma} \sqrt{\log n}, s^{-1} n^\gamma / \log n)\}$, for some $\gamma \in (0, 1/2]$. In addition, assume that λ satisfies $\lambda = o(b_s^{-1} n^{-\gamma} \log n)$ and $\lambda \gg n^{-\alpha} (\log n)^2$ with $\alpha = \min(\frac{1}{2}, 2\gamma - \alpha_0)$, $s = O(n^{\alpha_0})$, and $\alpha_0 < 1$, and that $\max_{1 \leq j \leq p} \|\mathbf{x}_j\|_\infty = o(n^\alpha / \sqrt{\log n})$ if the responses are unbounded.

Condition C: With a constant $c_1 > 0$, $P(\|\mathbf{a}^T \mathbf{Y} - \mathbf{a}^T \boldsymbol{\mu}(\boldsymbol{\beta}_0)\| > \|\mathbf{a}\|_2 \varepsilon) \leq 2e^{-c_1 \varepsilon^2}$ where $\varepsilon \in (0, \infty)$ if the responses are bounded and $\varepsilon \in (0, \|\mathbf{a}\|_2 / \|\mathbf{a}\|_\infty]$ if the responses are unbounded.

Conditions A–C are rewritten from Conditions 1–3 and the exponential bound (22) in Fan and Lv (2011), specifically for the LASSO situation. The following theorem is rewritten from Theorem 2 in that paper for the LASSO estimator.

Theorem 5—Assume that Conditions A–C are satisfied and $\log p = O(n^{1-2\alpha})$. Then, there exists a LASSO estimator $\hat{\beta}^\lambda$ such that with probability tending to one,

- a. $\hat{\beta}_2^\lambda = 0$;
- b. $\|\hat{\beta}_1^\lambda - \beta_{1,0}\|_\infty \leq n^{-\gamma} \log n$, with γ defined in Condition B.

In comparison, we present a parallel result of the pLASSO estimator. Similar to the previous subsection, we assume that the prior information in pLASSO is captured by a prior estimator $\hat{\beta}^p$ so that $\hat{Y}_i^p = b'(\mathbf{X}_i^T \hat{\beta}^p)$. We show that the weak oracle property can be achieved by imposing some extra conditions on the prior estimator (information).

Condition D: The prior estimator $\hat{\beta}^p$ satisfies that

$$\begin{aligned} \|\hat{\beta}_1^p - \beta_{1,0}\|_\infty &= o_p(\eta^{-1} n^{-\gamma} \log n), \\ \|\mathbb{X}^T \sum (\beta_0) \mathbb{X}_1 (\hat{\beta}_1^p - \beta_{1,0})\|_\infty &= o_p(n \lambda \eta^{-1}), \\ \|\mathbb{X}^T \sum (\beta_0) \mathbb{X}_2 \hat{\beta}_2^p\|_\infty &= o_p(n \lambda \eta^{-1}), \\ \|\mathbb{X}_1^T \sum (\beta_0) \mathbb{X}_2 \hat{\beta}_2^p\|_\infty &= o_p(b_s^{-1} \eta^{-1} n^{1-\gamma} \log n), \\ \|\hat{\beta}^p - \beta_0\|_2^2 &= o_p[\min(\lambda \eta^{-1}, b_s^{-1} \eta^{-1} n^{-\gamma} \log n)], \end{aligned}$$

and that

$$\max_{\beta \in \mathcal{N}_0^p} \max_{1 \leq j \leq p} \lambda_{\max} [\mathbb{X}^T \text{diag}\{|\mathbf{x}_j| \circ |\boldsymbol{\mu}''(\beta)|\} \mathbb{X}] = O_p(n),$$

where $\mathcal{N}_0^p$ is the line segment between β_0 and $\hat{\beta}^p$.

Theorem 6—Assume that Conditions A–D are satisfied and $\log p = O(n^{1-2\alpha})$. Then, there exists a pLASSO estimator $\hat{\beta}^{\lambda, \eta}$ such that with probability tending to one,

- a. $\hat{\beta}_2^{\lambda, \eta} = 0$;
- b. $\|\hat{\beta}_1^{\lambda, \eta} - \beta_{1,0}\|_\infty \leq n^{-\gamma} \log n / (1 + \eta)$, with γ defined in Condition B.

It is obvious that η plays an important role in Condition D. All the conditions involved with η in Condition D have the same explanation: how much we should believe in the prior information (measured by η) depends on how well the prior estimator $\hat{\beta}^p$ performs. When $\eta = 0$, those conditions are automatically satisfied and the conclusion of Theorem 6 becomes identical to that of Theorem 5.

Theorem 6 implies that, if the prior information is reasonably reliable, the pLASSO estimator can achieve the same property of variable selection consistency as the LASSO estimator. In addition, the pLASSO possesses a lower L_∞ -norm loss for the nonzero coefficients compared to the LASSO estimator, similar to what we concluded in Section 4.1 when the dimension p is fixed and low. Both theoretical conclusions show the improvement of efficiency of pLASSO over LASSO with reliable prior information.

4.3 Oracle Inequalities

In addition to statistical properties on variable selection and asymptotic distribution, a great deal of attention has been devoted to the oracle inequalities for the excess risk (prediction error) and/or estimation error of the LASSO estimator. In linear models, these results include Bunea et al. (2007) for prediction error of LASSO, Bickel et al. (2009) for prediction error of both LASSO and Dantzig selector as well as bounds on estimation error, and Koltchinskii et al. (2011) for a sharp oracle inequality of prediction error. Some parallel results have been developed for generalized linear models, including but not limited to, Van de Geer (2008), Bach (2010) and Kwemou (2012). In this subsection, we establish the oracle inequalities for pLASSO in the framework of generalized linear models, including its excess risk and estimation error.

We utilize the same notation as in Section 4.2. Specifically, the intercept β_0 is omitted, β_0 denotes the true values of the parameters β , and the pLASSO estimator $\hat{\beta}^{\wedge, \eta}$ is defined to be the minimizer of the criterion function (29). $\mathcal{J}(\beta) = \{j \in \{1, \dots, p\} : \beta_j \neq 0\}$ denotes the subscript set of nonzero components of β and $|\mathcal{J}(\beta)|$ denotes the cardinality of $\mathcal{J}(\beta)$. Thus, from Section 4.2, $\mathcal{J}(\beta_0) = \{1, \dots, s\}$ and $|\mathcal{J}(\beta_0)| = s$. In addition, we introduce the following notation:

$$\varepsilon = (\varepsilon_1, \dots, \varepsilon_n)^T = \mathbf{Y} - \boldsymbol{\mu}(\beta_0) \quad \text{and} \quad \hat{\varepsilon}^p = (\hat{\varepsilon}_1^p, \dots, \hat{\varepsilon}_n^p)^T = \hat{\mathbf{Y}}^p - \boldsymbol{\mu}(\beta_0), \quad (30)$$

where $\boldsymbol{\mu}(\beta_0) = [b'(\mathbf{X}_1^T \beta_0), \dots, b'(\mathbf{X}_n^T \beta_0)]^T$ as in Sections 4.1–4.2.

Following Van de Geer (2008), the excess risk of an estimator $\hat{\beta}$ in a generalized linear model is defined as

$$\mathcal{E}(\hat{\beta}) = R(\hat{\beta}) - R(\beta_0) = L[\hat{\beta}; \mathbb{X}, \boldsymbol{\mu}(\beta_0)] - L[\beta_0; \mathbb{X}, \boldsymbol{\mu}(\beta_0)]. \quad (31)$$

Before presenting the main result, let us review the restricted eigenvalue condition in Bickel et al. (2009) as follows.

Condition RE(r, c_0): For some integer r such that $1 \leq r \leq p$ and a positive number c_0 , the following condition holds:

$$\kappa(r, c_0) = \min_{\substack{J \subseteq \{1, \dots, p\}, \\ |J| \leq r}} \min_{\substack{\boldsymbol{\delta} \in \mathbb{R}^p, \boldsymbol{\delta} \neq 0, \\ \|\boldsymbol{\delta}_{J^c}\|_1 \leq c_0 \|\boldsymbol{\delta}_J\|_1}} \frac{\|\mathbb{X}\boldsymbol{\delta}\|_2}{\sqrt{n} \|\boldsymbol{\delta}_J\|_2} > 0.$$

Moreover, we impose two other regularity conditions.

Condition E: Assume that $\sup_{n \rightarrow \infty} \|\mathbf{X}_i\|_\infty < \infty$. Moreover, there exists a constant $M > 0$ such that $\sup_{n \rightarrow \infty} E(\varepsilon_i^m) \leq M^m$ for $m = 2, 3, 4, \dots$

Condition F: The parameter space under consideration is restricted to

$$\mathbb{B}_\delta = \{\boldsymbol{\beta} \in \mathbb{R}^p : \min_{1 \leq i \leq n} \min_{0 \leq t \leq 1} b''(\mathbf{X}_i^T[(1-t)\boldsymbol{\beta}_0 + t\boldsymbol{\beta}]) \geq 2\delta\} \text{ for a constant } \delta > 0.$$

The first part of Condition E was employed by Kwemou (2012) when they derived oracle inequalities for LASSO. The second part of Condition E, justified by Jiang and Zhang (2013), is used here to control the probability with which the oracle inequalities hold, by applying Bernstein's inequality [e.g., Lemma 2.2.11 in van der Vaart and Wellner (1996)]. It is noteworthy that this condition is automatically satisfied in logistic regression. Condition F essentially imposes a uniform lower bound on the function $b''(\cdot)$, and it depends on the function b . For linear regression, the lower bound is automatically satisfied as $b''(\cdot) \equiv 1$ so $\mathbb{B}_\delta = \mathbb{R}^p$ when we choose any $\delta > 0$. For logistic regression, $b''(\theta) = \exp(\theta)/[1 + \exp(\theta)]^2$. Intuitively, Condition F enforces that the probabilities of the Bernoulli variables are uniformly bounded away from 0 and 1. Rigorously speaking, Condition F requires that $\delta < 1/8$, and $\mathbf{X}_i^T \boldsymbol{\beta}_0$ and $\mathbf{X}_i^T \boldsymbol{\beta}$ are uniformly bounded by the interval $[\log\{(1 - \sqrt{1-8\delta})/(4\delta) - 1\}, \log\{(1 + \sqrt{1-8\delta})/(4\delta) - 1\}]$ for $1 \leq i \leq n$. When $\delta \rightarrow 0$, this interval tends to $\mathbb{R}$ so $\mathbb{B}_\delta \rightarrow \mathbb{R}^p$. For Poisson regression, $b''(\theta) = \exp(\theta)$, and Condition F is satisfied as long as $\mathbf{X}_i^T \boldsymbol{\beta}_0 \geq \log(2\delta)$ and $\mathbf{X}_i^T \boldsymbol{\beta} \geq \log(2\delta)$ for $1 \leq i \leq n$. Again, when $\delta \rightarrow 0$, $\log(2\delta) \rightarrow -\infty$ and thus $\mathbb{B}_\delta \rightarrow \mathbb{R}^p$.

Theorem 7—Fix a constant $\phi > 0$ and an integer $1 \leq r \leq p$. Let Condition RE($r, 3+4/\phi$) be satisfied. Assume that $\log p = \alpha(n)$. Suppose $b(\cdot)$ in (29) is twice continuously differentiable with $b''(\cdot) > 0$. Let $\lambda = An^{-1/2}(\sqrt{\log p} + \eta B_{\hat{\boldsymbol{\beta}}^p})$ for a large enough constant $A > 0$ and $B_{\hat{\boldsymbol{\beta}}^p} = n^{-1/2} \|\mathbf{X}^T \boldsymbol{\varepsilon}^p\|_\infty$. Under Conditions E and F, with probability tending to one, the pLASSO estimator $\hat{\boldsymbol{\beta}}^{\lambda, \eta}$ satisfies

$$\mathcal{E}(\hat{\boldsymbol{\beta}}^{\lambda, \eta}) \leq (1+\phi) \inf_{\substack{\boldsymbol{\beta} \in \mathbb{B}_\delta \\ J(\boldsymbol{\beta}) \leq r}} \left\{ \mathcal{E}(\boldsymbol{\beta}) + \frac{2C(\phi)A^2}{\delta \kappa^2(r, 3+4/\phi)} |J(\boldsymbol{\beta})| \left[\frac{1}{(1+\eta^2)} \frac{\log p}{n} + \frac{\eta^2}{(1+\eta)^2} \frac{B_{\hat{\boldsymbol{\beta}}^p}^2}{n} \right] \right\}, \quad (32)$$

with δ defined in Condition F and $C(\phi) = (2 + \phi)^2 / [\phi(1 + \phi)]$.

Theorem 7 provides a general oracle inequality for the excess risk of the pLASSO estimator. This result is parallel to those for LASSO, such as Theorem 2.1 in Van de Geer (2008). Compared to the well known oracle rate $r_{\text{LASSO}} = |J(\boldsymbol{\beta})| \log p/n$ for the LASSO estimator (Van de Geer, 2008; Bickel et al., 2009), the rate of pLASSO's excess risk has two components (by omitting the foregoing constants):

$$r_{\text{pLASSO},1} = \frac{1}{(1+\eta)^2} \frac{|J(\boldsymbol{\beta})| \log p}{n} \quad \text{and} \quad r_{\text{pLASSO},2} = \frac{\eta^2}{(1+\eta)^2} \frac{|J(\boldsymbol{\beta})| B_{\hat{\boldsymbol{\beta}}^p}^2}{n}.$$

$r_{\text{pLASSO},1}$ is a fraction of r_{LASSO} , and $r_{\text{pLASSO},2}$ depends on the prediction error of the prior estimator $\hat{\boldsymbol{\beta}}^{\text{p}} = \hat{\mathbf{Y}}^{\text{p}} - \boldsymbol{\mu}(\boldsymbol{\beta}_0)$.

