Statistics and Its Interface Volume 6 (2013) 285–299

High-dimensional regression and classification under a class of convex loss functions

Yuan Jiang* and Chunming Zhang†

The weighted L_1 penalty was used to revise the traditional Lasso in the linear regression model under quadratic loss. We make use of this penalty to investigate the high-dimensional regression and classification under a wide class of convex loss functions. We show that for the dimension growing nearly exponentially with the sample size, the penalized estimator possesses the oracle property for suitable weights, and its induced classifier is shown to be consistent to the optimal Bayes rule. Moreover, we propose two methods, called componentwise regression (CR) and penalized componentwise regression (PCR), for estimating weights. Both theories and simulation studies provide supporting evidence for the advantage of PCR over CR in high-dimensional regression and classification. The effectiveness of the proposed method is illustrated using real data sets.

Keywords and phrases: Convex loss, High-dimensional model, Optimal Bayes rule, Oracle property, Weighted L_1 penalty.

1. INTRODUCTION

Penalization was introduced to regularize overparameterized problems, and has succeeded in dealing with the challenges posed by analyzing high-dimensional data sets. L_1 penalty was used in Lasso by Tibshirani (1996) to simultaneously select variables and estimate parameters for a linear model with a fixed number of parameters. From then on, the L_1 penalty has exhibited its attractiveness in both theoretical and experimental perspectives by extensive studies (for example, Knight and Fu, 2000). However, Lasso can not select variables consistently without essential conditions (Meinshausen and Buhlmann, 2006; Zhao and Yu, 2006). To remedy this drawback, Zou (2006) introduced imposing adaptive weights on the L_1 penalties of the parameters, and showed that using the weighted L_1 penalties can enable consistent selection of variables under general conditions. Zou (2006) further showed that the resultant estimator achieves the oracle property (Donoho and Johnstone, 1994): it is asymptotically as efficient as the oracle estimator.

*Corresponding author.

†Zhang's research is supported in part by National Science Foundation grants DMS-07-05209 and DMS-1106586, and the Wisconsin Alumni Research Foundation.

As popular penalization methods, L_1 and weighted L_1 penalties have been further investigated in large/high-dimensional linear models and likelihood models. For the L_1 penalty, Meinshausen and Yu (2008) investigated the L_2-consistency of the Lasso estimator in linear models when the number of parameters p_n grows nearly exponentially with the sample size n; Zhang and Huang (2008) considered a rate consistency of the Lasso estimation when the high-dimensional linear model is sparse in the sense that most coefficients are small in absolute values. More recent contributions include Bach (2010) and Kakade et al. (2010). For the weighted L_1 penalty, Huang et al. (2008) applied the adaptive Lasso to a sparse high-dimensional linear model, and extended the oracle property of the penalized estimator. Other useful penalties with application to large/high-dimensional linear models and likelihood models include the smoothly clipped absolute deviation (SCAD) penalty discussed by Fan and Peng (2004) and Lv and Fan (2009).

However, as mentioned above, most research efforts on L_1 and weighted L_1 penalizations have been limited to linear models and likelihood models. One exception is van der Geer (2008), which dealt with Lasso in high-dimensional models with Lipschitz loss functions. However, that work focused on the prediction error of the Lasso estimator instead of the performance of the estimator itself. Other exceptions include Belloni and Chernozhukov (2009) and Zhou et al. (2009). However, they are particularly interested in quantile regression and gaussian graphical modeling respectively. Recently, to generalize the conventional penalized likelihood, Zhang et al. (2010) introduced penalized Bregman divergence. It used the concept of Bregman divergence, which unifies nearly all of the commonly used loss functions in the regression analysis and classification procedure (Zhang et al., 2009). For instance, an important application of Bregman divergence is the quasi-likelihood model (Wedderburn, 1974) which is popular when the underlying distribution of the observations is not fully specified. Zhang et al. (2010) studied the statistical properties of the penalized Bregman divergence estimator in conjunction with either nonconvex or convex penalties. The dimension p_n in that work has either a smaller or nearly the same order as the sample size n, depending on the choice of penalties. However, the high-dimensional setting, where p_n can grow faster than n, was not investigated by Zhang et al. (2010).

Therefore, it remains an open problem to broaden the scope of penalization in a high-dimensional setting. To fill

this gap, we study the penalized estimator and classifier for high-dimensional data in this work, where p_n can grow nearly exponentially with the sample size n. Compared with Zhang et al. (2010), we focus on those loss functions belonging to the class of convex Bregman divergence. We also restrict our investigations to the weighted L_1 penalization. Using convex loss functions and weighted L_1 penalties facilitates the derivation of the statistical properties for the penalized estimator and classifier in a high-dimensional setting. From the regression viewpoint, in high-dimensional regression models utilizing convex Bregman divergence as loss functions, the weighted L_1 penalized estimator is shown to possess the oracle property for suitable weights. From the classification viewpoint, the classifier induced by the penalized estimator is shown to be asymptotically consistent to the optimal Bayes rule.

Meanwhile, we discuss how to estimate the adaptive weights. The marginal regression method in Huang et al. (2008) is generalized to componentwise regression (CR) in our framework, which can provide satisfactory estimates of the weights under appropriate conditions. Furthermore, we propose a novel weight estimation method, called penalized componentwise regression (PCR). Compared with CR, PCR needs weaker conditions to estimate satisfactory weights. The numerical studies in this paper illustrate the application of the weighted L_1 penalty to high-dimensional regression and classification. The results verify the advantage of the weighted L_1 penalty over the L_1 penalty in model fitting, variable selection, and classification. The numerical results also indicate that PCR is preferable to CR for estimating the adaptive weights in practice.

The rest of the paper is organized as follows. Section 2 establishes the weighted L_1 penalization under convex Bregman divergence. We discuss the statistical properties of the penalized estimator for appropriate weights there. Section 3 studies the application of the weighted L_1 penalization to classification. Section 4 proposes two methods, CR and PCR, for estimating weights. Section 5 presents the results from simulations and real data examples. All technical details are included in the Appendix and supplemental materials (http://www.intlpress.com/SII/p/2013/6-2/SII-6-2-jiang-supplement.pdf).

2. REGRESSION UNDER CONVEX LOSS FUNCTIONS

In this section, we introduce the penalized Bregman divergence estimator similar to that in Zhang et al. (2010), and provide the oracle property of the weighted L_1 penalized estimator in a high-dimension setting.

2.1 Penalized Bregman divergence

Assume the set of training samples is given by $\mathcal{T}_n = \{(\mathtt{X}_{n1}, Y_{n1}), \ldots, (\mathtt{X}_{nn}, Y_{nn})\}$, observed independently from a common probability distribution. Let $(\mathtt{X}_n, Y_n)$ be the generic pair of a random realization from this distribution, where $\mathtt{X}_n = (X_1, \ldots, X_{p_n})^T$ is the input random vector and Y_n is the output random variable. We assume the following underlying model for the training data,

$$m(\mathtt{X}_n) = E(Y_n|\mathtt{X}_n) = F^{-1}(\beta_{n,0;0} + \mathtt{X}_n^T \boldsymbol{\beta}_{n;0}). \tag{1}$$

In (1), F is a known link function, $\beta_{n,0;0} \in \mathbb{R}^1$ and $\boldsymbol{\beta}_{n;0} = (\beta_{n,1;0}, \ldots, \beta_{n,p_n;0})^T \in \mathbb{R}^{p_n}$ are the unknown true parameters. Some parameters in $\boldsymbol{\beta}_{n;0}$ are assumed to be exactly zero, and we write $\boldsymbol{\beta}_{n;0} = (\boldsymbol{\beta}_{n;0}^{(\mathrm{I})T}, \boldsymbol{\beta}_{n;0}^{(\mathrm{II})T})^T$ without loss of generality, in which $\boldsymbol{\beta}_{n;0}^{(\mathrm{I})}$ is the part of nonzero parameters and $\boldsymbol{\beta}_{n;0}^{(\mathrm{II})} = \mathbf{0}$. s_n is used to denote the number of nonzero parameters, i.e., the length of $\boldsymbol{\beta}_{n;0}^{(\mathrm{I})}$.

Similar to Zhang et al. (2010), define the weighted L_1 penalized Bregman divergence estimator to be the minimizer of the following criterion function,

$$\ell_n(\beta_{n,0}, \boldsymbol{\beta}_n) = \frac{1}{n}\sum_{i=1}^{n} Q(Y_{ni}, F^{-1}(\beta_{n,0} + \mathtt{X}_{ni}^T \boldsymbol{\beta}_n)) + \lambda_n \sum_{j=1}^{p_n} w_{n,j}|\beta_{n,j}|, \tag{2}$$

where $\boldsymbol{\beta}_n = (\beta_{n,1}, \ldots, \beta_{n,p_n})^T$, $\lambda_n > 0$ is the tuning parameter, and $w_{n,1}, \ldots, w_{n,p_n}$ are the nonnegative weights for parameters $\beta_{n,1}, \ldots, \beta_{n,p_n}$. The loss function $Q(\cdot,\cdot)$ in (2) is the Bregman divergence proposed by Bregman (1967),

$$Q(\nu, \mu) = -q(\nu) + q(\mu) + (\nu - \mu)q'(\mu),$$

where q is a given concave function.

In general, $Q(Y, \mu)$ serves as a loss function for a random variable Y and its estimation μ. A large family of commonly used loss functions in regression and classification are Bregman divergence with suitably chosen generating functions q. For example, $q(\mu) = a\mu - \mu^2$ with a constant a results in the quadratic loss $Q(Y, \mu) = (Y - \mu)^2$. For a binary output variable Y, $q(\mu) = -\{\mu \log(\mu) + (1 - \mu)\log(1 - \mu)\}$ yields the deviance loss $Q(Y, \mu) = -\{Y \log(\mu) + (1 - Y)\log(1 - \mu)\}$; $q(\mu) = 2\{\mu(1 - \mu)\}^{1/2}$ generates the exponential loss $Q(Y, \mu) = \exp[-(Y - 1/2)\log\{\mu/(1-\mu)\}]$. We refer to Zhang et al. (2010) and the references therein for more details about the application of Bregman divergence.

For $\theta = F(\mu)$, define $\mathrm{q}_j(y; \theta) = (\partial^j/\partial\theta^j)Q(y, F^{-1}(\theta))$ for $j = 1, 2, \ldots$. Then,

$$\mathrm{q}_1(y; \theta) = (y - \mu)q''(\mu)/F'(\mu),$$
$$\mathrm{q}_2(y; \theta) = -q''(\mu)/\{F'(\mu)\}^2 + (y - \mu)A(\mu),$$

where $A(\mu) = \{q'''(\mu)F'(\mu) - q''(\mu)F''(\mu)\}/\{F'(\mu)\}^3$. The loss function Q in (2) is convex with respect to the parameters $\beta_{n,0}$ and $\boldsymbol{\beta}_n$ given that $\mathrm{q}_2(y; \theta) > 0$. In practice, a lot of commonly used loss functions are actually convex, e.g., quadratic loss with identity link for continuous output

variables, deviance loss or exponential loss with logit link for binary output variables, deviance loss with log link for counting output variables, etc. Therefore, we restrict our investigations to convex loss functions throughout this work.

