Regularized Estimation in Sparse High-dimensional Multivariate Regression, with Application to a DNA Methylation Study

Haixiang Zhang[1], Yinan Zheng[2], Grace Yoon[3], Zhou Zhang[2], Tao Gao[2], Brian Joyce[2], Wei Zhang[2], Joel Schwartz[4], Pantel Vokonas[5], Elena Colicino[6], Andrea Baccarelli[7], Lifang Hou[2], and Lei Liu[2*]

[1] *Center for Applied Mathematics, Tianjin University, Tianjin, 300072, China*

[2] *Department of Preventive Medicine, Northwestern University, Chicago IL 60611, USA*

[3] *Department of Statistics, Northwestern University, Chicago IL 60611, USA*

[4] *Department of Environmental Health, Harvard University, Boston, MA 02115, USA*

[5] *Department of Preventive Medicine, Icahn School of Medicine at Mount Sinai, New York, NY 10029, USA*

[6] *Normative Aging Study, Veterans Affairs Boston Healthcare System and Boston University, Boston, MA 02118, USA*

[7] *Department of Environmental Health Sciences, Columbia University, New York, NY 10032, USA*

[*] Corresponding author: lei.liu@northwestern.edu (L. Liu)

Summary. In this article, we consider variable selection for correlated high dimensional DNA methylation markers as multivariate outcomes. A novel weighted square-root LASSO procedure is proposed to estimate the regression coefficient matrix. A key feature of this method is tuning-insensitivity, which greatly simplifies the computation by obviating cross validation for penalty parameter selection. A precision matrix obtained via the constrained ℓ_1 minimization method (Cai et al. 2011) is used to account for the within-subject correlation among multivariate outcomes. Oracle inequalities of the regularized estimators are derived. The performance of our proposed method is illustrated via extensive simulation studies. We apply our method to study the relation between smoking and high dimensional DNA methylation markers in the Normative Aging Study (NAS).

Keywords. High-dimensional responses; Multivariate regression; Oracle inequality; Tuning-insensitive; Weighted square-root LASSO.

1 Introduction

With the development of modern technology for data collection, high-dimensional data have become increasingly common in many scientific research fields, e.g., genome-wide studies (Lin et al. 2015), biomedical sciences (Mukherjee et al. 2015), economics and finance (Basu and Michailidis 2015). Under these situations, the number of parameters is larger than the sample size, rendering traditional statistical procedures inappropriate. More recently, correlated data with high dimensional multivariate responses are often encountered in omics studies. Our motivating example is the Normative Aging Study, where methylation markers are taken as the multivariate outcomes. The methylation of DNA, where methyl groups are added to DNA at binding sites typically referred to as cytosine-phosphate-guanine (CpG) islands,

could affect the DNA expression. DNA methylation levels measured from probes close to one another are correlated (Moen et al. 2013), resulting in high dimensional multivariate outcomes. Correlation may also exist for DNA methylation markers having the similar function, e.g., related to exposure such as smoking. It is thus necessary to account for the within-subject correlation in the estimation procedure.

Our objective is to conduct selection of regression coefficients in high dimensional multivariate DNA methylation markers. There are two challenging issues for such a study: (1) how to conduct variable selection in the high dimensional setting; and (2) how to tackle the correlation among multivariate outcomes. For the first challenge, many studies have focused on penalized methods, such as the least absolute shrinkage and selection operator (LASSO, Tibshirani 1996), the smoothly clipped absolute deviation (SCAD, Fan and Li 2001), the elastic net (Zou and Hastie 2005), the adaptive LASSO (Zou 2006), and the minimax concave penalty (MCP, Zhang 2010). The penalized approach has been applied to many research topics, e.g., linear models (Wang and Leng 2007; Huang et al. 2011; Fan and Lv 2014), generalized linear models (van de Geer 2008; Jiang et al. 2016), survival models (Fan and Li 2002; Bradic et al. 2011; Lin and Lv 2013). Recently, Belloni et al. (2011) proposed a pivotal square-root LASSO method, which does not rely on the knowledge of the standard deviation for the error term. Later, Belloni et al. (2014) developed a self-tuning square-root LASSO method in high-dimensional nonparametric regression analysis. Liu and Wang (2017) proposed a new procedure for optimally estimating high dimensional Gaussian graphical models using the square-root LASSO. For more topics on variable selection, please refer to Bühlmann and van de Geer (2011).