Clearly, the result illustrates the trade-off between the variance reduction by incorporating the correct prior information and the additive bias due to the prior error, exactly in the same spirit as in Sections 4.1–4.2. On the one hand, let's consider the situation when the prior information has a high quality. Take linear regression as an example where $\varepsilon_1, \dots, \varepsilon_n$ independently follow an identical normal distribution. Suppose we know the support of $\boldsymbol{\beta}_0$,

then the best prior estimator is the oracle estimator $\hat{\boldsymbol{\beta}}^{\text{p}} = [(\hat{\boldsymbol{\beta}}_1^{\text{p}})^T, \mathbf{0}^T]^T$ with

$\hat{\boldsymbol{\beta}}_1^{\text{p}} = (\mathbb{X}_1^T \mathbb{X}_1)^{-1} \mathbb{X}_1^T \mathbf{Y}$. In this case, $\hat{\boldsymbol{\varepsilon}}^{\text{p}} = \mathbb{X}_1 (\mathbb{X}_1^T \mathbb{X}_1)^{-1} \mathbb{X}_1^T \boldsymbol{\varepsilon}$, so we have

$$B_{\hat{\boldsymbol{\varepsilon}}^{\text{p}}} = n^{-1/2} \|\mathbb{X}_1^T \mathbb{X}_1 (\mathbb{X}_1^T \mathbb{X}_1)^{-1} \mathbb{X}_1^T \boldsymbol{\varepsilon}\|_{\infty} = n^{-1/2} \max(\|\mathbb{X}_1^T \boldsymbol{\varepsilon}\|_{\infty}, \|\mathbb{X}_2^T \mathbb{X}_1 (\mathbb{X}_1^T \mathbb{X}_1)^{-1} \mathbb{X}_1^T \boldsymbol{\varepsilon}\|_{\infty}).$$

Firstly, the Bernstein's inequality [Lemma 2.2.11 in van der Vaart and Wellner (1996)] leads to $n^{-1/2} \|\mathbb{X}_1^T \boldsymbol{\varepsilon}\|_{\infty} = O_p(\sqrt{\log s})$. Secondly, noticing that $\mathbf{a}^T \boldsymbol{\varepsilon}$ follows a normal distribution for any vector $\mathbf{a}$ satisfying $\|\mathbf{a}\|_2 = 1$, it is easy to verify that

$$n^{-1/2} \|\mathbb{X}_2^T \mathbb{X}_1 (\mathbb{X}_1^T \mathbb{X}_1)^{-1} \mathbb{X}_1^T \boldsymbol{\varepsilon}\|_{\infty} = n^{-1/2} \max_{j \notin J(\boldsymbol{\beta}_0)} \|\boldsymbol{\Pi}_1 \mathbf{x}_j\|_2 O_p(\sqrt{\log(p-s)}),$$

where $\boldsymbol{\Pi}_1 = \mathbb{X}_1 (\mathbb{X}_1^T \mathbb{X}_1)^{-1} \mathbb{X}_1^T$. Therefore, $B_{\hat{\boldsymbol{\varepsilon}}^{\text{p}}}^2$ can have a much smaller order than $\log p$ with a large probability [depending on the maximum L_2 norm of the projections of $\mathbf{x}_j, j \notin J(\boldsymbol{\beta}_0)$, on the space spanned by $\mathbb{X}_1$], then the pLASSO excess risk rate is only a fraction of the LASSO rate. On the other hand, if the prior information is not reliable, a small choice of η will control the additional excess risk caused by the wrong information.

Following Theorem 7, Corollary 1 provides some specific oracle inequalities for excess risk and estimation bias for the pLASSO estimator. Except the constants in the bounds, the rates are essentially the same as in (32). Therefore, Corollary 1 provides the same insight for us to better understand the pLASSO method.

Corollary 1: Assume the same conditions as in Theorem 7, except that Condition RE($r, 3 + 4/\phi$) is replaced by Condition RE($s, 3$). Then, with probability tending to one, the pLASSO estimator $\hat{\boldsymbol{\beta}}^{\lambda, \eta}$ satisfies

$$\begin{aligned} \mathcal{E}(\hat{\boldsymbol{\beta}}^{\lambda, \eta}) &\leq \frac{8A^2}{\delta \kappa^2(s, 3)} s \left[\frac{1}{(1+\eta)^2} \frac{\log p}{n} + \frac{\eta^2}{(1+\eta)^2} \frac{B_{\hat{\boldsymbol{\varepsilon}}^{\text{p}}}^2}{n} \right], \\ \frac{1}{n} \|\mathbb{X} \hat{\boldsymbol{\beta}}^{\lambda, \eta} - \mathbb{X} \boldsymbol{\beta}_0\|_2^2 &\leq \frac{8A^2}{\delta^2 \kappa^2(s, 3)} s \left[\frac{1}{(1+\eta)^2} \frac{\log p}{n} + \frac{\eta^2}{(1+\eta)^2} \frac{B_{\hat{\boldsymbol{\varepsilon}}^{\text{p}}}^2}{n} \right], \\ \|\hat{\boldsymbol{\beta}}^{\lambda, \eta} - \boldsymbol{\beta}_0\|_1^2 &\leq \frac{128A^2}{\delta^2 \kappa^4(s, 3)} s^2 \left[\frac{1}{(1+\eta)^2} \frac{\log p}{n} + \frac{\eta^2}{(1+\eta)^2} \frac{B_{\hat{\boldsymbol{\varepsilon}}^{\text{p}}}^2}{n} \right]. \end{aligned}$$

5 SIMULATION

5.1 Unweighted Penalization Methods

In this subsection, we perform simulation studies to compare the following three estimators: the LASSO estimator $\hat{\beta}^L$ without using any prior information as in (3), the prior estimator $\hat{\beta}^P$ which fully trusts the prior information, and our proposed pLASSO estimator $\hat{\beta}^{L,\eta}$ which incorporates the prior information into LASSO as in (6). All three methods use L_1 -norm penalty, so we refer to them as unweighted penalization methods in contrast to those in Section 5.2.

As in Section 2.3, we utilize both linear regression and logistic regression in our simulations. For linear regression, the training data are generated from the following model:

$$\mathbf{X} \sim N(\mathbf{0}, \Sigma), \quad Y \sim N(\mathbf{X}^T \beta_0, \sigma^2), \quad (33)$$

where $\Sigma = (\sigma_{ij})$ with $\sigma_{ij} = \rho^{|i-j|}$, $i, j = 1, \dots, p$. In (33), we fix $\rho = 0.5$, $\sigma = 1$, and $\beta_0 = (-0.5, -2, 0.5, 2, -1.5, 1, 2, -1.5, 2, -2, 1, 1.5, -2, 1, 1.5, 0, \dots, 0)^T$, with $X_1, \dots, X_{15}$ relevant to the response.

With the same β_0 and Σ , the training data are generated from the following model for logistic regression:

$$\mathbf{X} \sim N(\mathbf{0}, \Sigma), \quad Y \sim \text{Bernoulli}[\text{logit}^{-1}(\mathbf{X}^T \beta_0)]. \quad (34)$$

It is noteworthy that the above design matrix $\mathbf{X}$ has the so-called “power decay correlation.” It has been shown in Corollary 3 of Zhao and Yu (2006) that $\mathbf{X}$ satisfies the strong irrepresentable condition in linear regression, which is our main condition for the weak oracle property (the second condition of Condition A in Section 4.2). Moreover, $\mathbf{X}$ also satisfies the restricted eigenvalue condition in Bickel et al. (2009) with a large probability, which is our main condition for the oracle inequalities [Condition RE(r, c_0) in Section 4.3]. For details, we refer to Example 1 in Raskutti et al. (2010).

In the above models, we set $n = 200$, $p = 1000$ in (33), and $n = 200$, $p = 400$ in (34). We adopt the form of prior information introduced in Section 2.2. A subset $\mathcal{S}^p$ of explanatory variables is provided and the criterion function in (8) is used to estimate $\hat{\beta}^P$. In our computation, twelve different sets $\mathcal{S}^p$ are used for comparison, classified in four groups:

$$\begin{aligned} \mathcal{G}_1: \mathcal{S}_1^p &= \{X_1, \dots, X_5\}, \mathcal{S}_2^p = \{X_1, \dots, X_{10}\}, \mathcal{S}_3^p = \{X_1, \dots, X_{15}\}; \\ \mathcal{G}_2: \mathcal{S}_4^p &= \{X_{16}, \dots, X_{20}\}, \mathcal{S}_5^p = \{X_{16}, \dots, X_{40}\}, \mathcal{S}_6^p = \{X_{16}, \dots, X_{60}\}; \\ \mathcal{G}_3: \mathcal{S}_7^p &= \mathcal{S}_2^p \cup \{X_{16}, \dots, X_{18}\}, \mathcal{S}_8^p = \mathcal{S}_2^p \cup \{X_{16}, \dots, X_{20}\}, \mathcal{S}_9^p = \mathcal{S}_2^p \cup \{X_{16}, \dots, X_{22}\}; \\ \mathcal{G}_4: \mathcal{S}_{10}^p &= \{X_1, \dots, X_3\} \cup \mathcal{S}_6^p, \mathcal{S}_{11}^p = \{X_1, \dots, X_5\} \cup \mathcal{S}_6^p, \mathcal{S}_{12}^p = \{X_1, \dots, X_7\} \cup \mathcal{S}_6^p. \end{aligned}$$

The sets $\mathcal{S}^p$ in the first and second groups ($\mathcal{G}_1$ and $\mathcal{G}_2$) respectively consist of all relevant and irrelevant predictors only. The sets in the third and fourth groups ($\mathcal{G}_3$ and $\mathcal{G}_4$) are mixed with relevant and irrelevant variables. The relevant predictors dominate the sets in the third group $\mathcal{G}_3$ while the irrelevant predictors dominate the sets in the fourth group $\mathcal{G}_4$.

We use the coordinate descent algorithm developed by Friedman et al. (2010) and the algorithm from Meier et al. (2008) for our linear regression setting and logistic regression setting, respectively. After an estimator is obtained, it is further refitted using only those selected predictors to improve the estimation/prediction accuracy. The tuning parameters λ and η are selected by a 3-fold cross validation to maximize the average empirical likelihood. Specifically, for the selection of λ , we adopt the “one standard error rule” (Breiman et al., 1984) to achieve a more parsimonious model. That is, the parameter λ is selected to be the largest possible which gives an average empirical likelihood within one standard error from the highest average empirical likelihood. The standard error is estimated from the three folds in cross validation. With an independently simulated test set of the same size as the training set, the following criteria are used to assess the performance of an estimator in the setting of linear regression:

- a. #CNZ, #INZ: the number of variables that are correctly and incorrectly selected, respectively;
- b. Bias: the L_1 norm bias of an estimator for the nonzero coefficients;
- c. MSR: the mean squared residuals of an estimator evaluated on the test data.

In addition to (a) and (b), the following criteria are used to assess the performance of an estimator in the setting of logistic regression:

- d. ME: the model error defined as $E_{\mathbf{X}}[\{\text{logit}^{-1}(\mathbf{X}^T \hat{\boldsymbol{\beta}}) - \text{logit}^{-1}(\mathbf{X}^T \boldsymbol{\beta}_0)\}^2]$. RME: the relative model error as the ratio of the model error to that of the LASSO estimator. In our simulation, the model error is approximated by the empirical average using the test set;
- e. MR: the misclassification rate of the classifier induced by an estimator, evaluated on the test data.

Therefore, criteria (a)–(c) apply to linear regression, while criteria (a), (b), (d) and (e) apply to logistic regression. In summary, the criteria in (a) consider the variable selection property of an estimator; criteria (b) and (d) take into account its estimation property; and criteria (c) and (e) evaluate its prediction performance on the test data. For each choice of $\mathcal{S}^p$, 100 training/test data sets are generated from each model. Figures 2–3 provide the boxplots of the criteria (a)–(c) for linear regression from the 100 runs; Figures 4–5 provide the boxplots of the criteria (a), (b), (d) and (e) for logistic regression. In the figures, we denote the three methods by LASSO, p, and pLASSO, respectively. In addition, we report the optimal selection of η by pLASSO from the 100 runs in Figure 6.

In the case where $\mathcal{S}^p$ includes only the relevant variables (group $\mathcal{G}_1$), the prior estimator performs the best among the three methods: it selects the most variables correctly and the fewest variables incorrectly; it also yields the lowest estimation bias (and/or smallest model error) and the smallest mean squared residuals (or misclassification rate). Moreover, pLASSO behaves better than LASSO in most criteria we have examined. However, pLASSO includes a similar (or even a larger) number of irrelevant variables to LASSO. For more discussion about this, we refer to Section 5.2 and the Discussion section.