2.2 Oracle property of the penalized estimator

Let $\widetilde{\boldsymbol{\beta}}_n = (\beta_{n,0}, \boldsymbol{\beta}_n^T)^T$, and correspondingly $\widetilde{\mathtt{X}}_n = (1, \mathtt{X}_n^T)^T$. Then, the criterion function (2) can be written as

(3)
$$\ell_n(\widetilde{\boldsymbol{\beta}}_n) = \frac{1}{n}\sum_{i=1}^{n} Q(Y_{ni}, F^{-1}(\widetilde{\mathtt{X}}_{ni}^T\widetilde{\boldsymbol{\beta}}_n)) + \lambda_n \sum_{j=1}^{p_n} w_{n,j}|\beta_{n,j}|,$$

and the penalized estimator is denoted by $\widehat{\widetilde{\boldsymbol{\beta}}}_n = (\widehat{\beta}_{n,0}, \widehat{\boldsymbol{\beta}}_n^T)^T$. In addition, we write the parameters $\widetilde{\boldsymbol{\beta}}_n$ into two parts as $\widetilde{\boldsymbol{\beta}}_n^{(\mathrm{I})} = (\beta_{n,0}, \beta_{n,1}, \ldots, \beta_{n,s_n})^T$ and $\boldsymbol{\beta}_n^{(\mathrm{II})} = (\beta_{n,s_n+1}, \ldots, \beta_{n,p_n})^T$. Accordingly, $\widetilde{\mathtt{X}}_n^{(\mathrm{I})} = (1, X_1, \ldots, X_{s_n})^T$ and $\mathtt{X}_n^{(\mathrm{II})} = (X_{s_n+1}, \ldots, X_{p_n})^T$.

Before presenting the oracle property of the penalized Bregman divergence estimator, we introduce some necessary notation as follows: first,

$$\mathbf{H}_{n;0} = -E\left[\frac{q''(m(\mathtt{X}_n))}{\{F'(m(\mathtt{X}_n))\}^2}\widetilde{\mathtt{X}}_n^{(\mathrm{I})}\widetilde{\mathtt{X}}_n^{(\mathrm{I})T}\right],$$
$$\boldsymbol{\Omega}_{n;0} = E\left[\mathrm{var}(Y_n|\mathtt{X}_n)\frac{\{q''(m(\mathtt{X}_n))\}^2}{\{F'(m(\mathtt{X}_n))\}^2}\widetilde{\mathtt{X}}_n^{(\mathrm{I})}\widetilde{\mathtt{X}}_n^{(\mathrm{I})T}\right];$$

second, with $\|\cdot\|$ and $\|\cdot\|_\infty$ denoting the L_2 and L_∞ norm respectively, define

$$\mathbf{P}_n(\widetilde{\boldsymbol{\beta}}_n^{(\mathrm{I})}) = \frac{1}{n}\sum_{i=1}^{n} \mathrm{q}_2(Y_{ni}; \widetilde{\mathtt{X}}_{ni}^{(\mathrm{I})T}\widetilde{\boldsymbol{\beta}}_n^{(\mathrm{I})})\mathtt{X}_{ni}^{(\mathrm{II})}\widetilde{\mathtt{X}}_{ni}^{(\mathrm{I})T},$$
$$\rho_n = \sup\{\|\mathbf{P}_n(\widetilde{\boldsymbol{\beta}}_n^{(\mathrm{I})})\mathbf{u}\|_\infty : \|\mathbf{u}\| = 1, \|\widetilde{\boldsymbol{\beta}}_n^{(\mathrm{I})} - \widetilde{\boldsymbol{\beta}}_{n;0}^{(\mathrm{I})}\| \le \log(n)\sqrt{s_n/n}\};$$

last, write the weights for $\widetilde{\boldsymbol{\beta}}_n^{(\mathrm{I})}$ as a diagonal matrix $\mathbf{W}_n = \mathrm{diag}(0, w_{n,1}, \ldots, w_{n,s_n})$, and define

$$w_{\max}^{(\mathrm{I})} = \max_{1\le j\le s_n} w_{n,j} \quad \text{and} \quad w_{\min}^{(\mathrm{II})} = \min_{s_n+1\le j\le p_n} w_{n,j}$$

to be the maximum weight for the nonzero parameters and the minimum weight for the zero parameters respectively.

Theorem 1. *Suppose* $s_n^5/n \to 0$ *and* $\log(p_n - s_n)/\min\{n, n\lambda_n^2(w_{\min}^{(\mathrm{II})})^2\} = o_P(1)$ *as* $n \to \infty$. *Assume Conditions* 1–4 *in the Appendix, and further* $w_{\max}^{(\mathrm{I})} = O_P\{1/(\lambda_n\sqrt{n})\}$, $(w_{\min}^{(\mathrm{II})})^{-1} = o_P\{\lambda_n\sqrt{n}/(\rho_n\sqrt{s_n})\}$. *With probability tending to one, there exists a global minimizer* $\widehat{\widetilde{\boldsymbol{\beta}}}_n$ *of* ℓ_n *in* (3) *which satisfies that*

(1) $\widehat{\boldsymbol{\beta}}_n^{(\mathrm{II})} = \mathbf{0}$,

(2) $\sqrt{n}A_n\boldsymbol{\Omega}_{n;0}^{-1/2}[\mathbf{H}_{n;0}(\widehat{\widetilde{\boldsymbol{\beta}}}_n^{(\mathrm{I})} - \widetilde{\boldsymbol{\beta}}_{n;0}^{(\mathrm{I})}) + \lambda_n\mathbf{W}_n\mathrm{sign}(\widetilde{\boldsymbol{\beta}}_{n;0}^{(\mathrm{I})})] \xrightarrow{\mathcal{L}} N(\mathbf{0}, G)$ *for any* $k\times(s_n+1)$ *matrix* A_n *such that* $A_nA_n^T \to G$ *with* G *being a* $k\times k$ *semi-positive definite matrix, where* $\mathrm{sign}(\widetilde{\boldsymbol{\beta}}_{n;0}^{(\mathrm{I})}) = \{\mathrm{sign}(\beta_{n,0;0}), \mathrm{sign}(\beta_{n,1;0}), \ldots, \mathrm{sign}(\beta_{n,s_n;0})\}^T$.

Theorem 1 provides the oracle property of the penalized Bregman divergence estimator under a class of convex loss functions. As mentioned in Section 2.1, the convexity of loss functions is assumed through Condition 1(b) in the Appendix. To illustrate more clearly the conditions and conclusions of this theorem, we make a few comments as follows.

First, ρ_n indicates how strongly the relevant and irrelevant input variables are correlated. For the purpose of illustration, let us consider the case of linear regression with the quadratic loss and identity link. In this case, $\rho_n = \sup\{\|\mathbf{P}_n\mathbf{u}\|_\infty : \|\mathbf{u}\| = 1\}$ with $\mathbf{P}_n = (2/n)\sum_{i=1}^n \mathtt{X}_{ni}^{(\mathrm{II})}\widetilde{\mathtt{X}}_{ni}^{(\mathrm{I})T}$. Suppose that the input variables are centralized and standardized. $\mathbf{P}_n$ is proportional to the sample correlation matrix between the relevant and irrelevant input variables. The sample correlation matrix plays a central role in the "irrepresentable condition" in Zhao and Yu (2007) and the "adaptive irrepresentable condition" in Huang et al. (2008).

Moreover, in the above case of linear regression with input variables centralized and standardized,

$$\rho_n \le \max_{s_n+1\le j\le p_n}\left\|\frac{2}{n}\sum_{i=1}^{n} X_{ij}\widetilde{\mathtt{X}}_{ni}^{(\mathrm{I})}\right\| \le (s_n+1)^{1/2}\max_{0\le k\le s_n, s_n+1\le j\le p_n}\left|\frac{2}{n}\sum_{i=1}^{n} X_{ij}X_{ik}\right|.$$

Specifically, if the relevant and irrelevant input variables are uncorrelated, under Condition 3 in the Appendix, $\rho_n = O_P([s_n\log\{s_n(p_n - s_n)\}/n]^{1/2})$ by Bernstein's inequality (Lemma 2.2.9 in van der Vaart and Wellner, 1996). It is noteworthy that Huang et al. (2008) also imposed a constraint on $\max_{0\le k\le s_n, s_n+1\le j\le p_n}|\frac{1}{n}\sum_{i=1}^n X_{ij}X_{ik}|$ in their "partial orthogonality condition".

Second, following the discussion of ρ_n as above, we assume in Theorem 1 the condition $(w_{\min}^{(\mathrm{II})})^{-1} = o_P\{\lambda_n\sqrt{n}/(\rho_n\sqrt{s_n})\}$ for the weights. This condition describes the following relationship: the stronger the correlations between the relevant and irrelevant input variables, the larger the minimum weight for irrelevant input variables needs to be. However, regardless of the correlation structure of the input variables, we can always prove that $\rho_n = O_P(\sqrt{s_n})$ under the conditions in Theorem 1 (the proof is in the Appendix). Therefore, $(w_{\min}^{(\mathrm{II})})^{-1} = o_P(\lambda_n\sqrt{n}/s_n)$ is always a sufficient condition for Theorem 1.

Third, regarding the dimension of the data, there are two conditions in Theorem 1. On the one hand, the condition $s_n^5/n \to 0$ imposes the constraint on the number of relevant input variables. However, it is worth mentioning that this

condition can be weakened to $s_n^2/n \to 0$ for part (1) of Theorem 1 (see Lemma 3 in the Appendix). On the other hand, the condition $\log(p_n - s_n)/\min\{n, n\lambda_n^2 (w_{\min}^{(\mathrm{II})})^2\} = o_P(1)$ imposes the constraint on the number of irrelevant input variables. Taking $s_n = n^{c_1}$ with $0 < c_1 < 1/5$ as an example, $w_{\min}^{(\mathrm{II})}$ can be $n^{c_2}/(\lambda_n \sqrt{n})$ with some $c_2 > c_1$ as mentioned above. Then, the constraint becomes $\log(p_n - s_n) = o\{\min(n^{2c_2}, n)\}$. This indicates that the allowed number of irrelevant input variables can grow almost exponentially fast of $\min(n^{2c_2}, n)$.

Last, the asymptotic distribution of the penalized Bregman divergence estimator depends on the Q-loss only through the second derivative of its generating q-function. The asymptotic covariance matrix of the penalized estimator is given by $\mathbf{H}_{n;0}^{-1} \mathbf{\Omega}_{n;0} \mathbf{H}_{n;0}^{-1}$ where both $\mathbf{H}_{n;0}$ and $\mathbf{\Omega}_{n;0}$ involve $q''(\cdot)$. It can be shown that this covariance matrix can achieve its lowest bound when the generating q-function satisfies the so-called generalized Bartlett identity (Bartlett, 1953; Zhang et al., 2010). This provides an insight of how different loss functions can impact the asymptotic behavior of their penalized estimators. We refer to the equation (11) in Zhang et al. (2010) and the discussions thereafter for details.

2.3 Comparison with a previous result

Huang et al. (2008) obtained the oracle property of the penalized estimator using weighted L_1 penalty in the framework of linear models. Since their quadratic loss function belongs to the class of convex loss functions discussed in our work, we would like to see how our result relates to theirs. Both results provide the oracle property of the penalized estimator, so we compare key conditions under which this conclusion is derived.

First, for the error terms $\epsilon_n = Y_n - m(\mathtt{X}_n)$, we restrict ϵ_n in Condition 3 in the Appendix; while the condition in Huang et al. (2008) is their condition (A1):

$$P(|\epsilon_n| > t) \le K \exp(-Ct^d) \quad \text{for } t \ge 0,$$

with certain constants $1 \le d \le 2$, $C > 0$ and $K > 0$. By the proof of Lemma 2.2.1 in van der Vaart and Wellner (1996), it is seen that their condition (A1) implies that $E\{\exp(D|\epsilon_n|^d)\} \le 2$ for $D = C/(1+K)$, and this further concludes that $E(|\epsilon_n|^{ld}) \le l!(1/D)^l$ for $l = 1, 2, \ldots$. Consequently by Liapounov's inequality (Shao, 2003, page 30), $E(|\epsilon_n|^l) \le (l!)^{1/d}\{(1/D)^{1/d}\}^l$ for $l = 1, 2, \ldots$. This implies our condition for ϵ_n, since $1 \le d \le 2$. Therefore, our condition for the error terms is slightly weaker than their condition (A1).