There is limited research to tackle the second challenge, especially when the responses are high-dimensional. Rothman et al. (2010) proposed an iterative algorithm for variable selection and estimation in high-dimensional multivariate regression using the LASSO penalty.

Sofer et al. (2014) considered variable selection for high-dimensional multivariate regression using the penalized likelihood method, but the dimensionality of the multiple responses is still much smaller than the sample size. Liu et al. (2015) proposed a calibrated multivariate regression method for high-dimensional multivariate regression models, but they only considered the uncorrelated error structure. Li, Nan and Zhu (2015) and Wilms and Croux (2017) studied the group LASSO for high-dimensional multivariate linear regression model. Most of these existing methods use cross-validation to choose tuning parameters over a full regularization path, which are computationally expensive and may potentially waste valuable training data. To deal with this problem, we will extend Belloni et al. (2011)'s (unweighted) square-root LASSO on a single outcome to multivariate outcomes with weighted square-root LASSO. The main advantage of our procedure over existing methods comes from the tuning-insensitive property, which is significantly faster than cross-validation. Another advantage is that we can use the entire dataset for variable selection, which may potentially learn a better model (Bishop et al., 2003).

The rest of the paper is organized as follows. In Section 2, we introduce the model and the weighted square-root LASSO procedure for multivariate linear regression with high-dimensional responses. In Section 3, we establish an error bound for the proposed estimator. In Section 4, we develop an efficient algorithm and conduct Monte Carlo simulations to assess the performance of our method. An empirical analysis of DNA methylation in the Normative Aging Study is presented in Section 5. Some concluding remarks are given in Section 6. All technical proofs are relegated to the Appendix.

2 Model and Estimation

Consider the sparse, high-dimensional multivariate linear regression model

$$Y_i = BX_i + \epsilon_i, \qquad i = 1, \cdots, n, \tag{2.1}$$

where $Y_i = (Y_{i1}, \cdots, Y_{ip})'$ is the p-dimensional response, e.g., DNA methylation markers; $X_i = (X_{i1}, \cdots, X_{iq})'$ is the q-dimensional covariates vector; $B = (B_1, \cdots, B_p)' \in \mathbb{R}^{p \times q}$ is the sparse regression coefficient matrix with $B_k = (\beta_{k1}, \cdots, \beta_{kq})' \in \mathbb{R}^q$, $k = 1, \cdots, p$; $\epsilon_i = (\epsilon_{i1}, \cdots, \epsilon_{ip})' \in \mathbb{R}^p$ is the random error term with mean 0 and covariance matrix Σ_p. Throughout this article, we assume that the predictor's dimension q is fixed but the dimension of response p could be larger than n. The sparse regression coefficient matrix B suggests that only a small number of coefficients are non-zero. Our interest is to estimate the coefficient matrix B and establish oracle inequalities for corresponding estimators, while account for the correlated outcomes.