In the other case where S^p includes only irrelevant variables (group $\mathcal{G}_2$), the prior estimator performs the worst, especially since it selects all those wrong variables in the prior set. pLASSO and LASSO are also better in identifying relevant variables. In terms of estimation bias and prediction accuracy, pLASSO performs comparably with LASSO, with a clear advantage over the prior estimator.

When the prior set includes a mixture of relevant variables and irrelevant variables (groups $\mathcal{G}_3$ and $\mathcal{G}_4$), the numerical performance depends on the “quality” of the prior set. On the one hand, benefitted from the higher-quality prior information in $\mathcal{G}_3$, the prior estimator and pLASSO reduce the estimation bias (and/or model error) and yield a better prediction on the test data. They also correctly include more relevant variables in the model. On the other hand, the lower-quality prior information in $\mathcal{G}_4$ forces the prior estimator to select all irrelevant variables in the prior set, not as appealing as LASSO or pLASSO. LASSO and pLASSO perform similarly with a lower estimation bias and more accurate prediction than the prior estimator.

In terms of the tuning parameter η , it is clearly seen from Figure 6 that the optimal selection of η conform with our intuition. In group $\mathcal{G}_1$, the tuning parameter η is usually selected at the highest value we set; while the optimal η is much smaller in group $\mathcal{G}_2$. For groups that have both relevant and irrelevant variables, η is selected at higher values in $\mathcal{G}_3$ than in $\mathcal{G}_4$, since there are more relevant variables in $\mathcal{G}_3$ and more irrelevant variables in $\mathcal{G}_4$.

In summary, the prior and LASSO estimators are the preferred choices when the prior information is of high and low quality, respectively. However, they are not very robust and can perform poorly in some other situations. The pLASSO method is a more robust choice regardless of the quality of the prior information.

5.2 Weighted Penalization Methods

Now, we include the adaptive LASSO method (Zou, 2006) in our simulation studies. Adaptive LASSO is a weighted penalization method which uses an initial estimator $\hat{\beta}^{\text{initial}}$ to weight the L_1 penalties. In our context, the adaptive LASSO estimator minimizes the following criterion function:

$$-\frac{1}{n} \sum_{i=1}^n [Y_i \mathbf{Z}_i^T \boldsymbol{\beta} - b(\mathbf{Z}_i^T \boldsymbol{\beta})] + \lambda \sum_{j=1}^p \frac{|\beta_j|}{1 + \tau |\beta_j^{\text{initial}}|}, \quad (35)$$

where λ and τ are tuning parameters. A referee suggested we investigate the case where the initial estimator is the prior estimator $\hat{\beta}^p$.

For a more comprehensive and fair comparison, we also include the cases where the initial estimator is chosen to be the LASSO and pLASSO estimator. Altogether, we have three weighted penalization methods to compare, which respectively use $\hat{\beta}^L$, $\hat{\beta}^p$ and $\hat{\beta}^{L,\eta}$ as initial estimators. For convenience, we refer to them as LASSO-A, p-A and pLASSO-A.

The simulation settings and computational algorithms utilized in this subsection are almost the same as in Section 5.1. Thus, we only point out the differences here. The tuning

parameters (including λ and τ) are still optimized using 3-fold cross validation. The relative model error (RME) of the p-A/pLASSO-A estimator is defined to be the ratio of the model error to that of the LASSO-A estimator. Except this minor difference, the criteria used to compare among methods are exactly the same as those in Section 5.1. Figures 7–8 provide the boxplots of the criteria (a) – (c) for linear regression from the 100 runs; Figures 9–10 provide the boxplots of the criteria (a), (b), (d) and (e) for logistic regression.

Overall, the weighted penalization methods perform better than the unweighted penalization methods in the corresponding settings. The weighted L_1 penalty improves all unweighted penalization methods in terms of variable selection, as well as estimation and prediction. It is noteworthy that pLASSO-A performs much better than pLASSO for eliminating irrelevant variables. As follows, we focus on the comparison within three weighted penalization methods. Their relative performance follows similar trends to their unweighted counterparts, although the differences are less obvious.

When $\mathcal{S}^p$ includes only the relevant variables (group $\mathcal{G}_1$), p-A and pLASSO-A outperform LASSO-A in all criteria including variable selection, estimation and prediction. The difference gets larger when the prior set size gets larger.

When $\mathcal{S}^p$ includes only the irrelevant variables (group $\mathcal{G}_2$), p-A is not forced to include all the irrelevant variables in $\mathcal{S}^p$. However, its performance can still be affected by the wrong information, especially when there are many irrelevant variables in $\mathcal{S}^p$ (e.g., rows 5–6 in Figure 7). In linear regression, LASSO-A and pLASSO-A exclude more irrelevant variables than p-A (e.g., rows 5–6 in Figure 7); while in logistic regression, the difference is not obvious (rows 4–6 in Figure 9). pLASSO-A also identifies more relevant variables than p-A. Moreover, LASSO-A and pLASSO-A outperform p-A in terms of the precision of the estimation and prediction.

When $\mathcal{S}^p$ includes a mixture of relevant variables and irrelevant variables (groups $\mathcal{G}_3$ and $\mathcal{G}_4$), the numerical performance depends again on the quality of the prior set. The results from prior set groups $\mathcal{G}_3$ and $\mathcal{G}_4$ are similar to those from groups $\mathcal{G}_1$ and $\mathcal{G}_2$, respectively. This is consistent to our observation in Section 5.1 for unweighted penalization methods.

6 REAL DATA ANALYSIS

In this section, we re-analyze the data from a study of bipolar disorder to illustrate how to carry out a real data analysis with pLASSO. Bipolar disorder is a serious and potentially life-threatening mood disorder (Merikangas et al., 2007). It is known that bipolar disorder has a substantial genetic component (Craddock and Forty, 2006). However, the underlying genetic mechanism of bipolar disorder remains elusive, despite significant research effort. There have been at least six genome-wide association studies reported in the literature so far (WTCCC, 2007; Baum et al., 2007; Ferreira et al., 2008; Scott et al., 2009; Sklar et al., 2008; Smith et al., 2009). While these studies revealed promising association signals, top findings from various studies do not show obvious overlap (Baum et al., 2008; Ollila et al., 2009).

In this work, we use the data collected in the Wellcome Trust Case Control Consortium (WTCCC) study as the primary data set. Then we choose the top findings from the other studies as prior information. Among the five prior studies, both Ferreira et al. (2008) and Scott et al. (2009) used meta-analysis by jointly analyzing an independently collected data set and the WTCCC data. Another study, Smith et al. (2009), focused on analyzing the data from two independent populations—individuals of European ancestry and individuals of African ancestry. Since we use the WTCCC data as the primary data and this data set includes only individuals of European ancestry, we do not utilize the results from the above three studies. Therefore, the other two studies—Baum et al. (2007) and Sklar et al. (2008)—serve as “previous” studies which provide us with prior information of potentially important variants for bipolar disorder. The following genes have been reported as top findings in these two studies: MYO5B, TSPAN8, EGFR, SORCS2, DFNB31, DGKH, VGCNL1, and NXN. In our context, we regard the SNPs in those genes as a previously identified set of variants whose information we can borrow in our current study. GeneALaCart (<http://www.genecards.org/BatchQueries/index.php>), the batch-querying platform of the GeneCards database, is used to extract the SNPs within these eight genes.

Certain quality control and prescreening procedures are performed before we carry out the association study for the WTCCC data. All SNPs with Minor Allele Frequency (MAF) less than 0.05 or failing the Hardy-Weinberg equilibrium test at p -value 0.0001 are excluded from further analysis. To reduce the computational burden, a univariate SNP analysis was performed to select a smaller set of SNPs that we are going to analyze with a regression model. Specifically, a covariate-adjusted (adjusted by age and gender) regression model using bipolar disorder as the response variable is run separately for each SNP after quality control. A SNP is included in the final analysis if either of the following two conditions is satisfied: (a) the p -value of the SNP from the logistic regression is smaller than 0.0001; (b) the SNP belongs to one of the eight previously identified genes and its logistic regression p -value is smaller than 0.1. With these selection criteria, the final data set in our analysis includes 916 SNPs in total, among which 174 SNPs compose the prior set S^p as in Section 2.2. The tuning parameters λ and η in both LASSO and pLASSO are selected by a 3-fold cross validation, together with the “one standard error rule” for λ . The final model is fitted using the optimal choice of tuning parameters. The algorithm in Meier et al. (2008) is employed in this example.

From our final results, the optimal choice of η is selected to be 0.2 in pLASSO. Compared with the weight of 1 for the data part in (6), it indicates that pLASSO imposes a mild belief on the prior information by balancing the trade-off between the data and the prior information. The relatively small value of η partially supports the findings that various studies might not agree with each other in detecting the genetic factors underlying bipolar disorder (Baum et al., 2008; Ollila et al., 2009). However, it also gives us some evidence that the data can be better fitted by incorporating prior information to some extent. Therefore, compared with LASSO based on the data themselves, pLASSO has the flexibility of deciding whether or not and to what extent we should incorporate the previous findings in a current study. In addition, both pLASSO and LASSO select a small value of λ , resulting in a large final model for this data set. This observation supports the common belief that bipolar disorder is a complex disease with multiple genetic variants of small effect sizes.

In GWAS, researchers are often looking for a list of top SNPs ranked by statistical significance, which deserves further investigation. With penalized regression, researchers are using different criteria to rank the SNPs, including effect size (Cho et al., 2009), self-developed significance (Wu et al., 2009), etc. To keep our situation simple, we provide a list of top SNPs sorted by effect size from pLASSO and LASSO separately. We will compare how two methods affect the list. Table 1 provides the top twenty SNPs with the largest effect sizes, along with the gene in which each SNP is located or its nearest gene within a distance of 60 kb (Johnson and O'Donnell, 2009).

It is seen that 15 SNPs are overlapping between these two lists, and they possess similar odds ratio estimates. This is due to the small choice of η from our calculation. In other words, two methods partially agree with each other when they are applied to the WTCCC data. It is also noteworthy that rs659991, a SNP in both two top lists, belongs to the gene DGKH which is one of the potential gene associated with bipolar disorder by Baum et al. (2007). This provides us with some evidence for the common genetic variants from the two studies (Baum et al., 2007; WTCCC, 2007).

There are 5 SNPs, however, in each list that are not matched between the two lists. This illustrates the effect of pLASSO compared with LASSO. In particular, one of the 5 unmatched SNPs selected by pLASSO, rs10982246, belongs to the gene DFNB31 as the previously identified information (Baum et al., 2007). Without considering the prior information, we would have missed this gene in the top findings solely by LASSO. Therefore, pLASSO provides a flexible framework to make analysis using information from multiple data sets.

It is also interesting to compare the results from multivariate analysis (LASSO/pLASSO) with univariate analysis in the original paper WTCCC (2007). As indicated in Table 1, some of our top SNPs/genes were identified to possess a strong or moderate association with bipolar disorder in WTCCC (2007). In summary, this example illustrates the applicability of pLASSO to, but not limited to, genetic studies. Certainly, our findings warrant further investigation, but we will leave this to our future work.

7 DISCUSSION

7.1 Summary

This paper presents pLASSO to incorporate prior information into penalized generalized linear models. pLASSO summarizes the prior information into an additional loss function term to achieve a balance between the prior information and the data, where the balance is controlled by a tuning parameter η . Notice that this is different from the elastic net method where the balance is between the L_1 and L_2 penalties (Zou and Hastie, 2005; Bunea, 2008). However, a shared feature is that neither the L_2 penalty nor our additional loss is meant to be standalone. Instead, they provide additional balance to improve the overall performance of the estimation process. Specifically, the L_2 penalty stabilizes the estimator and our additional loss increases its efficiency if the prior information is reliable.

Although we keep the settings simple, this represents the first attempt to examine such a scenario, to the best of our knowledge. A few comments are noteworthy. Firstly, the generalized linear models under consideration are coupled with the canonical link functions. Secondly, we keep the L_1 penalty as in LASSO and mainly compare pLASSO to LASSO but not other variable selection methods. We study both theoretically and empirically the effect of incorporating prior information by altering the loss function part. Both theoretical and numerical results show that pLASSO is more efficient when the prior information is of high quality, and is robust to the low quality prior information with a good choice of the tuning parameter η .

7.2 Tuning Parameters

For the choice of the tuning parameter η , we mainly investigate the performance of cross validation in our simulations. If incorporating the prior information improves the estimator's performance, cross validation will pick a relatively large tuning parameter η to improve the efficiency of the resultant estimator. Otherwise, cross validation will select a small tuning parameter to maintain the estimator's robustness. Simulation results support the effectiveness of cross validation for our purpose. Depending on the quality, pLASSO automatically incorporates the prior information to a proper extent and achieves estimation efficiency and prediction accuracy.

We should note that practical tuning methods, such as the cross validation used in our numerical studies, may not satisfy the conditions imposed on λ or η in our theory. Theoretical justification of practical tuning methods is known to be very challenging and remains to be an open problem, especially for high-dimensional data. We utilize a commonly-used practical tuning approach, and recognize that it is important to investigate this problem and pursue alternative approaches to selecting the tuning parameters. Methods that are computationally more efficient than cross-validation grid search are preferred and need to be pursued, e.g., the bisection method in Bunea and Barbu (2009).