Next, for the allowed number of irrelevant input variables $p_n - s_n$, we have two conditions $\log(p_n - s_n)/n = o(1)$ and $\log(p_n - s_n)/\{n\lambda_n^2 (w_{\min}^{(\mathrm{II})})^2\} = o_P(1)$. Since Huang et al. (2008) imposed a constraint $p_n = O(\exp(n^a))$ for a constant $0 < a < 1$, which is stronger than our first condition, we only focus on the second. As mentioned in Huang et al. (2008), their model can include the most covariates when the error terms have a sub-Gaussian tail [$d = 2$ in their condition (A1)]. When $d = 2$ in (A1), they imposed condition (A4):

$$\log(p_n - s_n)(M_{n2} + r_n^{-1})^2/(n\lambda_n^2) \to 0,$$

where M_{n2} controls the proxies $\eta_{n,j}$ of the true parameters $\beta_{n,j;0}$, and r_n describes the asymptotic rate of the difference between the initial estimators $\widehat{\beta}_{n,j}^*$ and these proxies. M_{n2} and r_n satisfy their condition (A2):

$$r_n \max_{1 \le j \le p_n} |\widehat{\beta}_{n,j}^* - \eta_{n,j}| = O_P(1), \quad r_n \to \infty,$$
$$\max_{s_n+1 \le j \le p_n} |\eta_{n,j}| \le M_{n2}.$$

According to (A2) and their definition $w_{n,j} = |\widehat{\beta}_{n,j}^*|^{-1}$,

$$(w_{\min}^{(\mathrm{II})})^{-1} = \max_{s_n+1 \le j \le p_n} |\widehat{\beta}_{n,j}^*| = O_P(M_{n2} + r_n^{-1}).$$

This, together with their condition (A4), implies that $\log(p_n - s_n)/\{n\lambda_n^2 (w_{\min}^{(\mathrm{II})})^2\} = o_P(1)$, exactly the same as our second condition about $p_n - s_n$. This observation suggests that our conditions for $p_n - s_n$ are weaker than those in Huang et al. (2008).

Last, it is slightly difficult to compare the allowed number of relevant input variables s_n. Our work requires that $s_n^2/n \to 0$ for the variable selection consistency, and $s_n^5/n \to 0$ for the oracle property of the penalized estimator; while in their work s_n depends on the above-mentioned r_n and λ_n, and has a lower order than n. This observation underscores the necessity of sparsity in high-dimensional regression.

3. APPLICATION TO CLASSIFICATION

In binary classification problems, Y_n only takes values 0 and 1. In this case, the penalized estimator $(\widehat{\beta}_{n,0}, \widehat{\boldsymbol{\beta}}_n^T)^T$ defined as the minimizer of (2) naturally induces the following classifier for a future input variable $\mathtt{x}_n$,

$$\widehat{\phi}_n(\mathtt{x}_n) = \mathrm{I}\{F^{-1}(\widehat{\beta}_{n,0} + \mathtt{x}_n^T \widehat{\boldsymbol{\beta}}_n) > 1/2\}. \tag{4}$$

In classification literature, the misclassification loss by a classification rule ϕ at a data point $(\mathtt{x}, y)$ is defined as $l(y, \phi(\mathtt{x})) = \mathrm{I}\{y \ne \phi(\mathtt{x})\}$. The risk of ϕ is the expected misclassification loss $R(\phi) = E\{l(Y, \phi(\mathtt{X}))\} = P\{\phi(\mathtt{X}) \ne Y\}$. The optimal Bayes rule, which minimizes the risk, is $\phi_{\mathrm{B}}(\mathtt{x}) = \mathrm{I}\{m(\mathtt{x}) > 1/2\}$.

The Bayes rule is denoted by $\phi_{n,\mathrm{B}}(\mathtt{x}_n) = \mathrm{I}\{m(\mathtt{x}_n) > 1/2\}$ in our setting. For a test sample $(\mathtt{X}_n, Y_n)$, which is an i.i.d. copy of the samples in the training set $\mathcal{T}_n$, the optimal Bayes risk is $R(\phi_{n,\mathrm{B}}) = P\{\phi_{n,\mathrm{B}}(\mathtt{X}_n) \ne Y_n\}$. Meanwhile, the conditional risk of the classifier $\widehat{\phi}_n$ is given by $R(\widehat{\phi}_n) = P\{\widehat{\phi}_n(\mathtt{X}_n) \ne Y_n | \mathcal{T}_n\}$.

A rule $\widehat{\phi}_n$ is called consistent if its conditional risk converges to the optimal Bayes risk in the sense that

$$E\{R(\widehat{\phi}_n)\} - R(\phi_{n,\mathrm{B}}) \to 0.$$

Theorem 2. *Under the conditions of Theorem 1, the classifier in (4) is consistent to the Bayes rule.*

Similar to Theorem 9 in Zhang et al. (2010), Theorem 2 verifies the classification consistency attained by the penalized Bregman divergence classifier. We omit its proof and refer to Zhang et al. (2010) for details.

4. ESTIMATION OF WEIGHTS

In this section, we explore how to estimate the weights $\{w_{n,j} : j = 1, \ldots, p_n\}$ before we apply the weighted L_1 penalization to high-dimensional regression and classification. For simplicity of notation, we denote $\gamma_n^{(\mathrm{I})} = \lambda_n\sqrt{n}$ and $\gamma_n^{(\mathrm{II})} = \lambda_n\sqrt{n}/(\rho_n\sqrt{s_n})$ hereafter.

4.1 Componentwise regression

Huang et al. (2008) applied marginal regression to estimate the weights in high-dimensional linear models. We show that their method can be generalized to our setting with a class of convex loss functions. Following Zhang et al. (2010), we call this generalization "componentwise regression (CR)".

In CR, an initial estimator $\widehat{\boldsymbol{\beta}}_n^{\mathrm{CR}}$ is computed to minimize the componentwise regression criterion function,

$$\ell_n^{\mathrm{CR}}(\boldsymbol{\beta}_n) = \frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{p_n} Q(Y_{ni}, F^{-1}(X_{ij}\beta_{n,j})). \tag{5}$$

Then, the weights are estimated by

$$\widehat{w}_{n,j} = |\widehat{\beta}_{n,j}^{\mathrm{CR}}|^{-1}, \quad j = 1, \ldots, p_n. \tag{6}$$

It is noteworthy that CR is the same as marginal regression when Q is the quadratic loss and F is the identity link. In addition to estimating weights in penalization methods, marginal regression is also very useful in high-dimensional variable screening (Fan and Lv, 2008; Fan et al., 2009; Fan and Song, 2010).

With the estimated weights $\{\widehat{w}_{n,j} : j = 1, \ldots, p_n\}$, define $\widehat{w}_{\max}^{(\mathrm{I})} = \max_{1\le j\le s_n}\widehat{w}_{n,j}$ and $\widehat{w}_{\min}^{(\mathrm{II})} = \min_{s_n+1\le j\le p_n}\widehat{w}_{n,j}$ to be the estimates of $w_{\max}^{(\mathrm{I})}$ and $w_{\min}^{(\mathrm{II})}$ respectively. Theorem 3 justifies the applicability of CR under certain conditions, with its proof included in the supplemental materials (http://www.intlpress.com/SII/p/2013/6-2/SII-6-2-jiang-supplement.pdf).

Theorem 3. *Suppose* $\gamma_n^{(\mathrm{I})} = O(1)$, $\sqrt{n}\gamma_n^{(\mathrm{I})} \to \infty$, $\sqrt{n}\gamma_n^{(\mathrm{II})} \to \infty$, $\log(s_n) = o(n\gamma_n^{(\mathrm{I})2})$ *and* $\log(p_n - s_n) = o\{\min(n\gamma_n^{(\mathrm{II})}, n\gamma_n^{(\mathrm{II})2})\}$. *Assume that* $E(\boldsymbol{X}_n) = \mathbf{0}$. *Under Conditions 1–6 in the Appendix, where* $\mathcal{A}_n = \gamma_n^{(\mathrm{I})}$ *and* $\mathcal{B}_n = \gamma_n^{(\mathrm{II})}$ *in Condition 5, the estimates* $\widehat{w}_{n,j}$ *in (6) satisfy that* $\widehat{w}_{\max}^{(\mathrm{I})} = O_P(1/\gamma_n^{(\mathrm{I})})$ *and* $(\widehat{w}_{\min}^{(\mathrm{II})})^{-1} = o_P(\gamma_n^{(\mathrm{II})})$ *as needed in Theorem 1.*

Condition 5 in the Appendix imposes a requirement on the marginal correlations between input variables and the output variable. Roughly, $\mathcal{A}_n$ represents the minimum absolute marginal correlation for the relevant input variables; $\mathcal{B}_n$ represents the maximum absolute marginal correlation for the irrelevant input variables. It seems natural to assume $\mathcal{B}_n = o(\mathcal{A}_n)$. However, it may not be essential. In Theorem 3, $\mathcal{A}_n = \gamma_n^{(\mathrm{I})}$ and $\mathcal{B}_n = \gamma_n^{(\mathrm{II})}$. Depending on ρ_n, $\mathcal{B}_n$ can have a larger order than $\mathcal{A}_n$ (e.g., when $\rho_n\sqrt{s_n} \to 0$). This implies that the irrelevant variables can have stronger correlations with the output variable than the relevant variables, given that the correlations between the irrelevant and relevant variables are weak enough.

To approximately achieve the condition $E(\boldsymbol{X}_n) = \mathbf{0}$ in Theorem 3, in application, the data can always be preprocessed by centralizing the input variables $X_1, \ldots, X_{p_n}$. This technique has been regularly used in previous works (e.g., Zou, 2006 and Huang et al., 2008).

4.2 Penalized componentwise regression

Following the discussion after Theorem 3, Condition 5 is relatively strong when $\mathcal{B}_n = o(\mathcal{A}_n)$. That is, the relevant variables have to be more strongly correlated with the output variable than the irrelevant variables. So, a natural question arises: can we relax Condition 5 by improving the weight estimation method? To this end, we propose an alternative method named "penalized componentwise regression (PCR)" as follows.

In PCR, we compute the initial estimator $\widehat{\boldsymbol{\beta}}_n^{\mathrm{PCR}}$ which minimizes a penalized version of the componentwise regression criterion function,

$$\ell_n^{\mathrm{PCR}}(\boldsymbol{\beta}_n) = \frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{p_n} Q(Y_{ni}, F^{-1}(X_{ij}\beta_{n,j})) + \kappa_n\sum_{j=1}^{p_n}|\beta_{n,j}|. \tag{7}$$

The weight estimates $\widehat{w}_{n,j}$ are then given by

$$\widehat{w}_{n,j} = |\widehat{\beta}_{n,j}^{\mathrm{PCR}}|^{-1}, \quad j = 1, \ldots, p_n. \tag{8}$$

We present the theoretical property of PCR in two different cases. Case (1): PCR is mainly proposed to weaken Condition 5 when $\mathcal{B}_n = o(\mathcal{A}_n)$. As $\mathcal{A}_n = \gamma_n^{(\mathrm{I})}$ and $\mathcal{B}_n = \gamma_n^{(\mathrm{II})}$ in Theorem 3, we impose the restriction that $\gamma_n^{(\mathrm{II})} = o(\gamma_n^{(\mathrm{I})})$ (i.e., $\rho_n\sqrt{s_n} \to \infty$) and see how PCR can weaken Condition 5. Case (2): PCR is not limited to be only applicable to Case (1) where $\gamma_n^{(\mathrm{II})} = o(\gamma_n^{(\mathrm{I})})$. We provide the justification for PCR in a general case without the restriction as well. Theorem 4 presents the applicability of PCR in

both cases, with its proof included in the supplemental materials (http://www.intlpress.com/SII/p/2013/6-2/SII-6-2-jiang-supplement.pdf).

Theorem 4. *Suppose* $\gamma_n^{(\mathrm{I})} = O(1)$, $\kappa_n = o(\gamma_n^{(\mathrm{I})})$ *and* $\log(s_n) = o(n\gamma_n^{(\mathrm{I})2})$. *Assume that* $E(\boldsymbol{X}_n) = \mathbf{0}$. *Assume Conditions* 1(*a*) *and* 2–5 *in the Appendix, where* $\mathcal{A}_n = \gamma_n^{(\mathrm{I})}$ *in Condition* 5.