Assume that (Y_i, X_i) are independently and identically distributed (i.i.d.) observations, $i = 1, \cdots, n$. If $\epsilon_i \sim N(0, \Sigma_p)$, the negative log-likelihood function is given by

$$\mathcal{L}(B, \Omega_p) = \frac{1}{2n} \left\{ n \log(2\pi) + n \log |\Omega_p^{-1}| + \sum_{i=1}^{n} (Y_i - BX_i)' \Omega_p (Y_i - BX_i) \right\}, \tag{2.2}$$

where $\Omega_p = \Sigma_p^{-1}$ is the precision matrix (Cai et al. 2011). Denote $(B_1', \cdots, B_p')'$ as $\beta = (\beta_1, \cdots, \beta_d)'$, where $d = pq$. Let $\mathcal{S} = \{j; \beta_j \neq 0\}$ be the true model with size $s = |\mathcal{S}|$. We use Ω_p to account for the within-subject correlation (Sofer et al. 2014), and propose the following criterion function:

$$Q(\beta; \Omega_p) = \frac{1}{n} \sum_{i=1}^{n} (Y_i - BX_i)' \Omega_p (Y_i - BX_i). \tag{2.3}$$

In practice, the precision matrix Ω_p can be estimated using the constrained ℓ_1 minimization method (Cai et al. 2011), which has been implemented in the R package *flare* (Li,

Zhao, Yuan and Liu, 2015). Basically, the estimation of the sparse inverse covariance matrix (precision matrix) Ω_p can be obtained by the following optimization problem:

$$\min \|\Omega_p\|,$$

subject to:

$$|\Sigma_n \Omega_p - I|_\infty \leq \gamma,$$

where $\gamma > 0$ is tuning parameter and $\Sigma_n = \frac{1}{n}\sum_{i=1}^{n}(Y_i - \tilde{B}X_i)'(Y_i - \tilde{B}X_i)$ with $\tilde{B}$ being a consistent estimator (e.g. ridge estimator).

Belloni et al. (2011) proposed an unweighted square-root LASSO for β with a scalar outcome. They showed that the penalty (tuning parameter) is pivotal, i.e., it does not rely on the knowledge of the error variance, nor do we need to pre-estimate it. Consequently, the estimation of β is insensitive to the tuning parameter, greatly simplifying the computation which often resorts to cross validation for tuning parameter selection. Furthermore, the square-root LASSO method achieves near-oracle performance for the estimation of β. In comparison, the ordinary LASSO needs to estimate the error variance (Belloni et al. 2011) to achieve near-oracle performance, which is a very challenging issue in high dimensional data.

Motivated by Belloni et al. (2011), the corresponding weighted square-root LASSO version for correlated outcomes in Model (2.1) is defined as

$$\hat{\beta} = \arg\min_{\beta \in \mathbb{R}^d} \left\{ \sqrt{Q(\beta; \Omega_p)} + \lambda \sum_{j=1}^{d} w_j |\beta_j| \right\}, \tag{2.4}$$

where $\lambda > 0$ is the tuning parameter, w_j is a known weight, $j = 1, \cdots, d$. We can set $w_j = 1/|\tilde{\beta}_j|$ along the lines of Zou (2006) with $\tilde{\beta}_j$ be the ridge estimator, $j = 1, \cdots, d$. Of note, we choose the ridge estimator rather than the ordinary least square estimator in the adaptive LASSO since p could be larger than n in Model (2.1). To obtain $\hat{\beta}$ in (2.4), we

consider the optimization problem,

$$\hat{\beta} = \arg \min_{\beta \in \mathbb{R}^d, \rho \geq 0} \left\{ \frac{Q(\beta; \Omega_p)}{2\rho} + \frac{\rho}{2} + \lambda \sum_{j=1}^{d} w_j |\beta_j| \right\}. \tag{2.5}$$

For $\rho \geq 0$, we have $\frac{Q(\beta;\Omega_p)}{2\rho} + \frac{\rho}{2} \geq \sqrt{Q(\beta; \Omega_p)}$, so the objective function in (2.5) is an upper bound of that in (2.4). The equality is attained if and only if $\rho = \sqrt{Q(\beta; \Omega_p)}$. Similar to Proposition 3.1 of Liu and Wang (2017), the optimizations in (2.4) and (2.5) yield the same solution $\hat{\beta}$. This relationship between (2.4) and (2.5) provides an efficient algorithm as described below.