7.3 pLASSO, LASSO, and adaptive LASSO

The distinction between pLASSO and LASSO is the additional loss function which measures the goodness of fit to the prior information (instead of the fit to the data in the usual loss function). Incorporating high-quality prior information will result in a more efficient estimator than LASSO, which is theoretically supported by a smaller asymptotic variance (Section 4.1), a smaller L_∞ loss (Section 4.2), and a lower excess risk rate (Section 4.3). It is also practically verified by simulations (Section 5.1). However, incorporating low-quality prior information will lead to the asymptotic bias (Section 4.1) or a higher excess risk rate (Section 4.3). Fortunately, we can adaptively select the tuning parameter η , making pLASSO robust to the low-quality prior information (Section 5.1).

Here, pLASSO only alters the loss function part but keeps the L_1 penalty exactly the same as in LASSO. In our numerical studies, pLASSO and adaptive LASSO does not dominate each other and have different advantages under different settings, and hence complement each other. Adaptive LASSO can incorporate the prior information into the penalty part through its adaptive weights, e.g., as in (35). It is interesting to explore in the future whether we can

incorporate prior information into both loss and penalty so that estimation efficiency and variable selection advantage can be simultaneously achieved. We have made preliminary progress in investigating the combination of pLASSO and adaptive LASSO (i.e., pLASSOA) in our simulation studies (Section 5.2), and demonstrated its advantage over adaptive LASSO. However, this issue warrants further studies.

Acknowledgments

The authors thank the editor, the associate editor, and two anonymous referees for their comments and suggestions that led to considerable improvements of the paper. This research is supported in part by grants R01 DA016750 and R01 DA029081 from the National Institutes of Health (NIH). The dataset used for the analyses described in this manuscript was obtained from the Wellcome Trust Case-Control Consortium. A full list of the investigators who contributed to the generation of the data is available from www.wtccc.org.uk. Funding for the project was provided by the Wellcome Trust under award 076113 and 085475.

References

- Bach F. Self-concordant analysis for logistic regression. *Electronic Journal of Statistics*. 2010; 4:384–414.
- Baum A, Akula N, Cabanero M, Cardona I, Corona W, Klemens B, Schulze T, Cichon S, Rietschel M, Nöthen M, et al. A genome-wide association study implicates diacylglycerol kinase eta (DGKH) and several other genes in the etiology of bipolar disorder. *Molecular Psychiatry*. 2007; 13:197–207. [PubMed: 17486107]
- Baum A, Hamshere M, Green E, Cichon S, Rietschel M, Nothen M, Craddock N, McMahon F. Meta-analysis of two genome-wide association studies of bipolar disorder reveals important points of agreement. *Molecular Psychiatry*. 2008; 13:466–467. [PubMed: 18421293]
- Bickel PJ, Ritov Y, Tsybakov AB. Simultaneous analysis of Lasso and Dantzig selector. *The Annals of Statistics*. 2009; 37:1705–1732.
- Breiman, L.; Friedman, J.; Olshen, R.; Stone, C. *Classification and Regression Trees*. Wadsworth International Group; 1984.
- Bunea F. Honest variable selection in linear and logistic regression models via ℓ_1 and $\ell_1 + \ell_2$ penalization. *Electronic Journal of Statistics*. 2008; 2:1153–1194.
- Bunea F, Barbu A. Dimension reduction and variable selection in case control studies via regularized likelihood optimization. *Electronic Journal of Statistics*. 2009; 3:1257–1287.
- Bunea F, Tsybakov A, Wegkamp M. Sparsity oracle inequalities for the Lasso. *Electronic Journal of Statistics*. 2007; 1:169–194.
- Cantor R, Lange K, Sinsheimer J. Prioritizing GWAS results: A review of statistical methods and recommendations for their application. *The American Journal of Human Genetics*. 2010; 86:6–22. [PubMed: 20074509]
- Cho S, Kim H, Oh S, Kim K, Park T. Elastic-net regularization approaches for genome-wide association studies of rheumatoid arthritis. *BMC proceedings*. 2009; 3:S25. [PubMed: 20018015]
- Collett, D. *Modelling Binary Data*. CRC Press; 2003.
- Craddock N, Forty L. Genetics of affective (mood) disorders. *European Journal of Human Genetics*. 2006; 14:660–668. [PubMed: 16721402]
- Efron B, Hastie T, Johnstone I, Tibshirani R. Least Angle Regression. *Annals of Statistics*. 2004; 32:407–451.
- Fahrmeir L, Kaufmann H. Consistency and Asymptotic Normality of the Maximum Likelihood Estimator in Generalized Linear Models. *Annals of Statistics*. 1985; 13:342–368.
- Fan J, Lv J. Non-concave penalized likelihood with NP-dimensionality. *IEEE Transactions on Information Theory*. 2011; 57:5467–5484. [PubMed: 22287795]
- Ferreira M, O'Donovan M, Meng Y, Jones I, Ruderfer D, Jones L, Fan J, Kirov G, Perlis R, Green E, et al. Collaborative genome-wide association analysis supports a role for ANK3 and CACNA1C in bipolar disorder. *Nature Genetics*. 2008; 40:1056–1058. [PubMed: 18711365]

- Friedman J, Hastie T, Tibshirani R. Special invited paper. Additive logistic regression: A statistical view of boosting. *Annals of Statistics*. 2000; 28:337–374.
- Friedman J, Hastie T, Tibshirani R. Regularization paths for generalized linear models via coordinate descent. *Journal of Statistical Software*. 2010; 33:1–22. [PubMed: 20808728]
- Jiang Y, Zhang CM. High-dimensional regression and classification under a class of convex loss functions. *Statistics and Its Interface*. 2013; 6:285–299.
- Johnson A, O'Donnell C. An open access database of genome-wide association results. *BMC medical genetics*. 2009; 10:6. [PubMed: 19161620]
- Knight K, Fu W. Asymptotics for lasso-type estimators. *The Annals of Statistics*. 2000; 28:1356–1378.
- Koltchinskii V, Lounici K, Tsybakov AB. Nuclear-norm penalization and optimal rates for noisy low-rank matrix completion. *The Annals of Statistics*. 2011; 39:2302–2329.
- Kwemou M. Non-asymptotic Oracle Inequalities for the Lasso and Group Lasso in high dimensional logistic model. 2012 arXiv preprint arXiv:1206.0710.
- McCullagh, P.; Nelder, J. *Generalized Linear Models*. Chapman & Hall/CRC; 1989.
- Meier L, van der Geer S, Bühlmann P. The group lasso for logistic regression. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*. 2008; 70:53–71.
- Meinshausen N, Yu B. Lasso-type recovery of sparse representations for high-dimensional data. *Annals of Statistics*. 2009; 37:246–270.
- Merikangas K, Akiskal H, Angst J, Greenberg P, Hirschfeld R, Petukhova M, Kessler R. Lifetime and 12-month prevalence of bipolar spectrum disorder in the National Comorbidity Survey replication. *Archives of general psychiatry*. 2007; 64:543–552. [PubMed: 17485606]
- Ollila H, Soronen P, Silander K, Palo O, Kieseppä T, Kaunisto M, Lönnqvist J, Peltonen L, Partonen T, Paunio T. Findings from bipolar disorder genome-wide association studies replicate in a Finnish bipolar family-cohort. *Molecular Psychiatry*. 2009; 14:351–353. [PubMed: 19308021]
- Raskutti G, Wainwright MJ, Yu B. Restricted eigenvalue properties for correlated Gaussian designs. *The Journal of Machine Learning Research*. 2010; 11:2241–2259.
- Rosset S, Zhu J. Piecewise linear regularized solution paths. *Annals of Statistics*. 2007; 35:1012–1030.
- Schapiro R, Rochery M, Rahim M, Gupta N. Boosting with prior knowledge for call classification. *Speech and Audio Processing, IEEE Transactions on*. 2005; 13:174–181.
- Scott L, Muglia P, Kong X, Guan W, Flickinger M, Upmanyu R, Tozzi F, Li J, Burmeister M, Absher D, et al. Genome-wide association and meta-analysis of bipolar disorder in individuals of European ancestry. *Proceedings of the National Academy of Sciences*. 2009; 106:7501–7506.
- Sklar P, Smoller J, Fan J, Ferreira M, Perlis R, Chambert K, Nimgaonkar V, McQueen M, Faraone S, Kirby A, et al. Whole-genome association study of bipolar disorder. *Molecular Psychiatry*. 2008; 13:558–569. [PubMed: 18317468]
- Smith E, Bloss C, Badner J, Barrett T, Belmonte P, Berrettini W, Byerley W, Coryell W, Craig D, Edenberg H, et al. Genome-wide association study of bipolar disorder in European American and African American individuals. *Molecular Psychiatry*. 2009; 14:755–763. [PubMed: 19488044]
- Tibshirani R. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*. 1996; 58:267–288.
- Van de Geer SA. High-dimensional generalized linear models and the Lasso. *The Annals of Statistics*. 2008; 36:614–645.
- van der Vaart, AW.; Wellner, JA. *Weak Convergence and Empirical Processes: With Applications to Statistics*. New York: Springer-Verlag; 1996.
- WTCCC. Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. *Nature*. 2007; 447:661–678. [PubMed: 17554300]
- Wu T, Chen Y, Hastie T, Sobel E, Lange K. Genome-wide association analysis by lasso penalized logistic regression. *Bioinformatics*. 2009; 25:714–721. [PubMed: 19176549]
- Wu T, Lange K. Coordinate descent algorithms for lasso penalized regression. *Annals of Applied Statistics*. 2008; 2:224–244.
- Wu Y. An ordinary differential equation-based solution path algorithm. *Journal of nonparametric statistics*. 2011; 23:185–199. [PubMed: 21532936]

- Zeggini E, Ioannidis J. Meta-analysis in genome-wide association studies. *Pharmacogenomics*. 2009; 10:191–201. [PubMed: 19207020]
- Zeggini E, Scott L, Saxena R, Voight B, Marchini J, Hu T, de Bakker P, Abecasis G, Almgren P, Andersen G, et al. Meta-analysis of genome-wide association data and large-scale replication identifies additional susceptibility loci for type 2 diabetes. *Nature Genetics*. 2008; 40:638–645. [PubMed: 18372903]
- Zhang H, Lu W. Adaptive Lasso for Cox's proportional hazards model. *Biometrika*. 2007; 94:691–703.
- Zhao P, Yu B. On model selection consistency of Lasso. *Journal of Machine Learning Research*. 2006; 7:2541–2563.
- Zou H. The adaptive lasso and its oracle properties. *Journal of the American Statistical Association*. 2006; 101:1418–1429.
- Zou H, Hastie T. Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*. 2005; 67:301–320.

8 APPENDIX

8.1 Proof of Theorem 2

By the continuity of $\hat{\beta}^{\lambda, \eta}$ and $\mathbf{G}^{\lambda, \eta}$ of η , it is impossible for the two groups $\mathcal{J}_1^{\lambda, \eta}$ and $\mathcal{J}_3^{\lambda, \eta}$ to exchange elements directly. Hence, any change of $\mathcal{A}^{\lambda, \eta}$ can only happen if $\mathcal{J}_2^{\lambda, \eta}$ is not empty, and that single element in $\mathcal{J}_2^{\lambda, \eta}$ will be the one either entering or deleted from $\mathcal{A}^{\lambda, \eta}$. Therefore, the “only if” parts for (a) and (b) are straightforward. We focus on the “if” parts as follows.

For (a), it is seen from (17) that

$$\hat{\beta}_{\mathcal{A}_k}^{\lambda, \eta} = \frac{1+\eta_k}{1+\eta} (\hat{\beta}_{\mathcal{A}_k}^{\lambda, \eta_k} - \hat{\beta}_{\mathcal{A}_k}^{\mathcal{A}_k, \mathbf{P}}) + \hat{\beta}_{\mathcal{A}_k}^{\mathcal{A}_k, \mathbf{P}}, \quad \eta \in (\eta_k, \eta_{k+1}).$$

So $\hat{\beta}_j^{\lambda, \eta}$ is a monotone function of $\eta \in (\eta_k, \eta_{k+1})$ for any $j \in \mathcal{A}_k$. Then if some j enters $\mathcal{J}_2^{\lambda, \eta}$ from $\mathcal{J}_1^{\lambda, \eta}$ at η , i.e., $\hat{\beta}_j^{\lambda, \eta}$ becomes 0 at η for some $j \in \mathcal{A}_k$, η must be a discontinuity point of $\mathcal{A}^{\lambda, \eta}$. That is, $\eta_{k+1} = \eta$. If not, $\eta < \eta_{k+1}$ and $\mathcal{A}^{\lambda, \eta}$ stays as $\mathcal{A}_k$. By the monotonicity, $\hat{\beta}_j^{\lambda, \eta}$ will change its sign, which is contradictive to the continuity of $\mathbf{G}^{\lambda, \eta}$ and the fact that $G_j^{\lambda, \eta} = -\lambda \text{sign}(\hat{\beta}_j^{\lambda, \eta})$ for any $j \in \mathcal{A}_k$. Moreover, since $\eta_{k+1} = \eta$ is proved, $\mathcal{A}_{k+1} = \mathcal{A}_k \setminus \{j\}$ is the only allowed change of $\mathcal{A}^{\lambda, \eta}$.