Case (1)*: If* $\gamma_n^{(\mathrm{II})} = o(\gamma_n^{(\mathrm{I})})$, *we assume that* $\gamma_n^{(\mathrm{II})} = o(\kappa_n)$, $\mathcal{B}_n = O(\kappa_n)$ *and* $\log(p_n - s_n) = o(n\kappa_n^2)$.

Case (2)*: In general, we assume Conditions* 1(*b*) *and* 6, $\mathcal{B}_n = O\{\max(\kappa_n, \gamma_n^{(\mathrm{II})})\}$, $\log(p_n - s_n) = o(n)$ *and* $\log(p_n - s_n) = o\{\max(n\kappa_n^2, n\gamma_n^{(\mathrm{II})2})\}$.

Then, in either case (1) *or case* (2), *there exist local minimizers* $\widehat{\beta}_{n,j}^{\mathrm{PCR}}$ *in* (7) *such that the corresponding estimates* $\widehat{w}_{n,j}$ *in* (8) *satisfy that* $\widehat{w}_{\max}^{(\mathrm{I})} = O_P(1/\gamma_n^{(\mathrm{I})})$ *and* $(\widehat{w}_{\min}^{(\mathrm{II})})^{-1} = o_P(\gamma_n^{(\mathrm{II})})$ *as needed in Theorem* 1.

We present the comparison between CR and PCR separately for case (1) and case (2). In case (1), first, Condition 5 is weakened. Compared with Theorem 3, the condition of $\mathcal{B}_n$ is relaxed to $\mathcal{B}_n = O(\kappa_n)$ from $\mathcal{B}_n = \gamma_n^{(\mathrm{II})}$, since $\gamma_n^{(\mathrm{II})} = o(\kappa_n)$ in this case. Second, Condition 1(b) is not required, which means PCR is applicable beyond the convex loss function framework. Third, PCR allows more input variables to be included in the model than CR. This is observed by comparing the conditions on $p_n - s_n$ in the two theorems.

In case (2), first, we observe that the restriction $\gamma_n^{(\mathrm{II})} = o(\gamma_n^{(\mathrm{I})})$ can be removed if we only consider convex loss functions. In this framework where CR is applicable, PCR is always applicable regardless of the relationship between $\gamma_n^{(\mathrm{I})}$ and $\gamma_n^{(\mathrm{II})}$. Second, the conditions regarding $\mathcal{B}_n$ and $p_n - s_n$ can also be weaker than those in Theorem 3, depending on the choice of κ_n.

In application, we treat κ_n in (7) as a tuning parameter, and use a tuning set or cross validation to select the optimal one. It is validated by our experimental studies that PCR is preferable to CR for estimating the weights, as we will see in Section 5.

5. NUMERICAL STUDIES

This section includes numerical studies which intend to evaluate the performance of high-dimensional regression and classification under different convex loss functions. We focus on the comparison between L_1 and weighted L_1 penalization methods. Meanwhile, the adaptive weights are estimated using both CR and PCR. The algorithm for the optimization with L_1 and weighted L_1 penalties has been studied thoroughly in the literature (Osborne et al., 2000; Efron et al., 2004; Rosset and Zhu, 2007), so we do not include the details here.

5.1 Simulations

This subsection contains two simulation studies: one for linear models which use quadratic loss, and the other for binary response models which use deviance loss and exponential loss. In addition to L_1 and weighted L_1 penalization methods, we also include a naive method using the adaptive weights for completeness, i.e., selecting a proportion of variables with top smallest weights. The proportion of variables selected by the naive method is regarded as a tuning parameter.

5.1.1 Quadratic loss

The data are generated from the following model,

$$\begin{aligned} &\mathtt{X}_n = (X_1, \ldots, X_{p_n})^T \sim N(\mathbf{0}, \boldsymbol{\Sigma}), \\ &\epsilon_n \sim N(0,1), \\ &Y_n = \beta_{n,0;0} + \mathtt{X}_n^T \boldsymbol{\beta}_{n;0} + \epsilon_n, \end{aligned}$$

where $\mathtt{X}_n$ is independent of ϵ_n, $\beta_{n,0;0} = 2.5$ and $\boldsymbol{\beta}_{n;0} = (2.5, 2.5, 1, 1, 1, 0, \ldots, 0)^T$. We set that $n = 100$, $s_n = 5$, $p_n = 500$ or $1,000$. The covariance matrix $\boldsymbol{\Sigma} = (\sigma_{ij})$ used to generate $\mathtt{X}_n$ is chosen as $\sigma_{ij} = \rho + (1-\rho)\,\mathrm{I}(i=j)$ (type I) or $\sigma_{ij} = \rho^{|i-j|}$ (type II), for $i = 1, \ldots, p_n$ and $j = 1, \ldots, p_n$.

In each setting, 100 sets of training data are generated. First, each set is normalized before computation so that the mean and standard deviation for each input variable X_j across the 100 samples are 0 and 1 respectively. Then, we compute the penalized estimators using the normalized training set. Last, the resultant estimators are transformed back to the original location and scale. The tuning parameters λ_n and κ_n in (3) and (7) are searched on a surface of grid points, and selected by minimizing the residual sum of squares evaluated on an independently generated tuning set with the same size as the training set. The proportion of variables selected by the naive method is also tuned similarly to λ_n and κ_n. The following results are evaluated in each setting.

Result I. Model fitting criterion MME. The model error (ME) is approximated by $\frac{1}{5000}\sum_{l=1}^{5000}\{\widehat{m}(\mathtt{x}_{nl}) - m(\mathtt{x}_{nl})\}^2$ at a sequence $\{\mathtt{x}_{nl}\}_{l=1}^{5000}$ simulated independently from the training set. We report the median of the MEs (MME) from the 100 training samples.

Result II. Variable selection criteria #CZ and #CNZ. We record the number of parameters which are correctly identified as zero when their true values are zero, and the number of parameters which are correctly identified as nonzero when their true values are nonzero. Reported are the averages (#CZ and #CNZ) from the 100 training samples.

Tables 1 and 2 summarize the simulation results when $p_n = 500$ and $p_n = 1,000$ respectively, from which we can draw the following conclusions.

First, the weighted L_1 penalty outperforms the L_1 penalty, in both model fitting and variable selection. The

Table 1. Simulation results for quadratic loss. $p_n = 500$

Penalty	MME	#CZ	#CNZ	MME	#CZ	#CNZ
	$\rho = 0$					
L_1	0.417	465.96	6.00			
weighted L_1 (CR)	0.280	482.51	5.95			
weighted L_1 (PCR)	0.258	482.36	5.95			
naive (CR)	1.270	481.18	5.14			
naive (PCR)	1.270	481.24	5.14			
	$\rho = 0.1$ (type I)			$\rho = 0.1$ (type II)		
L_1	0.366	471.53	6.00	0.375	470.04	6.00
weighted L_1 (CR)	0.241	479.01	6.00	0.197	485.53	5.98
weighted L_1 (PCR)	0.225	479.45	6.00	0.184	486.51	5.98
naive (CR)	1.124	478.25	5.23	0.456	482.86	5.55
naive (PCR)	1.122	478.28	5.23	0.441	482.97	5.55
	$\rho = 0.5$ (type I)			$\rho = 0.5$ (type II)		
L_1	0.352	471.83	6.00	0.232	478.96	6.00
weighted L_1 (CR)	0.314	476.01	6.00	0.093	491.63	6.00
weighted L_1 (PCR)	0.281	476.53	6.00	0.068	493.08	6.00
naive (CR)	1.400	480.06	4.49	0.081	492.08	6.00
naive (PCR)	1.400	480.06	4.49	0.071	492.45	6.00
	$\rho = 0.9$ (type I)			$\rho = 0.9$ (type II)		
L_1	0.332	472.00	5.66	0.137	486.43	6.00
weighted L_1 (CR)	0.324	473.29	5.68	0.072	492.13	6.00
weighted L_1 (PCR)	0.313	474.67	5.60	0.067	492.28	6.00
naive (CR)	0.495	486.05	3.96	0.074	492.63	5.97
naive (PCR)	0.495	486.05	3.96	0.075	492.59	5.97

Table 2. Simulation results for quadratic loss. $p_n = 1000$

Penalty	MME	#CZ	#CNZ	MME	#CZ	#CNZ
	$\rho = 0$					
L_1	0.566	960.07	6.00			
weighted L_1 (CR)	0.408	978.88	5.91			
weighted L_1 (PCR)	0.390	979.51	5.91			
naive (CR)	1.564	980.98	4.91			
naive (PCR)	1.564	981.04	4.91			
	$\rho = 0.1$ (type I)			$\rho = 0.1$ (type II)		
L_1	0.502	961.29	6.00	0.489	964.59	6.00
weighted L_1 (CR)	0.364	973.28	6.00	0.294	981.95	6.00
weighted L_1 (PCR)	0.338	973.76	6.00	0.261	982.57	6.00
naive (CR)	1.498	979.13	4.92	0.704	982.19	5.42
naive (PCR)	1.498	979.12	4.92	0.704	982.15	5.43
	$\rho = 0.5$ (type I)			$\rho = 0.5$ (type II)		
L_1	0.529	962.41	6.00	0.277	976.35	6.00
weighted L_1 (CR)	0.467	967.64	6.00	0.115	991.15	6.00
weighted L_1 (PCR)	0.440	967.80	6.00	0.078	991.94	6.00
naive (CR)	1.638	980.76	4.11	0.068	992.93	5.97
naive (PCR)	1.638	980.76	4.11	0.068	992.94	5.97
	$\rho = 0.9$ (type I)			$\rho = 0.9$ (type II)		
L_1	0.429	964.92	5.15	0.165	983.57	6.00
weighted L_1 (CR)	0.408	965.83	5.14	0.083	992.43	6.00
weighted L_1 (PCR)	0.421	966.79	5.13	0.075	992.58	6.00
naive (CR)	0.558	984.71	3.66	0.063	993.75	6.00
naive (PCR)	0.558	984.71	3.66	0.063	993.66	6.00

Table 3. Simulation results for deviance loss. $p_n = 500$

Penalty	MME	MMR	#CZ	#CNZ	MME	MMR	#CZ	#CNZ
	$\rho = 0$							
L_1	5.51	19.3	472.46	4.69				
weighted L_1 (CR)	4.60	17.9	486.23	4.08				
weighted L_1 (PCR)	4.07	17.0	490.61	3.73				
naive (CR)	3.90	16.5	494.46	3.18				
naive (PCR)	3.84	16.5	494.50	3.16				
	$\rho = 0.1$ (type I)				$\rho = 0.1$ (type II)			
L_1	5.05	17.3	472.60	4.56	5.03	18.1	471.41	5.02
weighted L_1 (CR)	4.27	16.1	485.10	4.31	4.24	16.9	486.26	4.50
weighted L_1 (PCR)	4.00	15.8	487.31	4.16	3.85	16.0	490.42	4.19
naive (CR)	4.18	16.0	493.87	3.23	3.94	16.1	494.52	3.48
naive (PCR)	4.09	15.9	493.77	3.29	3.93	16.1	494.45	3.52
	$\rho = 0.5$ (type I)				$\rho = 0.5$ (type II)			
L_1	3.84	13.5	476.06	3.74	3.72	13.8	472.61	5.38
weighted L_1 (CR)	3.77	13.3	482.21	3.58	2.04	11.5	489.08	5.38
weighted L_1 (PCR)	3.69	13.1	483.58	3.54	1.69	11.1	492.18	5.37
naive (CR)	6.75	16.6	492.25	2.37	1.89	11.3	494.60	4.78
naive (PCR)	6.13	16.2	491.94	2.52	1.87	11.3	494.56	4.80
	$\rho = 0.9$ (type I)				$\rho = 0.9$ (type II)			
L_1	1.72	9.1	481.98	1.81	2.35	10.1	478.81	4.77
weighted L_1 (CR)	1.88	9.4	485.37	1.62	0.88	8.2	492.39	4.92
weighted L_1 (PCR)	1.82	9.2	483.68	1.71	0.80	8.0	493.53	5.15
naive (CR)	3.44	10.9	492.11	1.23	1.07	8.3	494.73	4.57
naive (PCR)	3.06	10.5	491.94	1.26	0.97	8.3	494.72	4.59

weighted L_1 penalized estimator achieves smaller model errors, and it can correctly identify more zero parameters. In other words, using the weighted L_1 penalty fits a more accurate model, and it tends to exclude more irrelevant variables. Both methods show strong abilities to identify nonzero parameters in our simulation study, i.e., both can include almost all relevant variables in the model.