For given λ, we have the following procedure:

Step 0. Compute the precision matrix $\hat{\Omega}_p$ (using R package *flare*; Li, Zhao, Yuan and Liu, 2015) and $\hat{\beta}^{ridge}$ (using R package *glmnet*; Friedman et al., 2008). Set $\beta^{(0)} = \hat{\beta}^{ridge}$ and $\rho^{(0)} = \sqrt{Q(\beta^{(0)}; \hat{\Omega}_p)}$.

Step 1. Solve the optimization problem via coordinate descent algorithm (Friedman, et al., 2008)

$$\beta^{(k+1)} = \arg \min_{\beta \in \mathbb{R}^d} \left\{ \frac{Q(\beta; \hat{\Omega}_p)}{2\rho^{(k)}} + \frac{\rho^{(k)}}{2} + \lambda \sum_{j=1}^{d} w_j |\beta_j| \right\}.$$

Step 2. Update

$$\rho^{(k+1)} = \sqrt{Q(\beta^{(k+1)}; \hat{\Omega}_p)}.$$

Step 3. Repeat Steps 1 and 2 until convergence.

Of note, $\hat{\Omega}_p$ is consistent (Cai et al. 2011) and kept unchanged in the iteratively updated procedure, so the objective function in Step 1 is convex for β, which ensures fast convergence of the algorithm. The following Lemma 1 will present an explicit expression for λ, while the optimal value for the tuning parameter in (2.4) will be evaluated in Section 4 via empirical studies. The estimation procedure has been implemented in R (available upon request).

3 Theoretical results

In this section, we will establish the oracle inequality for $\hat{\beta}$ defined in (2.4). The following lemma lays the foundation for the tuning-insensitive property of the weighted square-root LASSO procedure, which is motivated by Bickel et al. (2009)'s choice on the penalty level for LASSO. Denote $W_{\min} = \min\{w_1, \cdots, w_d\}$, $W_{\max} = \max\{w_1, \cdots, w_d\}$ and $N = np$. We first have the following lemma.

Lemma 1. *Let* $\lambda = \frac{c}{W_{\min}}\sqrt{\frac{2a\log d}{N}}$ *with* $c > 1$ *and* $a > 2$. *Define*

$$\Omega = \{\lambda \geq \frac{c}{W_{\min}} \parallel \nabla Q^{1/2}(\beta; \Omega_p) \parallel_\infty\}, \tag{3.6}$$

then we have

$$P(\Omega) \geq 1 - \sqrt{\frac{2}{\pi a \log d}} \cdot d^{1-a\left(1-2\sqrt{\frac{(a-1)\log d}{N}}\right)} - d^{1-a},$$

where ∇ *is the gradient.*

Similar to Belloni et al. (2011), the prediction norm is given as $||e||_{2,N}^2 = \frac{1}{N}e'\mathbb{X}'\mathbb{X}e$, where $\mathbb{X}$ is defined in the Appendix and $e \in \mathbb{R}^d$. To derive the oracle inequalities, we define the compatibility factor (Huang et al. 2013) and restricted eigenvalue (RE, Bickel et al. 2009) as

$$\kappa(\xi, \mathcal{S}) = \inf_{0 \neq e \in \mathscr{C}(\xi, \mathcal{S})} \frac{s^{1/2}||e||_{2,N}}{||e_{\mathcal{S}}||_1} \quad \text{and} \quad \text{RE}(\xi, \mathcal{S}) = \inf_{0 \neq e \in \mathscr{C}(\xi, \mathcal{S})} \frac{||e||_{2,N}}{||e||_2},$$

respectively, where $\mathscr{C}(\xi, \mathcal{S}) = \{e \in \mathbb{R}^d : ||e_{\mathcal{S}^c}||_1 \leq \xi||e_{\mathcal{S}}||_1\}$ with $\xi = \frac{cW_{\max}+W_{\min}}{W_{\min}(c-1)}$. Hereafter, $\mathcal{A}^c$ denotes the complement of set $\mathcal{A}$; $v_{\mathcal{A}} = (v_j : j \in \mathcal{A})$ for a vector v. Then, we have the following conclusion.