For (b), by Theorem 1, we have that when $\eta \in (\eta_k, \eta_{k+1})$,

$$\begin{aligned} G_{\mathcal{A}_k}^{\lambda, \eta} &= n^{-1} (1+\eta_k) \mathbb{X}_{\mathcal{A}_k^c}^T \mathbb{X}_{\mathcal{A}_k} \hat{\beta}_{\mathcal{A}_k}^{\lambda, \eta_k} + n^{-1} (\eta - \eta_k) \mathbb{X}_{\mathcal{A}_k^c}^T \mathbb{X}_{\mathcal{A}_k} \hat{\beta}_{\mathcal{A}_k}^{\mathcal{A}_k, \mathbf{P}} - n^{-1} \mathbb{X}_{\mathcal{A}_k^c}^T (\mathbf{Y} + \eta \hat{\mathbf{Y}}^{\mathbf{P}}) \\ &= G_{\mathcal{A}_k}^{\lambda, \eta_k} + n^{-1} (\eta - \eta_k) \mathbb{X}_{\mathcal{A}_k^c}^T (\mathbb{X}_{\mathcal{A}_k} \hat{\beta}_{\mathcal{A}_k}^{\mathcal{A}_k, \mathbf{P}} - \hat{\mathbf{Y}}^{\mathbf{P}}). \end{aligned}$$

Hence $G_j^{\lambda,\eta}$ is a monotone function of $\eta \in (\eta_k, \eta_{k+1})$ for any $j \in \mathcal{A}_k^c$. Then if j enters $\mathcal{J}_2^{\lambda,\eta}$ from $\mathcal{J}_3^{\lambda,\eta}$ at η , i.e., $|G_j^{\lambda,\eta}|$ becomes λ for some $j \in \mathcal{A}_k^c$, η must be a discontinuity point of $\mathcal{A}^{\lambda,\eta}$. That is, $\eta_{k+1} = \eta$. If not, $\eta < \eta_{k+1}$ and $\mathcal{A}^{\lambda,\eta}$ stays as $\mathcal{A}_k$. By the monotonicity, $|G_j^{\lambda,\eta}|$ will be larger than λ which is impossible. Moreover, since $\eta_{k+1} = \eta$ is proved, $\mathcal{A}_{k+1} = \mathcal{A}_k \cup \{j\}$ is the only allowed change of $\mathcal{A}^{\lambda,\eta}$.

8.2 Proof of Theorem 3

Let $\phi = \sqrt{n}(\beta - \beta_0)$. In the LASSO criterion function (3), using Taylor's expansion,

$$\begin{aligned} nL(\beta; \mathbb{X}, \mathbf{Y}) &= nL(\beta_0 + n^{-1/2}\phi; \mathbb{X}, \mathbf{Y}) \\ &\simeq -\frac{1}{\sqrt{n}} \sum_{i=1}^n Y_i \mathbf{Z}_i^T \phi + \sum_{i=1}^n b(\mathbf{Z}_i^T \beta_0 + n^{-1/2} \mathbf{Z}_i^T \phi) \\ &\simeq -\frac{1}{\sqrt{n}} \sum_{i=1}^n [Y_i - b'(\mathbf{Z}_i^T \beta_0)] \mathbf{Z}_i^T \phi + \frac{1}{2n} \sum_{i=1}^n b''(\mathbf{Z}_i^T \beta_0) (\mathbf{Z}_i^T \phi)^2 + o_p(1) \end{aligned} \quad (\text{A.1})$$

$$\rightarrow_d -\mathbf{u}^T \phi + \frac{1}{2} \phi^T \boldsymbol{\Omega} \phi = Q(\phi), \quad (\text{A.2})$$

where $\simeq$ stands for the equality if ignoring an additive term independent with ϕ . (A.2) follows from Lemma 1 in Fahrmeir and Kaufmann (1985) by verifying their conditions (D) and (N). Furthermore, since $\sqrt{n}\lambda \rightarrow \lambda_0$,

$$n\lambda \sum_{j=1}^p (|\beta_j| - |\beta_{j,0}|) \rightarrow \lambda_0 \sum_{j=1}^p [\phi_j \text{sign}(\beta_{j,0}) I(\beta_{j,0} \neq 0) + |\phi_j| I(\beta_{j,0} = 0)] = P_{\lambda_0}(\phi).$$

Thus, the proof is finished using the same argument as in the proof of Theorem 2 in Knight and Fu (2000).

8.3 Proof of Theorem 4

We briefly outline the proof. As the pLASSO criterion function (6) includes the two loss functions $L(\beta; \mathbb{X}, \mathbf{Y})$ and $L(\beta; \mathbb{X}, \hat{\mathbf{Y}}^p)$, we need to evaluate them jointly. To do so, we first write $L(\beta; \mathbb{X}, \mathbf{Y})$ in terms of $\hat{\beta}$, the maximum likelihood estimator of β , i.e., the minimizer of the criterion function (2); and also write $L(\beta; \mathbb{X}, \hat{\mathbf{Y}}^p)$ in terms of $\hat{\beta}^p$, the prior estimator of β . Then, the two loss functions can be jointly evaluated by studying the joint distribution of $\hat{\beta}$ and $\hat{\beta}^p$. In another word, the relationship between the observed response $\mathbf{Y}$ and the prior response $\hat{\mathbf{Y}}^p$ can be investigated through jointly studying $\hat{\beta}$ and $\hat{\beta}^p$, which is feasible because the two estimators minimize (2) and (8) separately but based on the same data set. This is a key step in the proof. The detailed proof follows.

8.3.1 Case 1: $\eta \rightarrow \eta_0 = 0$

Let $\phi = \sqrt{n}(\beta - \beta_0)$. In the pLASSO criterion function (6), similar to (A.1),

$$nL(\beta; \mathbb{X}, \hat{\mathbf{Y}}^p) \simeq -\frac{1}{\sqrt{n}} \sum_{i=1}^n [\hat{Y}_i^p - b'(\mathbf{Z}_i^T \beta_0)] \mathbf{Z}_i^T \phi + \frac{1}{2n} \sum_{i=1}^n b''(\mathbf{Z}_i^T \beta_0) (\mathbf{Z}_i^T \phi)^2 + o_p(1). \quad (\text{A.3})$$

First, we write the first terms in (A.1) and (A.3) in terms of $\hat{\beta}$ and $\hat{\beta}^p$ respectively as follows:

$$\frac{1}{\sqrt{n}} \sum_{i=1}^n [Y_i - b'(\mathbf{Z}_i^T \beta_0)] \mathbf{Z}_i^T \phi = \sqrt{n}(\hat{\beta} - \beta_0)^T \Omega \phi + o_p(1), \quad (\text{A.4})$$

$$\frac{1}{\sqrt{n}} \sum_{i=1}^n [\hat{Y}_i^p - b'(\mathbf{Z}_i^T \beta_0)] \mathbf{Z}_i^T \phi = \sqrt{n}(\hat{\beta}^p - \beta_0)^T \Omega \phi + o_p(1). \quad (\text{A.5})$$

By verifying conditions (D) and (N) in Fahrmeir and Kaufmann (1985), (A.4) is a direct conclusion from Theorems 3 therein and also (3.13) in its proof. For (A.5),

$$\begin{aligned} \frac{1}{\sqrt{n}} \sum_{i=1}^n [\hat{Y}_i^p - b'(\mathbf{Z}_i^T \beta_0)] \mathbf{Z}_i^T \phi &= \frac{1}{\sqrt{n}} \sum_{i=1}^n b''(\mathbf{Z}_i^T \tilde{\beta}) \mathbf{Z}_i^T (\hat{\beta}^p - \beta_0) \mathbf{Z}_i^T \phi \\ &= \sqrt{n}(\hat{\beta}^p - \beta_0)^T \frac{1}{n} \sum_{i=1}^n b''(\mathbf{Z}_i^T \beta_0) \mathbf{Z}_i \mathbf{Z}_i^T \phi + o_p(1) \\ &= \sqrt{n}(\hat{\beta}^p - \beta_0)^T \Omega \phi + o_p(1). \end{aligned}$$

with $\tilde{\beta}$ between β_0 and $\hat{\beta}^p$. The above equalities hold due to the fact that $\sqrt{n}(\hat{\beta}^p - \beta_0) = O_p(1)$, which can be proved similarly to Theorem 3 since $\sqrt{n}\kappa \rightarrow \kappa_0 \geq 0$.

Next, we evaluate the asymptotic joint distribution of $\hat{\beta}$ and $\hat{\beta}^p$. It is observed that

$$\begin{aligned} &\left[\sqrt{n}(\hat{\beta} - \beta_0)^T, \sqrt{n}(\hat{\beta}^p - \beta_0)^T \right]^T \\ &= \arg \min_{\phi_1, \phi_2} [L(\beta_0 + n^{-1/2} \phi_1; \mathbb{X}, \mathbf{Y}) + L_{\kappa_0, \mathcal{S}^p}(\beta_0 + n^{-1/2} \phi_2; \mathbb{X}, \mathbf{Y})], \end{aligned}$$

and a similar argument to (A.1) and (A.2) leads to

$$nL(\beta_0 + n^{-1/2} \phi_1; \mathbb{X}, \mathbf{Y}) + nL_{\kappa_0, \mathcal{S}^p}(\beta_0 + n^{-1/2} \phi_2; \mathbb{X}, \mathbf{Y}) \rightarrow_d Q(\phi_1) + Q(\phi_2) + P_{\kappa_0, \mathcal{S}^p}(\phi_2).$$

Similar to Theorem 2 in Knight and Fu (2000), the above observations imply that

$$\begin{aligned} &\left[\sqrt{n}(\hat{\beta} - \beta_0)^T, \sqrt{n}(\hat{\beta}^p - \beta_0)^T \right]^T \rightarrow_d \arg \min [Q(\phi_1) + Q(\phi_2) + P_{\kappa_0, \mathcal{S}^p}(\phi_2)] \\ &= [\arg \min (Q)^T, \arg \min (Q + P_{\kappa_0, \mathcal{S}^p})^T]^T. \end{aligned}$$

Finally, from (A.1), (A.3)–(A.5), and the above asymptotic joint distribution,

$$\begin{aligned} & nL(\beta_0 + n^{-1/2}\phi; \mathbb{X}, \mathbf{Y}) + n\eta L(\beta_0 + n^{-1/2}\phi; \mathbb{X}, \hat{\mathbf{Y}}^P) \\ &= - \left[\sqrt{n}(\hat{\beta} - \beta_0) + n \sqrt{n}(\hat{\beta}^P - \beta_0) \right]^T \Omega \phi + \frac{1+\eta}{2n} \sum_{i=1}^n b''(\mathbf{Z}_i^T \beta_0) (\mathbf{Z}_i^T \phi)^2 + o_P(1) \\ &\rightarrow_d (1+\eta_0) \phi^T \Omega \phi / 2 - [\arg \min(Q) + \eta_0 \arg \min(Q + P_{\kappa_0, \mathcal{S}^P})]^T \Omega \phi. \end{aligned}$$

Then the proof is finished similar to that of Theorem 3. F

8.3.2 Case 2: $\eta \rightarrow \infty$

Let $\phi = \sqrt{n}(\beta - \beta_0)$. Similar to the proof in Case 1,

$$\begin{aligned} n\eta^{-1} L_{\lambda, \eta}(\beta; \mathbb{X}, \mathbf{Y}, \hat{\mathbf{Y}}^P) &= n\eta^{-1} L(\beta; \mathbb{X}, \mathbf{Y}) + nL(\beta; \mathbb{X}, \hat{\mathbf{Y}}^P) + n\eta^{-1} \lambda \sum_{j=1}^p |\beta_j| \\ &\rightarrow_d \phi^T \Omega \phi / 2 - [\arg \min(Q + P_{\kappa_0, \mathcal{S}^P})]^T \Omega \phi, \end{aligned}$$

ignoring an additive term independent with ϕ . So the conclusion is proved.

8.4 Proof of Theorem 6

Similar to Theorem 1 in Fan and Lv (2011), the Karush-Kuhn-Tucker (KKT) sufficient conditions for (29) are

$$\mathbb{X}_1^T [\mathbf{Y} - b'(\mathbb{X}_1 \hat{\beta}_1)] + \eta \mathbb{X}_1^T [\hat{\mathbf{Y}}^P - b'(\mathbb{X}_1 \hat{\beta}_1)] - n\lambda \text{sign}(\hat{\beta}_1) = 0, \quad (\text{A.6})$$

$$\|\mathbb{X}_2^T [\mathbf{Y} - b'(\mathbb{X}_1 \hat{\beta}_1)] + \eta \mathbb{X}_2^T [\hat{\mathbf{Y}}^P - b'(\mathbb{X}_1 \hat{\beta}_1)]\|_\infty < n\lambda, \quad (\text{A.7})$$

where $f(\mathbf{u}) = [f(u_1), \dots, f(u_n)]^T$ for any univariate function f and vector $\mathbf{u} = (u_1, \dots, u_n)^T$ for simplicity of notation hereafter. (A.6) and (A.7) imply that $\hat{\beta} = (\hat{\beta}_1^T, \mathbf{0}^T)^T$ is a pLASSO estimator.