The naive method which just selects variables with top smallest weights, however, is much worse than penalization methods in model fitting in most settings. For variable selection, the naive method tends to select a smaller model than penalization methods, keeping fewer relevant variables but excluding more irrelevant ones.

Second, when applied in the weighted L_1 penalization method, PCR achieves better results in model fitting than CR, resulting in smaller MMEs. Furthermore, PCR identifies slightly more zero parameters correctly than CR, which means that PCR can exclude more irrelevant variables from the model.

5.1.2 Deviance/exponential loss

The data are generated from

$$\mathtt{x}_n = (X_1, \ldots, X_{p_n})^T \sim N(\mathbf{0}, \boldsymbol{\Sigma}),$$
$$Y_n | \mathtt{x}_n = \mathtt{x}_n \sim \text{Bernoulli}(m(\mathtt{x}_n)).$$

The F-link function employed is $\text{logit}(m(\mathtt{x}_n)) = \beta_{n,0;0} + \mathtt{x}_n^T \boldsymbol{\beta}_{n;0}$, where $\beta_{n,0;0} = 2.5$ and $\boldsymbol{\beta}_{n;0} = (2.5, 2.5, 1, 1, 1, 0, \ldots, 0)^T$. The other settings are the same as in the previous subsection.

The simulation procedure is similar to that in Section 5.1.1. 100 sets of training data are generated. The penalized estimators are calculated from each normalized training set and then transformed back to the original location and scale. In calculation, both deviance loss and exponential loss are used. The tuning parameters λ_n, κ_n, and the proportion of variables selected by the naive method are all chosen by minimizing the empirical loss evaluated on a tuning set. Besides **Result I** and **Result II** (as in Section 5.1.1), we additionally report the following criterion.

Result III. Classification criterion MMR. We calculate the misclassification rate (MR) by evaluating the classifier on an independently generated test set with size 10,000. Reported are the medians of the MRs (MMR) from the 100 training samples.

Tables 3–6 present the simulation results for the two loss functions and two choices of p_n respectively. MME and MMR are recorded by their percentages in these tables (with the sign % omitted for conciseness). These results show the same patterns as the results in Tables 1–2 for comparing the performance in model fitting and variable selection. So we focus on comparing the performance in classification.

First, a comparison of the results indicates that the weighted L_1 penalty outperforms the L_1 penalty in classification, as a smaller misclassification rate is achieved by

Table 4. Simulation results for exponential loss. $p_n = 500$

Penalty	MME	MMR	#CZ	#CNZ	MME	MMR	#CZ	#CNZ
	$\rho = 0$							
L_1	5.06	18.8	478.68	4.39				
weighted L_1 (CR)	4.41	17.6	489.06	3.93				
weighted L_1 (PCR)	3.84	16.8	491.91	3.74				
naive (CR)	3.83	16.6	494.58	3.13				
naive (PCR)	3.85	16.6	494.58	3.09				
	$\rho = 0.1$ (type I)				$\rho = 0.1$ (type II)			
L_1	4.78	17.3	479.49	4.37	4.99	18.1	479.08	4.59
weighted L_1 (CR)	4.29	16.4	488.36	3.94	4.36	17.1	489.50	4.14
weighted L_1 (PCR)	4.12	16.1	489.56	3.86	3.98	16.4	491.89	3.90
naive (CR)	4.24	16.1	494.48	2.99	3.99	16.3	494.64	3.21
naive (PCR)	4.21	16.0	494.54	2.97	3.94	16.2	494.63	3.25
	$\rho = 0.5$ (type I)				$\rho = 0.5$ (type II)			
L_1	3.90	13.5	479.36	3.65	3.13	13.1	480.61	5.33
weighted L_1 (CR)	4.08	13.5	484.91	3.36	2.04	11.5	491.83	5.13
weighted L_1 (PCR)	4.08	13.4	485.69	3.29	1.78	11.3	493.51	5.13
naive (CR)	7.48	17.8	493.40	2.16	2.52	11.8	494.81	4.27
naive (PCR)	6.67	16.8	493.24	2.32	2.44	11.6	494.77	4.26
	$\rho = 0.9$ (type I)				$\rho = 0.9$ (type II)			
L_1	1.89	9.1	482.86	1.71	1.90	9.6	482.89	4.82
weighted L_1 (CR)	2.13	9.6	487.58	1.59	1.06	8.3	493.41	4.77
weighted L_1 (PCR)	1.97	9.3	485.91	1.68	0.93	8.1	493.62	5.03
naive (CR)	4.31	11.7	493.33	1.21	1.54	8.9	494.68	4.13
naive (PCR)	3.53	11.1	492.83	1.22	1.56	8.9	494.64	4.03

Table 5. Simulation results for deviance loss. $p_n = 1,000$

Penalty	MME	MMR	#CZ	#CNZ	MME	MMR	#CZ	#CNZ
	$\rho = 0$							
L_1	6.11	20.2	970.12	4.33				
weighted L_1 (CR)	5.08	18.3	985.80	3.89				
weighted L_1 (PCR)	4.31	17.3	990.99	3.48				
naive (CR)	3.82	16.6	994.69	2.99				
naive (PCR)	3.82	16.6	994.70	2.95				
	$\rho = 0.1$ (type I)				$\rho = 0.1$ (type II)			
L_1	5.77	18.7	972.05	4.21	5.88	19.1	971.07	4.57
weighted L_1 (CR)	4.83	17.4	984.04	4.04	4.66	17.5	986.39	4.22
weighted L_1 (PCR)	4.81	17.2	987.31	3.81	4.13	16.7	990.83	3.91
naive (CR)	4.49	16.5	993.70	3.05	4.04	16.4	994.54	3.21
naive (PCR)	4.49	16.5	993.83	3.01	4.01	16.4	994.55	3.22
	$\rho = 0.5$ (type I)				$\rho = 0.5$ (type II)			
L_1	4.27	14.2	972.65	3.28	4.35	14.7	970.59	5.17
weighted L_1 (CR)	4.08	14.1	980.02	3.15	2.56	12.4	988.94	5.18
weighted L_1 (PCR)	4.09	13.7	981.28	3.18	1.90	11.4	992.62	5.18
naive (CR)	7.06	17.2	992.10	2.10	2.53	12.1	994.68	4.44
naive (PCR)	6.54	17.0	992.05	2.17	2.53	12.1	994.70	4.43
	$\rho = 0.9$ (type I)				$\rho = 0.9$ (type II)			
L_1	1.70	9.1	980.39	1.64	2.72	10.7	975.49	4.61
weighted L_1 (CR)	1.92	9.4	984.93	1.49	1.07	8.4	991.93	4.78
weighted L_1 (PCR)	1.79	9.2	983.30	1.55	0.85	8.2	993.30	4.99
naive (CR)	3.74	10.8	991.37	1.20	1.37	8.7	994.86	4.37
naive (PCR)	3.16	10.5	991.80	1.26	1.39	8.7	994.87	4.25

Table 6. Simulation results for exponential loss. $p_n = 1,000$

Penalty	MME	MMR	#CZ	#CNZ	MME	MMR	#CZ	#CNZ
	$\rho = 0$							
L_1	5.89	20.2	978.29	4.07				
weighted L_1 (CR)	4.89	18.3	988.38	3.68				
weighted L_1 (PCR)	4.31	17.4	991.87	3.53				
naive (CR)	4.02	16.7	994.52	3.06				
naive (PCR)	3.96	16.7	994.62	2.98				
	$\rho = 0.1$ (type I)				$\rho = 0.1$ (type II)			
L_1	5.05	17.9	975.49	4.24	5.35	18.9	977.54	4.41
weighted L_1 (CR)	4.53	16.9	987.15	3.81	4.33	17.1	988.64	3.94
weighted L_1 (PCR)	4.39	16.5	989.02	3.77	4.03	16.5	991.67	3.79
naive (CR)	4.09	16.0	994.10	3.03	4.04	16.4	994.56	3.27
naive (PCR)	4.04	16.0	994.15	3.03	4.05	16.5	994.61	3.16
	$\rho = 0.5$ (type I)				$\rho = 0.5$ (type II)			
L_1	4.11	14.0	977.45	3.43	3.73	14.1	975.15	5.25
weighted L_1 (CR)	4.14	13.8	983.79	3.21	2.40	12.1	990.59	5.04
weighted L_1 (PCR)	4.01	13.8	984.52	3.20	1.95	11.4	992.87	5.01
naive (CR)	7.55	17.9	992.72	2.04	2.75	12.6	994.68	4.19
naive (PCR)	6.73	16.8	992.57	2.19	2.75	12.5	994.66	4.20
	$\rho = 0.9$ (type I)				$\rho = 0.9$ (type II)			
L_1	1.82	9.2	982.40	1.62	2.21	10.0	980.39	4.82
weighted L_1 (CR)	2.14	9.7	987.37	1.42	1.23	8.6	993.17	4.57
weighted L_1 (PCR)	1.96	9.5	985.23	1.53	1.07	8.3	993.85	4.88
naive (CR)	4.20	11.4	992.16	1.20	1.84	9.1	994.77	4.15
naive (PCR)	3.73	11.0	992.56	1.20	1.76	9.1	994.77	3.95

using the weighted L_1 penalty. In addition, the naive method performs reasonably well when the correlations among input variables are weak or moderate. However, when the correlations increase, its performance is derogated obviously and becomes worse than either penalization method.

Second, when applied in the weighted L_1 penalization method, PCR achieves further advantages in classification compared with CR, since PCR possesses slightly smaller misclassification rates.

Third, there is no significant evidence for the different impacts caused by different loss functions on the high-dimensional regression and classification in this simulation. The penalized estimator/classifier performs similarly under deviance loss and exponential loss.

5.2 Real data

This subsection illustrates high-dimensional regression and classification with convex loss functions using two real data sets—the MNIST data and the lymphoma data.

5.2.1 MNIST data

The MNIST database, one of the most famous databases in digit recognition, was created by LeCun et al. (1998). The handwritten digits were size normalized and centered in a $28 \times 28 = 784$ pixel image. The resultant image contains grey levels in each pixel. Thus each image can be regarded as a 784 dimensional vector of grey levels, and each pixel's information is used as an input variable. The digit categories of the images (digit 0 to digit 9) are regarded as the output variable.

We only use the samples with digit 6 or digit 9 in this study, since we are focusing on binary classification problem (the same two digits were chosen in Wang et al., 2006). The MNIST dataset contains $6,876$ ($5,918$ training and 958 test) samples of digit 6 and $6,958$ ($5,949$ training and $1,009$ test) samples of digit 9. We randomly choose $2n_1$ "balanced" samples ("balanced" means that n_1 samples of digit 6 and n_1 samples of digit 9) from the training set, which form our training set; again we randomly choose $2n_2$ "balanced" test samples to form our test set. Since we consider high-dimensional models in this study, we choose n_1 to be 25, 50 and 75 in our experiment. Also, n_2 is set to be 800. The data are normalized before further investigation, so that the mean and standard deviation for each input variable across the samples are 0 and 1 respectively.

The penalized estimators are all obtained from the training set only, and their corresponding classifiers are then evaluated on the test set. The tuning parameters λ_n and κ_n are selected by minimizing the misclassification rate with 3-fold cross validation of the training set. The following results of the estimators/classifiers are evaluated.

Result I'. Cross validation error. When cross validation is used to select λ_n and κ_n, for each fold of samples, a classifier is built using the other two folds, and a misclassification rate can be evaluated on the validation fold. Cross validation

error is the average of these three misclassification rates, using the optimal tuning parameters.