Lemma 2. *Denote* $\hat{e} = \hat{\beta} - \beta$, *then on the event* Ω, *we have* $\hat{e} \in \mathscr{C}(\xi, \mathcal{S})$. That is, the L_1 norm of the variables not relevant should be less than a multiple of those relevant.

The following theorem gives the upper bound for the estimation error.

Theorem 1. *Let* $c > 1$, $\xi = \frac{cW_{\max}+W_{\min}}{W_{\min}(c-1)}$, *suppose that* $\frac{W_{\max}\lambda s^{1/2}}{\kappa(\xi,\mathcal{S})} \leq \zeta < 1$ *and* $Q(\beta;\Omega_p) \leq K^2$ *with* $K > 0$. *Then on the event* Ω,

$$\mathrm{RE}(\xi,\mathcal{S})||\hat{\beta}-\beta||_2 \leq ||\hat{\beta}-\beta||_{2,N} \leq 2\left(W_{\max}+\frac{W_{\min}}{c}\right)\frac{\lambda s^{1/2}K}{(1-\zeta^2)\kappa(\xi,\mathcal{S})}.$$

Remark 1. The condition $Q(\beta;\Omega_p) \leq K^2$ with $K > 0$ is mild, since $Q(\beta;\Omega_p)$ converges to 1 in probability as $n \to \infty$.

Remark 2. This result and Lemma 1 show that by choosing $\lambda = \frac{\eta}{W_{min}}\sqrt{\frac{\log d}{N}}$ with some $\eta = c\sqrt{2a} > 2$, the obtained $\hat{\beta}$ achieves the near-oracle rate of convergence (Belloni et al. 2011). Since the choice of η does not rely on any unknown parameters or quantities, we call the property tuning-insensitive. Empirically, it is found that setting $\eta = 8$ works well in most cases we encountered, which will be verified via simulation.

4 Simulation studies

In this section, we will conduct simulation studies to validate the proposed methodology. We assume that X is from $N_q(0,\Sigma_X)$, where $q = 2$ and $\Sigma_X = (\sigma_{ij})$ is given by $\sigma_{ij} = 0.7^{|i-j|}$. The random error term is generated from $N_p(0,\Sigma_E)$, we consider the following two settings for the error covariance:

Case (a). AR(1) error covariance: $\Sigma_{E,st} = 0.8^{|s-t|}$.

Case (b). Fractional Gaussian Noise (FGN) error covariance:

$$\Sigma_{E,st} = 0.5\left((|s-t|+1)^{2H} - 2|s-t|^{2H} + |(|s-t|-1)|^{2H}\right)$$

with Hurst parameter $H = 0.9$, so the correlation is (0.74, 0.63, 0.58, 0.55, 0.52, 0.50, 0.49, 0.48, 0.46, 0.45) for distance $|s-t| = 1:10$. It is noted that the inverse error covariance for Case (a) is a tri-diagonal sparse matrix, while Case (b) has a dense inverse error covariance. Set $\beta = (1, 0.6, 0.3, 1.2, 0.8, 0.5, 0, \cdots, 0)'$, i.e., the first six elements are non-zero, while the

rest are all 0. We take $p = 200$ and 300, respectively. All simulation results are based on 100 replications with $n = 100$ and 200.

According to Theorem 1, the tuning parameter $\lambda = \frac{\eta}{W_{min}}\sqrt{\frac{\log d}{N}}$ with some $\eta > 2$. To choose the suitable λ, we consider various values of η, and the corresponding performance is plotted in Figures 1 - 4. We can see that tuning parameter's effect is limited and the procedure has better properties when $\eta \in [6, 10]$. Thus, we suggest $\eta = 8$ for the weighted square-root LASSO (WSR-LASSO) method in both simulation and real application. For comparison, we also consider the LASSO and square-root LASSO (SR-LASSO) with weight $w_j = 1$ in (2.4).