8.4.1 Step 1: Existence of a solution to (A.6)

We prove that (A.6) has a solution $\hat{\beta}_1$ inside the hypercube

$$\mathcal{N} = \{\beta_1 \in \mathbb{R}^s : \|\beta_1 - \beta_{1,0}\|_\infty = n^{-\gamma} \log n / (1+\eta)\}.$$

For any $\beta_1 \in \mathcal{N}$, since $d_n = n^{-\gamma} \log n$, we have $\text{sign}(\beta_1) = \text{sign}(\beta_{1,0})$ and

$$\min_{1 \leq j \leq s} |\beta_j| \geq \min_{1 \leq j \leq s} |\beta_{j,0}| - n^{-\gamma} \log n / (1+\eta) \geq \frac{1+2\eta}{1+\eta} n^{-\gamma} \log n.$$

With a slight abuse of notation, let $\boldsymbol{\gamma}(\boldsymbol{\beta}) = \mathbb{X}^T b(\mathbb{X}\boldsymbol{\beta})$, $\boldsymbol{\gamma}(\boldsymbol{\beta}_1) = \mathbb{X}^T b(\mathbb{X}\boldsymbol{\beta}_1)$, $\boldsymbol{\xi} = \mathbb{X}^T [\mathbf{Y} - b(\mathbb{X}\boldsymbol{\beta}_0)]$ and $\boldsymbol{\xi}^P = \mathbb{X}^T [\hat{\mathbf{Y}}^P - b(\mathbb{X}\boldsymbol{\beta}_0)]$, then (A.6) is equivalent to $\boldsymbol{\Psi}_1(\hat{\boldsymbol{\beta}}_1) = \mathbf{0}$ where

$$\boldsymbol{\Psi}_1(\boldsymbol{\beta}_1) = (1+\eta)[\boldsymbol{\gamma}_1(\boldsymbol{\beta}_1) - \boldsymbol{\gamma}_1(\boldsymbol{\beta}_{1,0})] - \boldsymbol{\xi}_1 - \eta \boldsymbol{\xi}_1^P + n\lambda \text{sign}(\boldsymbol{\beta}_{1,0}).$$

Similar to the proof of Theorem 2 in Fan and Lv (2011), $\|\boldsymbol{\xi}_1\|_\infty = O_P(\sqrt{n \log n})$, and

$$\boldsymbol{\gamma}_1(\boldsymbol{\beta}_1) - \boldsymbol{\gamma}_1(\boldsymbol{\beta}_{1,0}) = \mathbb{X}_1^T \sum (\boldsymbol{\beta}_0) \mathbb{X}_1 (\boldsymbol{\beta}_1 - \boldsymbol{\beta}_{1,0}) + \mathbf{r}_1,$$

where $\mathbf{r}_1 = (r_1, \dots, r_s)^T$ and for each $j = 1, \dots, s$,

$$r_j = \frac{1}{2} (\boldsymbol{\beta}_1 - \boldsymbol{\beta}_{1,0})^T \frac{\partial^2}{\partial \boldsymbol{\beta}_1 \partial \boldsymbol{\beta}_1^T} \gamma_j(\tilde{\boldsymbol{\beta}}_{1,j}) (\boldsymbol{\beta}_1 - \boldsymbol{\beta}_{1,0}),$$

with $\tilde{\boldsymbol{\beta}}_{1,j}$ lying on the line segment between $\boldsymbol{\beta}_1$ and $\boldsymbol{\beta}_{1,0}$. Therefore,

$$[\mathbb{X}_1^T \sum (\boldsymbol{\beta}_0) \mathbb{X}_1]^{-1} \boldsymbol{\Psi}_1(\boldsymbol{\beta}_1) = (1+\eta)(\boldsymbol{\beta}_1 - \boldsymbol{\beta}_{1,0}) + \mathbf{u},$$

where

$$\mathbf{u} = [\mathbb{X}_1^T \sum (\boldsymbol{\beta}_0) \mathbb{X}_1]^{-1} [(1+\eta)\mathbf{r}_1 - \boldsymbol{\xi}_1 - \eta \boldsymbol{\xi}_1^P + n\lambda \text{sign}(\boldsymbol{\beta}_{1,0})].$$

In addition, we have similarly that

$$\boldsymbol{\xi}_1^P = \boldsymbol{\gamma}_1(\hat{\boldsymbol{\beta}}^P) - \boldsymbol{\gamma}_1(\boldsymbol{\beta}_0) = \mathbb{X}_1^T \sum (\boldsymbol{\beta}_0) [\mathbb{X}_1(\hat{\boldsymbol{\beta}}_1^P - \boldsymbol{\beta}_{1,0}) + \mathbb{X}_2 \hat{\boldsymbol{\beta}}_2^P] + \mathbf{r}_1^P,$$

where $\mathbf{r}_1^P = (r_1^P, \dots, r_s^P)^T$ and for each $j = 1, \dots, s$,

$$r_j^P = \frac{1}{2} (\hat{\boldsymbol{\beta}}^P - \boldsymbol{\beta}_0)^T \frac{\partial^2}{\partial \boldsymbol{\beta} \partial \boldsymbol{\beta}^T} \gamma_j(\tilde{\boldsymbol{\beta}}_j) (\hat{\boldsymbol{\beta}}^P - \boldsymbol{\beta}_0),$$

with $\tilde{\boldsymbol{\beta}}_j$ lying between $\hat{\boldsymbol{\beta}}$ and $\boldsymbol{\beta}_0$.

This leads to $\mathbf{u} = \mathbf{u}_1 + \mathbf{u}_2 + \mathbf{u}_3$ with

$$\begin{aligned}\mathbf{u}_1 &= [\mathbb{X}_1^T \Sigma(\boldsymbol{\beta}_0) \mathbb{X}_1]^{-1} [(1+\eta) \mathbf{r}_1 - \boldsymbol{\xi}_1 + n\lambda \text{sign}(\boldsymbol{\beta}_{1,0})], \\ \mathbf{u}_2 &= -\eta(\hat{\boldsymbol{\beta}}_1^p - \boldsymbol{\beta}_{1,0}) - \eta[\mathbb{X}_1^T \Sigma(\boldsymbol{\beta}_0) \mathbb{X}_1]^{-1} \mathbb{X}_1^T \Sigma(\boldsymbol{\beta}_0) \mathbb{X}_2 \hat{\boldsymbol{\beta}}_2^p, \\ \mathbf{u}_3 &= -\eta[\mathbb{X}_1^T \Sigma(\boldsymbol{\beta}_0) \mathbb{X}_1]^{-1} \mathbf{r}_1^p.\end{aligned}$$

Using the similar arguments as in the proof of Theorem 2 in Fan and Lv (2011),

$$\|\mathbf{u}_1\|_\infty = O(b_s n^{-1}) O_P[sn^{1-2\gamma}(\log n)^2 / (1+\eta) + \sqrt{n \log n} + n\lambda].$$

It is also seen that

$$\begin{aligned}\|\mathbf{u}_2\|_\infty &\leq \eta \|\hat{\boldsymbol{\beta}}_1^p - \boldsymbol{\beta}_{1,0}\|_\infty + \eta \|[\mathbb{X}_1^T \Sigma(\boldsymbol{\beta}_0) \mathbb{X}_1]^{-1}\|_\infty \|\mathbb{X}_1^T \Sigma(\boldsymbol{\beta}_0) \mathbb{X}_2 \hat{\boldsymbol{\beta}}_2^p\|_\infty, \\ \|\mathbf{u}_3\|_\infty &= O(\eta b_s n^{-1}) \max_{\boldsymbol{\beta} \in \mathcal{N}_0^p} \max_{1 \leq j \leq s} \lambda_{\max}[\mathbb{X}^T \text{diag}\{|\mathbf{x}_j| \circ |b'''(\mathbb{X}\boldsymbol{\beta})|\} \mathbb{X}] \|\hat{\boldsymbol{\beta}}^p - \boldsymbol{\beta}_0\|_2^2 \\ &= O_P(\eta b_s) \|\hat{\boldsymbol{\beta}}^p - \boldsymbol{\beta}_0\|_2^2.\end{aligned}$$

As $\|\boldsymbol{\beta}_1 - \boldsymbol{\beta}_{1,0}\|_\infty = n^{-\gamma} \log n / (1 + \eta)$ for $\boldsymbol{\beta}_1 \in \mathcal{N}$, it is already proved that $\|\mathbf{u}_1\|_\infty = o_P[(1 + \eta) \|\boldsymbol{\beta}_1 - \boldsymbol{\beta}_{1,0}\|_\infty]$ by Condition B. In addition, under Condition D, it is verified that $\|\mathbf{u}_2\|_\infty = o_P[(1 + \eta) \|\boldsymbol{\beta}_1 - \boldsymbol{\beta}_{1,0}\|_\infty]$ and $\|\mathbf{u}_3\|_\infty = o_P[(1 + \eta) \|\boldsymbol{\beta}_1 - \boldsymbol{\beta}_{1,0}\|_\infty]$ for $\boldsymbol{\beta}_1 \in \mathcal{N}$. Therefore, the existence of $\hat{\boldsymbol{\beta}}_1$ inside $\mathcal{N}$ satisfying $\Psi_1(\hat{\boldsymbol{\beta}}_1) = 0$ is implied by the Miranda's existence theorem as similarly argued by Fan and Lv (2011).

8.4.2 Step 2: Verification of condition (A.7)

Let $\hat{\boldsymbol{\beta}}_1$ be the solution of (A.6). Note that (A.7) is equivalent to $\|\Psi_2\|_\infty < n\lambda$, where

$$\begin{aligned}\Psi_2 &= \mathbb{X}_2^T [\mathbf{Y} - b'(\mathbb{X}_1 \hat{\boldsymbol{\beta}}_1)] + \eta \mathbb{X}_2^T [\hat{\mathbf{Y}}^p - b'(\mathbb{X}_1 \hat{\boldsymbol{\beta}}_1)] \\ &= \boldsymbol{\xi}_2 + \eta \boldsymbol{\xi}_2^p - (1+\eta) [\boldsymbol{\gamma}_2(\hat{\boldsymbol{\beta}}_1) - \boldsymbol{\gamma}_2(\boldsymbol{\beta}_{1,0})].\end{aligned}$$

Similar to Fan and Lv (2011), $\|\boldsymbol{\xi}_2\|_\infty = O_P(n^{1-\alpha} \sqrt{\log n})$, and

$$\boldsymbol{\gamma}_2(\hat{\boldsymbol{\beta}}_1) - \boldsymbol{\gamma}_2(\boldsymbol{\beta}_{1,0}) = \mathbb{X}_2^T \Sigma(\boldsymbol{\beta}_0) \mathbb{X}_1 (\hat{\boldsymbol{\beta}}_1 - \boldsymbol{\beta}_{1,0}) + \mathbf{r}_2,$$

where $\mathbf{r}_2 = (r_{s+1}, \dots, r_p)^T$ and for each $j = s + 1, \dots, p$,

$$r_j = \frac{1}{2} (\hat{\boldsymbol{\beta}}_1 - \boldsymbol{\beta}_{1,0})^T \frac{\partial^2}{\partial \boldsymbol{\beta}_1 \partial \boldsymbol{\beta}_1^T} \gamma_j(\tilde{\boldsymbol{\beta}}_{1,j}) (\hat{\boldsymbol{\beta}}_1 - \boldsymbol{\beta}_{1,0}),$$

with $\tilde{\beta}_1$ lying between $\hat{\beta}_1$ and $\beta_{1,0}$. It follows from Fan and Lv (2011) that

$$\begin{aligned}\|\mathbf{r}_2\|_\infty &\leq \frac{1}{2} \max_{\beta_1 \in \mathcal{N}_0} \max_{s+1 \leq j \leq p} \lambda_{\max}[\mathbb{X}_1^T \text{diag}\{|\mathbf{x}_j| \circ |b'''(\mathbb{X}_1 \beta_1)|\} \mathbb{X}_1] \|\hat{\beta}_1 - \beta_{1,0}\|_2^2 \\ &= O[sn^{1-2\gamma}(\log n)^2/(1+\eta)^2].\end{aligned}$$

In addition, by Step 1,

$$\hat{\beta}_1 - \beta_{1,0} = (1+\eta)^{-1} [\mathbb{X}_1^T \sum (\beta_0) \mathbb{X}_1]^{-1} [\xi_1 + \eta \xi_1^p - (1+\eta) \mathbf{r}_1 - n\lambda \text{sign}(\beta_{1,0})],$$

where $\|\xi_1\|_\infty = O_P(\sqrt{n \log n})$ and $\|\mathbf{r}_1\|_\infty = O[sn^{1-2\gamma}(\log n)^2/(1+\eta)^2]$ from Step 1. Moreover,

$$\|\xi_1^p\|_\infty \leq \|\mathbb{X}_1^T \sum (\beta_0) \mathbb{X}_1 (\hat{\beta}_1^p - \beta_{1,0})\|_\infty + \|\mathbb{X}_1^T \sum (\beta_0) \mathbb{X}_2 \hat{\beta}_2^p\|_\infty + \|\mathbf{r}_1^p\|_\infty,$$

with $\|\mathbf{r}_1^p\|_\infty = O_P(n) \|\hat{\beta}^p - \beta_0\|_2^2$ from Step 1. Therefore,

$$\begin{aligned}&\|\mathbb{X}_2^T \sum (\beta_0) \mathbb{X}_1 (\hat{\beta}_1 - \beta_{1,0})\|_\infty \\ &= (1+\eta)^{-1} \|\mathbb{X}_2^T \sum (\beta_0) \mathbb{X}_1 [\mathbb{X}_1^T \sum (\beta_0) \mathbb{X}_1]^{-1}\|_\infty [\|\xi_1\|_\infty + \eta \|\xi_1^p\|_\infty + (1+\eta) \|\mathbf{r}_1\|_\infty + n\lambda] \\ &= (1+\eta)^{-1} C[O_P(\sqrt{n \log n}) + O[sn^{1-2\gamma}(\log n)^2/(1+\eta)] + \eta \|\mathbb{X}_1^T \sum (\beta_0) \mathbb{X}_1 (\hat{\beta}_1^p - \beta_{1,0})\|_\infty \\ &\quad + \eta \|\mathbb{X}_1^T \sum (\beta_0) \mathbb{X}_2 \hat{\beta}_2^p\|_\infty + O_P(\eta n) \|\hat{\beta}^p - \beta_0\|_2^2 + n\lambda].\end{aligned}$$