Result II'. Test error. A classifier is obtained from regression estimates using the training sample. Test error is the misclassification rate by evaluating this classifier on the test set.

Result III'. Number of selected variables. It is the number of the relevant pixels whose coefficients are estimated as nonzero.

The final results are tabulated in Table 7, with "CVE" for cross validation error, "TE" for test error, and "# pixels" for the number of selected pixels. The results of the penalized estimators/classifiers are similar under deviance loss and exponential loss. Using weighted L_1 penalty with PCR performs the best in terms of CVE and TE, and most time it selects the fewest pixels. However, with the increasing size of the training set, the differences between the three methods become smaller.

5.2.2 Lymphoma data

Alizadeh et al. (2000) identified two molecularly distinct forms of diffuse large B-cell lymphoma (DLBCL) by studying the lymphoma data. These two forms of DLBCL, called "germinal centre B-like DLBCL" and "activated B-like DLBCL", had gene expression patterns indicative of different stages of B-cell differentiation.

The publicly available dataset contains $4,026$ genes across 47 samples, of which 24 are germinal centre B-like DLBCL and 23 are activated B-like DLBCL. However, there are a few missing values in the data, so we use the k-NN (k-nearest neighbors) method to impute the missing expression data. After imputing, the data are normalized so that the mean and standard deviation for each gene across the 47 samples are 0 and 1 respectively.

We randomly divide the data into a training set with 31 samples (16 cases of germinal centre B-like DLBCL and 15 cases of activated B-like DLBCL) and a test set with 16 samples (8 cases of germinal centre B-like DLBCL and 8 cases of activated B-like DLBCL). The penalized estimators are all obtained from the training set only, and their corresponding classifiers are then evaluated on the test set. The tuning parameters λ_n and κ_n are selected by 3-fold cross validation with the training set. The same types of results as **Result I'**, **Result II'** and **Result III'** in the analysis of MNIST data are recorded.

100 random training/test splits of the whole data are performed and the above procedure is repeated to record 100 CVEs, 100 TEs and 100 numbers of selected genes. Table 8 tabulates the median of CVEs, the median of TEs and the average number of selected genes. The penalized estimators/classifiers perform consistently for two loss functions. Weighted L_1 penalty with PCR performs the best in terms of CVE and TE, and approximately only half of the number of genes are selected by weighted L_1 penalty compared with L_1 penalty.

APPENDIX A. CONDITIONS

We present the conditions for our main theoretical results. Some of the conditions below are purely technical and serve only to provide theoretical understanding of the newly proposed methodology. We have no intent to make the conditions the weakest possible. Throughout the Appendix, $\|\cdot\|$ is used only for the L_2 norm, and $\|\cdot\|_1$ and $\|\cdot\|_\infty$ denote the L_1 norm and the L_∞ norm respectively.

Condition 1. Condition 1 consists of two parts: 1(a) and 1(b).

1(a). q and F are smooth functions, satisfying that $q''(\cdot) < 0$ and $F'(\cdot) \neq 0$.

1(b). $\mathrm{q}_2(y;\theta) > 0$ for all $\theta \in \mathbb{R}$ and all y in the range of Y_n.

Condition 2. $\sup_{n\geq 1} \|\boldsymbol{\beta}^{(\mathrm{I})}_{n;0}\|_1 < \infty$, and $\sqrt{n/s_n}\beta^{(\mathrm{I})}_{\min} \to \infty$ where $\beta^{(\mathrm{I})}_{\min} = \min_{1\leq j\leq s_n} |\beta_{n,j;0}|$.

Condition 3. $\sup_{n\geq 1} \|\mathtt{X}_n\|_\infty = B_X < \infty$; for a constant $H > 0$, $\sup_{n\geq 1} E\{|Y_n - m(\mathtt{X}_n)|^l\} \leq l! H^l$ for $l = 2, 3, 4, \ldots$.

Condition 4. Define

$$\mathbf{H}_n(\widetilde{\boldsymbol{\beta}}^{(\mathrm{I})}_n) = E[\mathrm{q}_2(Y_n; \widetilde{\mathtt{X}}^{(\mathrm{I})T}_n \widetilde{\boldsymbol{\beta}}^{(\mathrm{I})}_n)\widetilde{\mathtt{X}}^{(\mathrm{I})}_n \widetilde{\mathtt{X}}^{(\mathrm{I})T}_n].$$

The eigenvalues of $\mathbf{H}_n(\widetilde{\boldsymbol{\beta}}^{(\mathrm{I})}_n)$ are uniformly bounded away from 0 for any $\widetilde{\boldsymbol{\beta}}^{(\mathrm{I})}_n$ satisfying that $\|\widetilde{\boldsymbol{\beta}}^{(\mathrm{I})}_n - \widetilde{\boldsymbol{\beta}}^{(\mathrm{I})}_{n;0}\| \leq \log(n)\sqrt{s_n/n}$. Meanwhile, the eigenvalues of $\boldsymbol{\Omega}_{n;0}$ are uniformly bounded away from 0.

Condition 5. For two nonnegative sequences s_{n1} and s_{n2}, we use $s_{n1} \succeq s_{n2}$ to denote that there exists a constant $c > 0$ such that $s_{n1} \geq c s_{n2}$ for all $n \geq 1$. Then, $\min_{1\leq j\leq s_n} |E(X_j Y_n)| \succeq \mathcal{A}_n$ and $\max_{s_n+1\leq j\leq p_n} |E(X_j Y_n)| = o(\mathcal{B}_n)$ for two positive sequences $\mathcal{A}_n$ and $\mathcal{B}_n$.

Condition 6. With a constant $\epsilon > 0$, $\inf_{n, s_n+1\leq j\leq p_n} E\{\mathrm{q}_2(Y_n; \gamma^{(\mathrm{II})}_n \epsilon_j X_j) X_j^2\} \geq \eta > 0$ for any constants $\epsilon_j \in (0, \epsilon)$ with $s_n + 1 \leq j \leq p_n$.

APPENDIX B. SOME TECHNICAL LEMMAS

This section presents some technical lemmas that are used in proving the main results.

B.1 Lemma 1

Lemma 1. *Under Conditions* 1(*a*), 2 *and* 3, *there exist some positive constants* C_1 *and* C_2 *both independent of* j *such that, with* $t > 0$,

$$P\left(\left|\sum_{i=1}^n q_1(Y_{ni}; \widetilde{\boldsymbol{X}}^T_{ni}\widetilde{\boldsymbol{\beta}}_{n;0})X_{ij}\right| > t\right) \leq 2\exp\left(\frac{-t^2}{C_1 n + C_2 t}\right),$$

for $1 \leq j \leq p_n$.

Table 7. MNIST data results

n_1	Penalty	deviance loss			exponential loss		
		CVE	TE	# pixels	CVE	TE	# pixels
25	L_1	4.17%	9.63%	11	4.17%	9.63%	11
	weighted L_1 (CR)	4.17%	7.75%	5	6.25%	7.88%	7
	weighted L_1 (PCR)	0.00%	7.88%	3	0.00%	7.94%	4
50	L_1	3.03%	7.69%	25	3.03%	7.75%	22
	weighted L_1 (CR)	4.04%	6.63%	19	4.04%	8.25%	14
	weighted L_1 (PCR)	2.02%	6.25%	14	2.02%	6.25%	15
75	L_1	2.00%	2.56%	37	2.00%	2.62%	34
	weighted L_1 (CR)	1.33%	2.88%	27	2.67%	3.06%	21
	weighted L_1 (PCR)	2.00%	2.50%	29	2.67%	2.50%	27

Table 8. Lymphoma data results

Penalty	deviance loss			exponential loss		
	CVE	TE	# genes	CVE	TE	# genes
L_1	16.7%	12.5%	16.04	13.3%	12.5%	13.73
weighted L_1 (CR)	3.33%	6.25%	7.75	3.33%	12.5%	7.17
weighted L_1 (PCR)	3.33%	6.25%	8.36	0.00%	12.5%	7.72

Proof. This lemma is derived directly from Bernstein's inequality (Lemma 2.2.11 in van der Vaart and Wellner, 1996). Under Conditions 1(a), 2 and 3, $\sup_{n\geq 1}|m(\mathtt{X}_n)| < \infty$, then $\sup_{n\geq 1}|(q''/F')(m(\mathtt{X}_n))| = A < \infty$. Let $Z_{ij} = \mathrm{q}_1(Y_{ni}; \widetilde{\mathtt{X}}_{ni}^T\widetilde{\boldsymbol{\beta}}_{n;0})X_{ij}$. Then $Z_{1j},\ldots,Z_{nj}$ are i.i.d. with mean 0, and the moment condition in Bernstein's inequality is satisfied as

$$E(|Z_{ij}|^l) \leq l!(AHB_X)^l \leq l!M^{l-2}v_i/2, \quad l = 2,3,4,\ldots,$$

with $M = AHB_X$ and $v_i \equiv 2(AHB_X)^2$. So, Berstein's inequality implies that

$$P\left(\left|\sum_{i=1}^n Z_{ij}\right| > t\right) \leq 2\exp\left\{\frac{-t^2}{2(nv_1+Mt)}\right\}.$$

Lemma 1 is proved by setting $C_1 = 2v_1$ and $C_2 = 2M$. □

B.2 Lemma 2

We define an "oracle subproblem" of (3) as

(B.1)
$$\ell_n^O(\widetilde{\boldsymbol{\beta}}_n^{(\mathrm{I})}) = \frac{1}{n}\sum_{i=1}^n Q(Y_{ni}, F^{-1}(\widetilde{\mathtt{X}}_{ni}^{(\mathrm{I})T}\widetilde{\boldsymbol{\beta}}_n^{(\mathrm{I})})) + \lambda_n\sum_{j=1}^{s_n} w_{n,j}|\beta_{n,j}|.$$

The criterion function in (B.1) is called an oracle subproblem of (3) since it only includes those relevant input variables.

Lemma 2. *Under Conditions* 1(*a*) *and* 2–4, *assume further that* $s_n^2/n \to 0$ *and* $w_{\max}^{(\mathrm{I})} = O_P\{1/(\lambda_n\sqrt{n})\}$. *Then, with probability tending to one, there exists a local minimizer* $\widehat{\widetilde{\mathbf{b}}}_n^{(\mathrm{I})}$ *of* ℓ_n^O *in* (B.1) *satisfies that* $\|\widehat{\widetilde{\mathbf{b}}}_n^{(\mathrm{I})} - \widetilde{\boldsymbol{\beta}}_{n;0}^{(\mathrm{I})}\| = O_P(\sqrt{s_n/n})$.

Proof. Let $\alpha_n = \sqrt{s_n/n}$ and $\mathbf{u}_n = (u_{n0}, u_{n1}, \ldots, u_{ns_n})^T \in \mathbb{R}^{s_n+1}$. It suffices to show that for any given $\epsilon > 0$, there is a constant C_ϵ large enough such that, for large n,

$$P\Big(\inf_{\|\mathbf{u}_n\|=C_\epsilon} \ell_n^O(\widetilde{\boldsymbol{\beta}}_{n;0}^{(\mathrm{I})} + \alpha_n\mathbf{u}_n) > \ell_n^O(\widetilde{\boldsymbol{\beta}}_{n;0}^{(\mathrm{I})})\Big) \geq 1-\epsilon.$$

By Taylor's expansion, $\ell_n^O(\widetilde{\boldsymbol{\beta}}_n^{(\mathrm{I})} + \alpha_n\mathbf{u}_n) - \ell_n^O(\widetilde{\boldsymbol{\beta}}_{n;0}^{(\mathrm{I})}) = I_1 + I_2$ where

$$\begin{aligned} I_1 &= I_{1,1} + I_{1,2} \\ &= \frac{\alpha_n}{n}\sum_{i=1}^n \mathrm{q}_1(Y_{ni}; \widetilde{\mathtt{X}}_{ni}^{(\mathrm{I})T}\widetilde{\boldsymbol{\beta}}_{n;0}^{(\mathrm{I})})(\widetilde{\mathtt{X}}_{ni}^{(\mathrm{I})T}\mathbf{u}_n) \\ &\quad + \frac{\alpha_n^2}{2n}\sum_{i=1}^n \mathrm{q}_2(Y_{ni}; \widetilde{\mathtt{X}}_{ni}^{(\mathrm{I})T}\widetilde{\boldsymbol{\beta}}_n^{(\mathrm{I})*})(\widetilde{\mathtt{X}}_{ni}^{(\mathrm{I})T}\mathbf{u}_n)^2, \\ I_2 &= \lambda_n\sum_{j=1}^{s_n} w_{nj}\{|\beta_{n,j;0} + \alpha_n u_{nj}| - |\beta_{n,j;0}|\}, \end{aligned}$$

with $\widetilde{\boldsymbol{\beta}}_n^{(\mathrm{I})*}$ located between $\widetilde{\boldsymbol{\beta}}_{n;0}^{(\mathrm{I})}$ and $\widetilde{\boldsymbol{\beta}}_{n;0}^{(\mathrm{I})} + \alpha_n\mathbf{u}_n$.