Tables 1 and 2 report the results, which include the rate that the correct model (CMR) is selected $I\{\hat{\mathcal{S}} = \mathcal{S}\}$, the false positive rate (FPR) $|\hat{\mathcal{S}}\backslash\mathcal{S}|/|\hat{\mathcal{S}}|$, the false negative rate (FNR) $|\mathcal{S}\backslash\hat{\mathcal{S}}|/(d - |\hat{\mathcal{S}}|)$, and the model error (ME) $\mathrm{tr}[(\hat{B} - B)\Sigma_X(\hat{B} - B)']$ (Yuan and Lin 2007). It can be seen that LASSO and SR-LASSO have similar performance, which is in line with the conclusion of Belloni et al. (2011). Moreover, the WSR-LASSO method has a higher rate of selecting the correct model and a smaller model error than LASSO and SR-LASSO. The results on the false positive and negative rates also suggest that the proposed method is preferred over LASSO and SR-LASSO in practice.

5 Application

We apply our proposed methodology to the DNA methylation (DNAm) data from the US Department of Veterans Affairs' Normative Aging Study (NAS). We exclude participants who (i) were non-white or had missing information on race to minimize potential confounding effects of genetic ancestry, or (ii) had any cancer diagnosed and history of stroke or coronary artery disease as their blood methylation profiles could have been affected. A total of 169 individuals with samples collected at their first blood draw remain for analysis.

We are interested in the effect of smoking on DNA methylation. Gao et al. (2015) conducted a literature review on the DNA methylation change in response to active smoking exposure in adults. From it, we consider methylation markers at a total of 151 cytosine-phosphate-guanine (CpG) dinucleotides which had been reported multiple (≥ 2) times in the literature. In our studies, the correlations among the 151 CpG sites have a range of $[-0.6440, 0.9369]$, manifesting high correlations among these CpGs.

We are interesting in the smoking pack year (packyr)'s effect on these DNAm markers. We also include age and BMI in the model. In total, we need to estimate 453 (151×3) regression coefficients. We use the proposed method in Section 2 with tuning parameter $\lambda = 0.0026$.

In Table 3, we compare our results to the original 151 CpGs listed by Gao et al. (2015). In Table 4 we report the selected CpGs and coefficient estimates. Thirty three CpGs among a total of 151 are selected by our method in the NAS data. We can see that CpGs reported more frequently in the literature are also more likely to be chosen by our method in the NAS data. For example, among 8 CpGs reported at least 7 times, 5 (62.5%) are selected by our method from the NAS data, while only 15 out of 89 (16.9%) CpGs reported two times in literature are selected by our method. The decreasing trend in Table 3 shows the consistency of our method with the literature. Of note, our method correctly identifies the top two CpGs - cg03636183 and cg05575921 (located in F2RL3 and AHRR genes), which have been reported 12 and 11 times in literature, respectively.

6 Concluding remarks

We proposed a weighted square-root LASSO method for high-dimensional multivariate regression models. We estimated the precision matrix by the CLIME method to account for the correlations between responses and obtained oracle inequalities for the estimator. Sim-

ulation studies were provided to illustrate the proposed procedure. We applied the method to study the relation between smoking and high dimensional DNA methylation markers.