Similar to ξ_1^p in Step 1,

$$\xi_2^p = \gamma_2(\hat{\beta}^p) - \gamma_2(\beta_0) = \mathbb{X}_2^T \sum (\beta_0) [\mathbb{X}_1 (\hat{\beta}_1^p - \beta_{1,0}) + \mathbb{X}_2 \hat{\beta}_2^p] + \mathbf{r}_2^p,$$

where $\mathbf{r}_2^p = (r_{s+1}^p, \dots, r_p^p)^T$ and for each $j = s+1, \dots, p$,

$$r_j^p = \frac{1}{2} (\hat{\beta}^p - \beta_0)^T \frac{\partial^2}{\partial \beta \partial \beta^T} \gamma_j(\tilde{\beta}_j) (\hat{\beta}^p - \beta_0),$$

with $\tilde{\beta}_j$ lying between $\hat{\beta}$ and β_0 . Thus,

$$\|\xi_2^p\|_\infty \leq \|\mathbb{X}_2^T \sum (\beta_0) \mathbb{X}_1 (\hat{\beta}_1 - \beta_{1,0})\|_\infty + \|\mathbb{X}_2^T \sum (\beta_0) \mathbb{X}_2 \hat{\beta}_2^p\|_\infty + O_P(n) \|\hat{\beta}^p - \beta_0\|_2^2.$$

It is implied that $\|\xi_2\|_\infty = o(n\lambda)$ from Condition B and $\eta\|\xi_2^p\|_\infty = o_P(n\lambda)$ from Condition D. In addition, $\|(1 + \eta)[\gamma_2(\hat{\beta}_1) - \gamma_2(\beta_{1,0})]\|_\infty = Cn\lambda + o_P(n\lambda)$ from Conditions B and D. Therefore, with probability tending to one, (A.7) holds with the solution $\hat{\beta}_1$ in Step 1.

8.5 Proof of Theorem 7

Recall from Condition F that the parameter space under consideration is restricted to $\mathbb{B}_\delta$. As $\hat{\beta}^{\lambda,\eta}$ minimizes (29) for $\beta \in \mathbb{B}_\delta$, the following inequality holds for any $\beta \in \mathbb{B}_\delta$,

$$L(\hat{\beta}^{\lambda,\eta}; \mathbb{X}, \mathbf{Y}) + \eta L(\hat{\beta}^{\lambda,\eta}; \mathbb{X}, \hat{\mathbf{Y}}^p) + \lambda \sum_{j=1}^p |\hat{\beta}_j^{\lambda,\eta}| \leq L(\beta; \mathbb{X}, \mathbf{Y}) + \eta L(\beta; \mathbb{X}, \hat{\mathbf{Y}}^p) + \lambda \sum_{j=1}^p |\beta_j|. \quad (\text{A.8})$$

With $R(\beta)$ defined in (31), by noticing that

$$\begin{aligned} L(\beta; \mathbb{X}, \mathbf{Y}) - R(\beta) &= -\frac{1}{n} \sum_{i=1}^n [Y_i - b'(\mathbf{X}_i^T \beta_0)] \mathbf{X}_i^T \beta, \\ L(\beta; \mathbb{X}, \hat{\mathbf{Y}}^p) - R(\beta) &= -\frac{1}{n} \sum_{i=1}^n [\hat{Y}_i^p - b'(\mathbf{X}_i^T \beta_0)] \mathbf{X}_i^T \beta, \end{aligned}$$

(A.8) can be rewritten as

$$(1 + \eta) \mathcal{E}(\hat{\beta}^{\lambda,\eta}) \leq (1 + \eta) \mathcal{E}(\beta) + I_1 + I_2 + I_3, \quad (\text{A.9})$$

where

$$I_1 = \frac{1}{n} \sum_{i=1}^n \varepsilon_i \mathbf{X}_i^T (\hat{\beta}^{\lambda,\eta} - \beta), \quad I_2 = \frac{\eta}{n} \sum_{i=1}^n \hat{\varepsilon}_i^p \mathbf{X}_i^T (\hat{\beta}^{\lambda,\eta} - \beta), \quad I_3 = \lambda \sum_{j=1}^p (|\beta_j| - |\hat{\beta}_j^{\lambda,\eta}|),$$

with ε_i and $\hat{\varepsilon}_i^p$ defined in (30). By direct calculations,

$$\begin{aligned} |I_1| &= \left| \sum_{j=1}^p \frac{1}{n} \sum_{i=1}^n \varepsilon_i X_{ij} (\hat{\beta}_j^{\lambda,\eta} - \beta_j) \right| \leq \sum_{j=1}^p |\hat{\beta}_j^{\lambda,\eta} - \beta_j| \max_{1 \leq j \leq p} \left| \frac{1}{n} \sum_{i=1}^n \varepsilon_i X_{ij} \right|, \\ |I_2| &= \left| \sum_{j=1}^p \frac{\eta}{n} \sum_{i=1}^n \hat{\varepsilon}_i^p X_{ij} (\hat{\beta}_j^{\lambda,\eta} - \beta_j) \right| \leq \sum_{j=1}^p |\hat{\beta}_j^{\lambda,\eta} - \beta_j| \max_{1 \leq j \leq p} \left| \frac{\eta}{n} \sum_{i=1}^n \hat{\varepsilon}_i^p X_{ij} \right|. \end{aligned}$$

Recall that $\lambda = An^{-1/2} (\sqrt{\log p} + \eta B_{\hat{\varepsilon}^p})$ where $B_{\hat{\varepsilon}^p} = n^{-1/2} \|\mathbf{x}^T \hat{\varepsilon}^p\|_\infty$. Consider the following two events,

$$\mathcal{A}_1 = \left\{ \max_{1 \leq j \leq p} \left| \frac{1}{n} \sum_{i=1}^n \varepsilon_i X_{ij} \right| \leq \frac{\lambda}{4} \right\}, \quad \mathcal{A}_2 = \left\{ \max_{1 \leq j \leq p} \left| \frac{\eta}{n} \sum_{i=1}^n \hat{\varepsilon}_i^p X_{ij} \right| \leq \frac{\lambda}{4} \right\}.$$

Our derivation will be restricted on $\mathcal{A}_1 \cap \mathcal{A}_2$ hereafter, and we will bound the probability of $(\mathcal{A}_1 \cap \mathcal{A}_2)^c$ at the end of the proof. On $\mathcal{A}_1 \cap \mathcal{A}_2$,

$$|I_1| \leq \frac{\lambda}{4} \sum_{j=1}^p |\hat{\beta}_j^{\lambda, \eta} - \beta_j|, \quad |I_2| \leq \frac{\lambda}{4} \sum_{j=1}^p |\hat{\beta}_j^{\lambda, \eta} - \beta_j|.$$

Then (A.9) leads to

$$(1+\eta)\mathcal{E}(\hat{\beta}^{\lambda, \eta}) + \frac{\lambda}{2} \sum_{j=1}^p |\hat{\beta}_j^{\lambda, \eta} - \beta_j| \leq (1+\eta)\mathcal{E}(\beta) + \lambda \sum_{j=1}^p (|\beta_j| - |\hat{\beta}_j^{\lambda, \eta}| + |\hat{\beta}_j^{\lambda, \eta} - \beta_j|). \quad (\text{A.10})$$

When $j \notin \mathcal{J}(\beta)$, $\beta_j = 0$, then $|\beta_j| - |\hat{\beta}_j^{\lambda, \eta}| + |\hat{\beta}_j^{\lambda, \eta} - \beta_j| = 0$. Then, (A.10) leads to

$$(1+\eta)\mathcal{E}(\hat{\beta}^{\lambda, \eta}) + \frac{\lambda}{2} \sum_{j=1}^p |\hat{\beta}_j^{\lambda, \eta} - \beta_j| \leq (1+\eta)\mathcal{E}(\beta) + 2\lambda \sum_{j \in \mathcal{J}(\beta)} |\hat{\beta}_j^{\lambda, \eta} - \beta_j|. \quad (\text{A.11})$$

Define another event $\mathcal{B} = \{2\lambda \sum_{j \in \mathcal{J}(\beta)} |\hat{\beta}_j^{\lambda, \eta} - \beta_j| \leq \varphi(1+\eta)\mathcal{E}(\beta)\}$ for the constant $\varphi > 0$ defined in Theorem 7. On the one hand, on event $\mathcal{B}$, (A.11) implies that

$$\mathcal{E}(\hat{\beta}^{\lambda, \eta}) \leq (1+\varphi)\mathcal{E}(\beta), \text{ for any } \beta \in \mathbb{B}_\delta.$$

Thus, the inequality (32) in Theorem 7 is proved on event $\mathcal{B}$.

On the other hand, on event $\mathcal{B}^c$, (A.11) leads to

$$\frac{\lambda}{2} \sum_{j=1}^p |\hat{\beta}_j^{\lambda, \eta} - \beta_j| \leq 2\lambda(1+1/\varphi) \sum_{j \in \mathcal{J}(\beta)} |\hat{\beta}_j^{\lambda, \eta} - \beta_j|. \quad (\text{A.12})$$

Denoted by $\delta = \hat{\beta}^{\lambda, \eta} - \beta$, (A.12) implies that

$$\frac{1}{2} \|\delta_{\mathcal{J}(\beta)^c}\|_1 + \frac{1}{2} \|\delta_{\mathcal{J}(\beta)}\|_1 \leq 2(1+1/\varphi) \|\delta_{\mathcal{J}(\beta)}\|_1, \text{ i.e., } \|\delta_{\mathcal{J}(\beta)^c}\|_1 \leq (3+4/\varphi) \|\delta_{\mathcal{J}(\beta)}\|_1.$$

Under Condition RE($r, 3 + 4/\varphi$), for any β satisfying $|\mathcal{J}(\beta)| \leq r$,

$$\kappa^2(r, 3+4/\varphi) \|\delta_{\mathcal{J}(\beta)}\|_2^2 \leq \frac{1}{n} \|\mathbb{X}\delta\|_2^2 = \frac{1}{n} \|\mathbb{X}\hat{\beta}^{\lambda, \eta} - \mathbb{X}\beta\|_2^2. \quad (\text{A.13})$$

From (A.11) and (A.13), we have that

$$\begin{aligned}
 (1+\eta)\mathcal{E}(\hat{\beta}^{\lambda,\eta}) &\leq (1+\eta)\mathcal{E}(\beta) + 2\lambda \sqrt{|J(\beta)|} \|\delta_{J(\beta)}\|_2 \\
 &\leq (1+\eta)\mathcal{E}(\beta) + \frac{2\lambda \sqrt{|J(\beta)|}}{\sqrt{n\kappa(r,3+4/\varphi)}} \|\mathbb{X}\hat{\beta}^{\lambda,\eta} - \mathbb{X}\beta\|_2.
 \end{aligned} \quad (\text{A.14})$$

Furthermore, Lemma 1 shows that, for $\beta \in \mathbb{B}_s$,

$$\mathcal{E}(\hat{\beta}^{\lambda,\eta}) \geq \frac{1}{n} \delta \|\mathbb{X}\hat{\beta}^{\lambda,\eta} - \mathbb{X}\beta_0\|_2^2 \quad \text{and} \quad \mathcal{E}(\beta) \geq \frac{1}{n} \delta \|\mathbb{X}\beta - \mathbb{X}\beta_0\|_2^2,$$

Then, (A.14) becomes

$$\begin{aligned}
 (1+\eta)\mathcal{E}(\hat{\beta}^{\lambda,\eta}) &\leq (1+\eta)\mathcal{E}(\beta) + \frac{2\lambda \sqrt{|J(\beta)|}}{\sqrt{\delta\kappa(r,3+4/\varphi)}} \left[\sqrt{\mathcal{E}(\hat{\beta}^{\lambda,\eta})} + \sqrt{\mathcal{E}(\beta)} \right] \\
 &\leq (1+\eta)\mathcal{E}(\beta) + 2b \frac{\lambda^2 |J(\beta)|}{\delta\kappa^2(r,3+4/\varphi)} + \frac{1}{b} \mathcal{E}(\hat{\beta}^{\lambda,\eta}) + \frac{1}{b} \mathcal{E}(\beta),
 \end{aligned} \quad (\text{A.15})$$

where we have applied the inequality $2uv \leq bu^2 + v^2/b$ for $b > 0$. With $b > 1/(1+\eta)$, (A.15) implies

$$\mathcal{E}(\hat{\beta}^{\lambda,\eta}) \leq \frac{b+b\eta+1}{b+b\eta-1} \mathcal{E}(\beta) + \frac{2b^2\lambda^2|J(\beta)|}{(b+b\eta-1)\delta\kappa^2(r,3+4/\varphi)}.$$

Setting $(b+b\eta+1)/(b+b\eta-1) = 1+\phi$ leads to $b = (2+\phi)/[\phi(1+\eta)]$ and therefore,

$$\begin{aligned}
 \mathcal{E}(\hat{\beta}^{\lambda,\eta}) &\leq (1+\phi) \left[\mathcal{E}(\beta) + \frac{(2+\phi)^2\lambda^2|J(\beta)|}{\phi(1+\eta)\delta(1+\eta)^2\kappa^2(r,3+4/\varphi)} \right] \\
 &\leq (1+\phi) \left(\mathcal{E}(\beta) + \frac{2C(\varphi)A^2|J(\beta)|}{\delta\kappa^2(r,3+4/\varphi)} \left[\frac{1}{(1+\eta)^2} \frac{\log p}{n} + \frac{\eta^2}{(1+\eta)^2} \frac{B_{\hat{\epsilon}p}^2}{n} \right] \right),
 \end{aligned}$$

for any $\beta \in \mathbb{B}_s$ satisfying that $|J(\beta)| \leq r$. Therefore, the inequality (32) in Theorem 7 is proved on event $\mathcal{B}^c$.