It is seen that

$$\begin{aligned} |I_{1,1}| &\leq \alpha_n\Big\|\frac{1}{n}\sum_{i=1}^n \mathrm{q}_1(Y_{ni}; \widetilde{\mathtt{X}}_{ni}^{(\mathrm{I})T}\widetilde{\boldsymbol{\beta}}_{n;0}^{(\mathrm{I})})\widetilde{\mathtt{X}}_{ni}^{(\mathrm{I})}\Big\|\|\mathbf{u}_n\| \\ &= O_P(s_n/n)\|\mathbf{u}_n\|, \\ I_{1,2} &= \frac{s_n}{2n}\mathbf{u}_n^T E\Big[\mathrm{q}_2(Y_n; \widetilde{\mathtt{X}}_n^{(\mathrm{I})T}\widetilde{\boldsymbol{\beta}}_n^{(\mathrm{I})*})\widetilde{\mathtt{X}}_n^{(\mathrm{I})}\widetilde{\mathtt{X}}_n^{(\mathrm{I})T}\Big]\mathbf{u}_n \\ &\quad + O_P(s_n^2/n^{3/2})\|\mathbf{u}_n\|^2, \\ I_2 &\geq -\lambda_n\alpha_n\sum_{j=1}^{s_n} w_{nj}|u_{nj}| \geq -\lambda_n(s_n/n^{1/2})w_{\max}^{(\mathrm{I})}\|\mathbf{u}_n\|. \end{aligned}$$

As $\|\widetilde{\boldsymbol{\beta}}_n^{(\mathrm{I})*} - \widetilde{\boldsymbol{\beta}}_{n,0}^{(\mathrm{I})}\| \le C_\epsilon \alpha_n$, for n large enough, $\|\widetilde{\boldsymbol{\beta}}_n^{(\mathrm{I})*} - \widetilde{\boldsymbol{\beta}}_{n;0}^{(\mathrm{I})}\| \le \log(n)\sqrt{s_n/n}$. The eigenvalues of $E[\mathrm{q}_2(Y_n; \widetilde{\mathbf{X}}_n^{(\mathrm{I})T}\widetilde{\boldsymbol{\beta}}_n^{(\mathrm{I})*})\widetilde{\mathbf{X}}_n^{(\mathrm{I})}\widetilde{\mathbf{X}}_n^{(\mathrm{I})T}]$ are therefore bounded away from 0 by Condition 4. Together with the conditions $s_n^2/n \to 0$ and $w_{\max}^{(\mathrm{I})} = O_P\{1/(\lambda_n\sqrt{n})\}$, we can choose C_ϵ large enough and set $\|\mathbf{u}_n\| = C_\epsilon$, such that $\{s_n/(2n)\}\mathbf{u}_n^T E[\mathrm{q}_2(Y_n; \widetilde{\mathbf{X}}_n^{(\mathrm{I})T}\widetilde{\boldsymbol{\beta}}_n^{(\mathrm{I})*})\widetilde{\mathbf{X}}_n^{(\mathrm{I})}\widetilde{\mathbf{X}}_n^{(\mathrm{I})T}]\mathbf{u}_n$ dominates all other terms in I_1 and I_2 with probability $1-\epsilon$ when n is large enough. □

B.3 Lemma 3

Lemma 3. *In addition to the conditions in Lemma 2, assume Condition 1(b),* $\log(p_n - s_n)/\min\{n, n\lambda_n^2(w_{\min}^{(\mathrm{II})})^2\} = o_P(1)$, *and* $(w_{\min}^{(\mathrm{II})})^{-1} = o_P\{\lambda_n\sqrt{n}/(\rho_n\sqrt{s_n})\}$. *Then, with probability tending to one, there exists a global minimizer* $\widehat{\widetilde{\boldsymbol{\beta}}}_n$ *of* ℓ_n *in (3) satisfying that* $\widehat{\widetilde{\boldsymbol{\beta}}}_n^{(\mathrm{I})}$ *is the minimizer* $\widehat{\widetilde{\mathbf{b}}}_n^{(\mathrm{I})}$ *of the oracle subproblem (B.1) in Lemma 2 and that* $\widehat{\boldsymbol{\beta}}_n^{(\mathrm{II})} = \mathbf{0}$.

Proof. Under Condition 1(b), $\widehat{\widetilde{\mathbf{b}}}_n^{(\mathrm{I})} = (\widehat{b}_{n,0}, \ldots, \widehat{b}_{n,s_n})^T$ is also a global minimizer due to the convexity of ℓ_n^O, satisfying the Karush-Kuhn-Tucker necessary conditions (Theorem A.1 in Wright, 1997):

$$\begin{cases} n^{-1}\sum_{i=1}^n \mathrm{q}_1(Y_{ni}; \widetilde{\mathbf{X}}_{ni}^{(\mathrm{I})T}\widehat{\widetilde{\mathbf{b}}}_n^{(\mathrm{I})}) = 0, \\ n^{-1}\sum_{i=1}^n \mathrm{q}_1(Y_{ni}; \widetilde{\mathbf{X}}_{ni}^{(\mathrm{I})T}\widehat{\widetilde{\mathbf{b}}}_n^{(\mathrm{I})})X_{ij} = -\lambda_n w_{n,j}\mathrm{sign}(\widehat{b}_{n,j}), \\ \quad \text{for } j \text{ with } \widehat{b}_{n,j} \neq 0, \\ |n^{-1}\sum_{i=1}^n \mathrm{q}_1(Y_{ni}; \widetilde{\mathbf{X}}_{ni}^{T}\widehat{\widetilde{\mathbf{b}}}_n^{(\mathrm{I})})X_{ij}| \le \lambda_n w_{n,j}, \\ \quad \text{for } j \text{ with } \widehat{b}_{n,j} = 0. \end{cases}$$

Comparing these conditions with the Karush-Kuhn-Tucker sufficient conditions (Theorem A.2 in Wright, 1997) for the convex criterion function ℓ_n in (3), it is seen that $(\widehat{\widetilde{\mathbf{b}}}_n^{(\mathrm{I})T}, \mathbf{0}^T)^T$ will be a global minimizer of ℓ_n if

$$\widehat{b}_{n,j} \neq 0, \quad \text{for } 1 \le j \le s_n, \tag{B.2}$$

$$\left|\frac{1}{n}\sum_{i=1}^n \mathrm{q}_1(Y_{ni}; \widetilde{\mathbf{X}}_{ni}^{(\mathrm{I})T}\widehat{\widetilde{\mathbf{b}}}_n^{(\mathrm{I})})X_{ij}\right| \le \lambda_n w_{n,j}, \quad \text{for } s_n+1 \le j \le p_n. \tag{B.3}$$

We separately verify (B.2) and (B.3) as follows.

First, Lemma 2 concludes that $\|\widehat{\widetilde{\mathbf{b}}}_n^{(\mathrm{I})} - \widetilde{\boldsymbol{\beta}}_{n;0}^{(\mathrm{I})}\| = O_P(\sqrt{s_n/n})$. Therefore,

$$\begin{aligned} &P\Big(\mathrm{sign}(\widehat{b}_{n,j}) \neq \mathrm{sign}(\beta_{n,j;0}) \text{ for some } j \in \{1, \ldots, s_n\}\Big) \\ &\le P\Big(|\widehat{b}_{n,j} - \beta_{n,j;0}| \ge |\beta_{n,j;0}| \text{ for some } j \in \{1, \ldots, s_n\}\Big) \\ &\le P\Big(\max_{1\le j\le s_n} |\widehat{b}_{n,j} - \beta_{n,j;0}| \ge \beta_{\min}^{(\mathrm{I})}\Big) \to 0, \end{aligned}$$

by the assumption that $\sqrt{n/s_n}\beta_{\min}^{(\mathrm{I})} \to \infty$. Thus (B.2) holds with probability tending to one.

Second, by Taylor's expansion, (B.3) holds if we can prove

$$P\left(\max_{s_n+1\le j\le p_n}\left|\frac{1}{n}\sum_{i=1}^n \mathrm{q}_1(Y_{ni}; \widetilde{\mathbf{X}}_{ni}^{(\mathrm{I})T}\widetilde{\boldsymbol{\beta}}_{n;0}^{(\mathrm{I})})X_{ij}\right| > \frac{\lambda_n}{2}w_{\min}^{(\mathrm{II})}\right) \to 0, \tag{B.4}$$

$$P\left(\max_{s_n+1\le j\le p_n}\left|\frac{1}{n}\sum_{i=1}^n \mathrm{q}_2(Y_{ni}; \widetilde{\mathbf{X}}_{ni}^{(\mathrm{I})T}\widetilde{\boldsymbol{\beta}}_n^{(\mathrm{I})*})\{\widetilde{\mathbf{X}}_{ni}^{(\mathrm{I})T}(\widehat{\widetilde{\mathbf{b}}}_n^{(\mathrm{I})} - \widetilde{\boldsymbol{\beta}}_{n;0}^{(\mathrm{I})})\} \times X_{ij}\right| > \frac{\lambda_n}{2}w_{\min}^{(\mathrm{II})}\right) \to 0, \tag{B.5}$$

with $\widetilde{\boldsymbol{\beta}}_n^{(\mathrm{I})*}$ located between $\widetilde{\boldsymbol{\beta}}_{n;0}^{(\mathrm{I})}$ and $\widehat{\widetilde{\mathbf{b}}}_n^{(\mathrm{I})}$.

We first prove (B.4). As in the proof of Lemma 1, denote $Z_{ij} = \mathrm{q}_1(Y_{ni}; \widetilde{\mathbf{X}}_{ni}^{(\mathrm{I})T}\widetilde{\boldsymbol{\beta}}_{n;0}^{(\mathrm{I})})X_{ij}$. Since $\log(p_n - s_n)/\{n\lambda_n^2(w_{\min}^{(\mathrm{II})})^2\} = o_P(1)$, it follows that for any $\epsilon > 0$, $P\{\log(p_n - s_n) > \epsilon n\lambda_n^2(w_{\min}^{(\mathrm{II})})^2\} = o(1)$. Thus,

$$\begin{aligned} &P\left(\max_{s_n+1\le j\le p_n}\left|\frac{1}{n}\sum_{i=1}^n Z_{ij}\right| > \frac{\lambda_n}{2}w_{\min}^{(\mathrm{II})}\right) \\ &\le P\left(\max_{s_n+1\le j\le p_n}\left|\sum_{i=1}^n Z_{ij}\right| > \sqrt{\frac{n\log(p_n - s_n)}{4\epsilon}}\right) + o(1). \end{aligned} \tag{B.6}$$

By Lemma 1 and Bonferroni inequality,

$$\begin{aligned} &P\left(\max_{s_n+1\le j\le p_n}\left|\sum_{i=1}^n Z_{ij}\right| > \sqrt{\frac{n\log(p_n - s_n)}{4\epsilon}}\right) \\ &\le 2(p_n - s_n)\exp\left\{\frac{-n\log(p_n - s_n)/(4\epsilon)}{C_1 n + C_2\sqrt{n\log(p_n - s_n)/(4\epsilon)}}\right\} \\ &= 2\exp\left\{\frac{-n\log(p_n - s_n)}{C_1 n + C_2\sqrt{n\log(p_n - s_n)/(4\epsilon)}} \times \left(\frac{1}{4\epsilon} - C_1 - C_2\sqrt{\frac{\log(p_n - s_n)}{4n\epsilon}}\right)\right\}. \end{aligned} \tag{B.7}$$

Since $\log(p_n - s_n)/n = o(1)$, we can choose $\epsilon > 0$ small enough, such that for large enough n, $1/(4\epsilon) - C_1 - C_2\sqrt{\log(p_n - s_n)/(4n\epsilon)} > 0$. Then the upper bound in (B.7) is $o(1)$. This, together with (B.6), implies (B.4).