The last part of the proof is to bound the probability of $(\mathcal{A}_1 \cap \mathcal{A}_2)^c$. Recall that

$\lambda = An^{-1/2}(\sqrt{\log p} + \eta B_{\hat{\epsilon}p})$. With the same notation as in Condition E, define $M' = \max[M \sup_{n,i} \|\mathbf{X}_i\|_\infty, 2(M \sup_{n,i} \|\mathbf{X}_i\|_\infty)^2] < \infty$. For $\mathcal{A}_1$, applying Bernstein's inequality (Lemma 2.2.11 in van der Vaart and Wellner (1996)) leads to

$$\begin{aligned}
 P(\mathcal{A}_1^c) &\leq \sum_{j=1}^p P \left(\left| \sum_{i=1}^n \epsilon_i X_{ij} \right| > \frac{A}{4} \sqrt{n \log p} \right) \\
 &\leq 2p \exp \left\{ -\frac{1}{2} \frac{(A^2/16)n \log p}{M' [n+(A/4)\sqrt{n \log p}]} \right\} \\
 &\leq 2 \exp \left\{ -\frac{1}{2} \frac{(A^2/16-2M')n \log p - 2M' (A/4) \sqrt{n} (\log p)^{3/2}}{M' [n+(A/4)\sqrt{n \log p}]} \right\}.
 \end{aligned}$$

It is seen that $P(\mathcal{A}_1^c) \rightarrow 0$ when $A > 0$ is large enough and $\log p = o(n)$. In addition, for $\mathcal{A}_2$,

$$\max_{1 \leq j \leq p} \left| \frac{\eta}{n} \sum_{i=1}^n \hat{\varepsilon}_i^p X_{ij} \right| = n^{-1/2} \eta B_{\hat{\varepsilon}^p} \leq \frac{\lambda}{4},$$

as long as $A \geq 4$. Then $P(\mathcal{A}_2^c) = 0$. Therefore, $P[(\mathcal{A}_1 \cap \mathcal{A}_2)^c] \rightarrow 0$.

8.6 Proof of Corollary 1

In the proof of Theorem 7, setting $\beta = \beta_0$ in (A.11) leads to

$$\|\delta_{0,J_0^c}\|_1 \leq 3\|\delta_{0,J_0}\|_1, \quad (\text{A.16})$$

where $\delta_0 = \hat{\beta}^{\lambda,\eta} - \beta_0$, $J_0 = \mathcal{J}(\beta_0)$ with $|J_0| = s$. Thus, Condition RE($r, 3 + 4/\phi$) becomes RE($s, 3$) and $\kappa(r, 3 + 4/\phi)$ becomes $\kappa(s, 3)$ in the proof. Also $\mathcal{B} = \emptyset$ when $\beta = \beta_0$. Meanwhile, setting $\beta = \beta_0$ in (A.14) we have that

$$\mathcal{E}(\hat{\beta}^{\lambda,\eta}) \leq \frac{4\lambda^2 |J_0|}{\delta(1+\eta)^2 \kappa^2(s, 3)}.$$

In addition, by Lemma 1,

$$\frac{1}{n} \|\mathbb{X} \hat{\beta}^{\lambda,\eta} - \mathbb{X} \beta_0\|_2^2 \leq \frac{1}{\delta} \mathcal{E}(\hat{\beta}^{\lambda,\eta}) \leq \frac{4\lambda^2 |J_0|}{\delta^2(1+\eta)^2 \kappa^2(s, 3)}.$$

Furthermore, (A.16) implies that $\|\delta_0\|_1 \leq 4\|\delta_{0,J_0}\|_1 \leq 4\sqrt{|J_0|}\|\delta_{0,J_0}\|_2$. Therefore, under Condition RE($s, 3$),

$$\|\hat{\beta}^{\lambda,\eta} - \beta_0\|_1^2 \leq \frac{16|J_0|}{n\kappa^2(s, 3)} \|\mathbb{X} \hat{\beta}^{\lambda,\eta} - \mathbb{X} \beta_0\|_2^2 \leq \frac{64\lambda^2 |J_0|^2}{\delta^2(1+\eta)^2 \kappa^4(s, 3)}.$$

Therefore, all results in Corollary 1 are proved by noticing that

$$\lambda^2 \leq 2A^2 \left(\frac{\log p}{n} + \eta^2 \frac{B_{\hat{\varepsilon}^p}^2}{n} \right).$$

8.7 A Lemma

Lemma 1—Under Condition F , for any $\beta \in \mathbb{B}_\delta$,

$$\mathcal{E}(\beta) \geq \frac{1}{n} \delta \|\mathbb{X}\beta - \mathbb{X}\beta_0\|_2^2.$$

Proof: For $t \in [0, 1]$, consider the following function

$$g(t) = \frac{1}{n} \sum_{i=1}^n b(\mathbf{X}_i^T [t\beta + (1-t)\beta_0]).$$

Clearly, g is twice continuously differentiable. By Taylor's expansion,

$$g(1) - g(0) - g'(0) = \int_0^1 g''(t)(1-t) dt.$$

Note that the left-hand side of the above equation is just $\mathcal{E}(\beta)$, and the right-hand side is

$$\frac{1}{n} \sum_{i=1}^n [\mathbf{X}_i^T (\beta - \beta_0)]^2 \int_0^1 b''(\mathbf{X}_i^T [t\beta + (1-t)\beta_0])(1-t) dt \geq \frac{1}{n} \delta \sum_{i=1}^n [\mathbf{X}_i^T (\beta - \beta_0)]^2,$$

under Condition F . The proof is completed.

Figure 1.

An Example of pLASSO Solution Path. This path is illustrated using the diabetes data in Efron et al. (2004), with η on the horizontal axis and coefficient estimates on the vertical axis. There are 10 predictors in the diabetes data, each of which is standardized to have a unit norm. We choose a prior set $\mathcal{S}^p = \{X_5, X_6, X_7, X_8\}$. λ is fixed at 316.07 with an initial set $\mathcal{A}^{\lambda,0} = \{3, 4, 9\}$. Each vertical line in the plot represents a change of the active set, with the inclusion/deletion of a variable noted at the top of the plot, and the corresponding η value noted at the bottom. The last vertical line is hypothetical to show the limit of coefficients when $\eta \rightarrow \infty$.

Figure 2.

Simulation results of unweighted penalization methods for linear regression. Rows 1–3 correspond to prior sets $\mathcal{S}_1^P - \mathcal{S}_3^P$ in $\mathcal{G}_1$ respectively, and rows 4–6 correspond to prior sets $\mathcal{S}_4^P - \mathcal{S}_6^P$ in $\mathcal{G}_2$ respectively. Columns 1–4 correspond to #CNZ, #INZ, Bias and MSR, respectively.

Figure 3.

Simulation results of unweighted penalization methods for linear regression. Rows 1–3 correspond to prior sets $\mathcal{S}_7^P - \mathcal{S}_9^P$ in $\mathcal{G}_3$ respectively, and rows 4–6 correspond to prior sets $\mathcal{S}_{10}^P - \mathcal{S}_{12}^P$ in $\mathcal{G}_4$ respectively. Columns 1–4 correspond to #CNZ, #INZ, Bias and MSR, respectively.

Figure 4.

Simulation results of unweighted penalization methods for logistic regression. Rows 1–3 correspond to prior sets $\mathcal{S}_1^P - \mathcal{S}_3^P$ in $\mathcal{G}_1$ respectively, and rows 4–6 correspond to prior sets $\mathcal{S}_4^P - \mathcal{S}_6^P$ in $\mathcal{G}_2$ respectively. Columns 1–5 correspond to #CNZ, #INZ, Bias, RME and MR, respectively.

Figure 5. Simulation results of unweighted penalization methods for logistic regression. Rows 1–3 correspond to prior sets $\mathcal{S}_7^P - \mathcal{S}_9^P$ in $\mathcal{G}_3$ respectively, and rows 4–6 correspond to prior sets $\mathcal{S}_{10}^P - \mathcal{S}_{12}^P$ in $\mathcal{G}_4$ respectively. Columns 1–5 correspond to #CNZ, #INZ, Bias, RME and MR, respectively.

Figure 6. Optimal selection of the tuning parameter η by pLASSO in linear regression (upper panel) and logistic regression (lower panel).

Figure 7. Simulation results of weighted penalization methods for linear regression. Rows 1–3 correspond to prior sets $\mathcal{J}_1^P - \mathcal{J}_3^P$ in $\mathcal{G}_1$ respectively, and rows 4–6 correspond to prior sets $\mathcal{J}_4^P - \mathcal{J}_6^P$ in $\mathcal{G}_2$ respectively. Columns 1–4 correspond to #CNZ, #INZ, Bias and MSR, respectively.

Figure 8.

Simulation results of weighted penalization methods for linear regression. Rows 1–3 correspond to prior sets $\mathcal{S}_7^P - \mathcal{S}_9^P$ in $\mathcal{G}_3$ respectively, and rows 4–6 correspond to prior sets $\mathcal{S}_{10}^P - \mathcal{S}_{12}^P$ in $\mathcal{G}_4$ respectively. Columns 1–4 correspond to #CNZ, #INZ, Bias and MSR, respectively.

Figure 9.

Simulation results of weighted penalization methods for logistic regression. Rows 1–3 correspond to prior sets $\mathcal{S}_1^P - \mathcal{S}_3^P$ in $\mathcal{G}_1$ respectively, and rows 4–6 correspond to prior sets $\mathcal{S}_4^P - \mathcal{S}_6^P$ in $\mathcal{G}_2$ respectively. Columns 1–5 correspond to #CNZ, #INZ, Bias, RME and MR, respectively.

Figure 10.

Simulation results of weighted penalization methods for logistic regression. Rows 1–3 correspond to prior sets $\mathcal{S}_7^P - \mathcal{S}_9^P$ in $\mathcal{G}_3$ respectively, and rows 4–6 correspond to prior sets $\mathcal{S}_{10}^P - \mathcal{S}_{12}^P$ in $\mathcal{G}_4$ respectively. Columns 1–5 correspond to #CNZ, #INZ, Bias, RME and MR, respectively.

Table 1

Real Data Results. Top 20 SNPs with the largest effect size from LASSO and pLASSO separately.

SNP	LASSO				pLASSO			
	Chr.	Gene	OR	SNP	Chr.	Gene	OR	
Matched								
rs2226284	1	LRRC7	1.51	rs2226284	1	LRRC7	1.30	
rs6691577*	1	LRRC7	0.67	rs6691577*	1	LRRC7	0.77	
rs4407218*	2	-	1.25	rs4407218*	2	-	1.18	
rs10496366	2	-	0.81	rs10496366	2	-	0.85	
rs12640371	4	-	1.26	rs12640371	4	-	1.20	
rs4451017	5	-	1.28	rs4451017	5	-	1.22	
rs10041150	5	-	1.24	rs10041150	5	-	1.17	
rs2609653*	8	-	1.24	rs2609653*	8	-	1.17	
rs884672	10	SORCS3	0.78	rs884672	10	SORCS3	0.84	
rs10994548	10	RHOBTB1	0.81	rs10994548	10	RHOBTB1	0.85	
rs659991	13	DGKH	1.28	rs659991	13	DGKH	1.32	
rs7159947	14	PARP2	1.22	rs7159947	14	PARP2	1.17	
rs7176592	15	C15orf23	1.24	rs7176592	15	C15orf23	1.18	
rs12938916	17	MRPS23	1.36	rs12938916	17	MRPS23	1.28	
rs2837588	21	DSCAM	1.24	rs2837588	21	DSCAM	1.18	
Unmatched								
rs6668411	1	DAB1	0.82	rs12472797	2	-	0.86	
rs6895541	5	FBXL21	1.22	rs11103407	9	COL5A1	1.17	
rs17053171	6	-	0.81	rs10982246*	9	DFNB31	0.86	
rs6987445	8	-	1.22	rs17600642	10	ADAMTS14	1.17	
rs4984405	15	-	1.22	rs6574988	14	-	1.17	

The SNPs/genes in the prior set are in bold characters. Chr.: chromosome; Gene: within or nearby gene; OR: odds ratio estimate;

* identified in WTCCC (2007) with a strong or moderate association with bipolar disorder.