We then prove (B.5).

$$\begin{aligned} &\max_{s_n+1\le j\le p_n}\left|\frac{1}{n}\sum_{i=1}^n \mathrm{q}_2(Y_{ni}; \widetilde{\mathbf{X}}_{ni}^{(\mathrm{I})T}\widetilde{\boldsymbol{\beta}}_n^{(\mathrm{I})*})\{\widetilde{\mathbf{X}}_{ni}^{(\mathrm{I})T}(\widehat{\widetilde{\mathbf{b}}}_n^{(\mathrm{I})} - \widetilde{\boldsymbol{\beta}}_{n;0}^{(\mathrm{I})})\}X_{ij}\right| \\ &\le \sup_{\|\mathbf{u}\|=1}\max_{s_n+1\le j\le p_n}\left|\frac{1}{n}\sum_{i=1}^n \mathrm{q}_2(Y_{ni}; \widetilde{\mathbf{X}}_{ni}^{(\mathrm{I})T}\widetilde{\boldsymbol{\beta}}_n^{(\mathrm{I})*})(\widetilde{\mathbf{X}}_{ni}^{(\mathrm{I})T}\mathbf{u})X_{ij}\right| \\ &\quad \times \|\widehat{\widetilde{\mathbf{b}}}_n^{(\mathrm{I})} - \widetilde{\boldsymbol{\beta}}_{n;0}^{(\mathrm{I})}\| \end{aligned}$$

$$\leq \sup_{\|\mathbf{u}\|=1} \left\| \mathbf{P}_n(\widetilde{\boldsymbol{\beta}}_n^{(\mathrm{I})*})\mathbf{u} \right\|_\infty O_P(\sqrt{s_n/n}) = O_P(\rho_n\sqrt{s_n/n}),$$

from the definition of $\mathbf{P}_n(\widetilde{\boldsymbol{\beta}}_n^{(\mathrm{I})})$ and ρ_n, and the fact that $\|\widetilde{\boldsymbol{\beta}}_n^{(\mathrm{I})*} - \widetilde{\boldsymbol{\beta}}_{n;0}^{(\mathrm{I})}\| = O_P(\sqrt{s_n/n})$. (B.5) is therefore proved by $(w_{\min}^{(\mathrm{II})})^{-1} = o_P\{\lambda_n\sqrt{n}/(\rho_n\sqrt{s_n})\}$. □

APPENDIX C. PROOF OF THEOREM 1

Part (1) of Theorem 1 is proved by Lemma 3. Part (2) of Theorem 1 is a direct conclusion by applying part (ii) of Theorem 6 in Zhang et al. (2010) to the oracle subproblem (B.1).

In addition, let us prove that $\rho_n = O_P(\sqrt{s_n})$ as mentioned in the discussion after Theorem 1:

$$\begin{aligned}
\rho_n &= \sup\Big\{ \Big\| \mathbf{P}_n(\widetilde{\boldsymbol{\beta}}_n^{(\mathrm{I})})\mathbf{u} \Big\|_\infty : \\
&\qquad \|\mathbf{u}\| = 1, \|\widetilde{\boldsymbol{\beta}}_n^{(\mathrm{I})} - \widetilde{\boldsymbol{\beta}}_{n;0}^{(\mathrm{I})}\| \leq \log(n)\sqrt{s_n/n} \Big\} \\
&\leq \sup\Big\{ \frac{1}{n}\sum_{i=1}^n \Big| \mathrm{q}_2(Y_{ni}; \widetilde{\mathbf{X}}_{ni}^{(\mathrm{I})T}\widetilde{\boldsymbol{\beta}}_n^{(\mathrm{I})}) \Big| \sup_{\|\mathbf{u}\|=1} \Big| \widetilde{\mathbf{X}}_{ni}^{(\mathrm{I})T}\mathbf{u} \Big| B_X : \\
&\qquad \|\widetilde{\boldsymbol{\beta}}_n^{(\mathrm{I})} - \widetilde{\boldsymbol{\beta}}_{n;0}^{(\mathrm{I})}\| \leq \log(n)\sqrt{s_n/n} \Big\} \\
&= O_P(\sqrt{s_n}),
\end{aligned}$$

as $n^{-1}\sum_{i=1}^n |\mathrm{q}_2(Y_{ni}; \widetilde{\mathbf{X}}_{ni}^{(\mathrm{I})T}\widetilde{\boldsymbol{\beta}}_n^{(\mathrm{I})})| = O_P(1)$ uniformly for any $\widetilde{\boldsymbol{\beta}}_n^{(\mathrm{I})}$ satisfying $\|\widetilde{\boldsymbol{\beta}}_n^{(\mathrm{I})} - \widetilde{\boldsymbol{\beta}}_{n;0}^{(\mathrm{I})}\| \leq \log(n)\sqrt{s_n/n}$.

Received 30 August 2012

REFERENCES

Alizadeh, A. A., Eisen, M. B., Davis, R. E., Ma, C., Lossos, I. S., Rosenwald, A., Boldrick, J. C., Sabet, H., Tran, T., Yu, X., Powell, J. I., Yang, L., Marti, G. E., Moore, T., Hudson, J. Jr., Lu, L., Lewis, D. B., Tibshirani, R., Sherlock, G., Chan, W. C., Greiner, T. C., Weisenburger, D. D., Armitage, J. O., Warnke, R., Levy, R., Wilson, W., Grever, M. R., Byrd, J. C., Botstein, D., Brown, P. O., and Staudt, L. M. (2000). Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling. *Nature* **403** 503–511.

Bach, F. (2010). Self-concordant analysis for logistic regression. *Electr. J. Statist.* **4** 384–414. MR2645490

Bartlett, M. S. (1953). Approximate confidence intervals. *Biometrika* **40** 12–19. MR0056889

Belloni, A. and Chernozhukov, V. (2011). ℓ_1-penalized quantile regression in high-dimensional sparse models. *Ann. Statist.* **39** 82–130. MR2797841

Brègman, L. M. (1967). A relaxation method of finding a common point of convex sets and its application to the solution of problems in convex programming. *U.S.S.R. Comput. Math. and Math. Phys.* **7** 620–631. MR0215617

Donoho, D. L. and Johnstone, I. M. (1994). Ideal spatial adaption by wavelet shrinkage. *Biometrika* **81** 425–455. MR1311089

Efron, B., Hastie, T., Johnstone, I., and Tibshirani, R. (2004). Least angle regression. *Ann. Statist.* **32** 407–499. MR2060166

Fan, J. and Lv, J. (2008). Sure independence screening for ultrahigh dimensional feature space. *J. Roy. Statist. Soc. Ser. B* **70** 849–911. MR2530322

Fan, J. and Peng, H. (2004). Nonconcave penalized likelihood with a diverging number of parameters. *Ann. Statist.* **32** 928–961. MR2065194

Fan, J., Samworth, R., and Wu, Y. (2009). Ultrahigh dimensional feature selection: beyond the linear model. *J. Machine Learning Res.* **10** 2013–2038. MR2550099

Fan, J. and Song, R. (2010). Sure independence screening in generalized linear models with NP-dimensionality. *Ann. Statist.* **38** 3567–3604. MR2766861

Huang, J., Ma, S. G., and Zhang, C.-H. (2008). Adaptive Lasso for sparse high-dimensional regression models. *Statist. Sinica* **18** 1603–1618. MR2469326

Kakade, S. M., Shamir, O., Sridharan, K., and Tewari, A. (2010). Learning exponential families in high-dimensions: strong convexity and sparsity. *Proceedings of the 13th International Conference on Artificial Intelligence and Statistics* **9** 381–388.

Knight, K. and Fu, W. J. (2000). Asymptotics for Lasso-type estimators. *Ann. Statist.* **28** 1356–1378. MR1805787

LeCun, Y., Bottou, L., Bengio, Y., and Haffner, P. (1998). Gradient-based learning applied to document recognition. *Proc. IEEE* **86** 2278–2324.

Lv, J. and Fan, Y. (2009). A unified approach to model selection and sparse recovery using regularized least squares. *Ann. Statist.* **37** 3498–3528. MR2549567

Meinshausen, N. and Buhlmann, P. (2006). High dimensional graphs and variable selection with the Lasso. *Ann. Statist.* **34** 1436–1462. MR2278363

Meinshausen, N. and Yu, B. (2006). Lasso-type recovery of sparse representations for high-dimensional data. *Ann. Statist.* **37** 246–270. MR2488351

Osborne, M. R., Presnell, B., and Turlach, B. A. (2000). A new approach to variable selection in least squares problems. *IMA Journal of Numerical Analysis* **20** 389–403. MR1773265

Rosset, S. and Zhu, J. (2007). Piecewise linear regularized solution paths. *Ann. Statist.* **35** 1012–1030. MR2341696

Shao, J. (2003). *Mathematical Statistics*, 2nd ed. Springer-Verlag, New York. MR2002723

Tibshirani, R. (1996). Regression shrinkage and selection via the Lasso. *J. Roy. Statist. Soc. Ser. B* **58** 267–288. MR1379242

van de Geer, S. A. (2008). High-dimensional generalized linear models and the Lasso. *Ann. Statist.* **36** 614–645. MR2396809

van der Vaart, A. W. and Wellner, J. A. (1996). *Weak Convergence and Empirical Processes: With Applications to Statistics*. Springer-Verlag, New York. MR1385671

Wang, L., Zhu, J., and Zou, H. (2006). The doubly regularized support vector machine. *Statist. Sinica* **16** 589–615. MR2267251

Wedderburn, R. W. M. (1974). Quasi-likelihood functions, generalized linear models, and the Gauss-Newton method. *Biometrika* **61** 439–447. MR0375592

Wright, S. J. (1997). *Primal-Dual Interior-Point Methods*. SIAM, Philadelphia. MR1422257

Zhang C.-H. and Huang J. (2008). The sparsity and bias of the Lasso selection in high-dimensional linear regression. *Ann. Statist.* **36** 1567–1594. MR2435448

Zhang, C. M., Jiang, Y., and Shang, Z. (2009). New aspects of Bregman divergence in regression and classification with parametric and nonparametric estimation. *Canad. J. Statist.* **37** 119–139. MR2509465

Zhang, C. M., Jiang, Y., and Chai, Y. (2010). Penalized Bregman divergence for large dimensional regression and classification. *Biometrika* **97** 551–566. MR2672483

Zhao, P. and Yu, B. (2007). On model selection consistency of Lasso. *J. Machine Learning Res.* **7** 2541–2567. MR2274449

Zhou, S., van de Geer, S. A., and Buhlmann, P. (2009). Adaptive Lasso for high dimensional regression and Gaussian graphical modeling. *arXiv:0903.2515v1.*

Zou, H. (2006). The adaptive Lasso and its oracle properties. *J. Amer. Statist. Assoc.* **101** 1418–1429. MR2279469

Yuan Jiang
Department of Statistics
Oregon State University
Corvallis, OR 97331-4606
USA
E-mail address: yuan.jiang@stat.oregonstate.edu

Chunming Zhang
Department of Statistics
University of Wisconsin-Madison
Madison, WI 53706-1510
USA
E-mail address: cmzhang@stat.wisc.edu