There exist several topics to research in the future. First, the estimation of high-dimensional Gaussian graphical models is an active area of research (Cai et al. 2011; Cai and Yuan 2012; Fan et al. 2013). It is of great interest to consider the joint estimation of regression coefficients and the precision matrix in (2.1). A possible solution is to add a penalty term in (2.4) for the elements of the precision matrix. Second, although it is assumed that the dimension of the covariates q is fixed, it is straightforward to extend the proposed procedure to the high-dimensional covariates setting, the main difficulty lies in the computational burden due to the ultra-high dimensional parameters. Third, statistical inference on the weighted square-root LASSO is an important and interesting topic. Fourth, since our method obviates the burden of cross-validation, our method is computational efficient: for the DNA methylation data in Section 5, it took R software about 10.3 seconds to converge in a personal computer. As a reviewer suggested, it would be of interest to make the proposed algorithm scalable to genome-wide response and predictor markers, which can be implemented in high performance computing facilities. Fifth, we are interested in high-dimensional mediation analysis (Zhang et al. 2016) to determine whether high-dimensional DNA methylation markers mediate the path from intervention (e.g. diet, physical exercise) to health outcomes.

Acknowledgements

We would like to thank the Editor, the Associate Editor and two reviewers for their helpful comments and suggestions, which helped us improve the article substantially.

Funding

This work was supported by AHA 14SFRN20480260, 12GRNT12070254, and National Institute of Environmental Health Sciences grant R01ES021357, R01ES021733, and R01ES015172, National Natural Science Foundation of China (Nos. 11301212, 11401146), and China Postdoctoral Science Foundation (No. 2014M550861). The VA Normative Aging Study is supported by the Cooperative Studies Program/Epidemiology Research and Information Center of the US Department of Veterans Affairs.

Appendix

For notational simplicity, let $\mathbb{Y} = (Y_1', \cdots, Y_n')' \in \mathbb{R}^N$ and $\mathcal{E} = (\epsilon_1', \cdots, \epsilon_n')' \in \mathbb{R}^N$ with $N = np$. For $\mathcal{A} = (a_{ij}) \in \mathbb{R}^{m\times n}$ and $\mathcal{B} = (b_{ij}) \in \mathbb{R}^{p\times q}$, the $\mathcal{A} \otimes \mathcal{B} \in \mathbb{R}^{mp\times nq}$ is defined as

$$\mathcal{A} \otimes \mathcal{B} = \begin{bmatrix} a_{11}\mathcal{B} & a_{12}\mathcal{B} & \cdots & a_{1n}\mathcal{B} \\ a_{21}\mathcal{B} & a_{22}\mathcal{B} & \cdots & a_{2n}\mathcal{B} \\ \vdots & \vdots & \vdots & \vdots \\ a_{m1}\mathcal{B} & a_{m2}\mathcal{B} & \cdots & a_{mn}\mathcal{B} \end{bmatrix}.$$

Let $\mathbb{X} = (\mathcal{X}_1', \cdots, \mathcal{X}_n')'$ with $\mathcal{X}_k = I_p \otimes X_k'$, $k = 1, \cdots, n$. Denote Λ as the $N \times N$ block diagonal matrix with the i-th diagonal component Ω_p, $i = 1, \cdots, n$. Then (2.3) can be rewritten as

$$Q(\beta; \Lambda) = \frac{1}{N}(\mathbb{Y} - \mathbb{X}\beta)'\Lambda(\mathbb{Y} - \mathbb{X}\beta). \tag{6.1}$$

In the following, we denote $Q(\beta; \Lambda)$ as $Q(\beta)$. We first need the following lemma.

Lemma 3. (Laurent and Massart, 2000). *Let $X \sim \chi^2_d$, then for $0 \le t < 1/2$, we have that*

$$P(X \le \{1 - t\}d) \le \exp\left(-\frac{1}{4}dt^2\right).$$

Proof of Lemma 1. Denote $\Lambda^{1/2}\mathbb{Y} = \Lambda^{1/2}\mathbb{X}'\beta + \Lambda^{1/2}\mathcal{E}$ as $\tilde{\mathbb{Y}} = \tilde{\mathbb{X}}'\beta + \tilde{\mathcal{E}}$, where $\tilde{\mathcal{E}}$ follows the N-dimensional multivariate normal distribution with mean $\mathbf{0}$ and covariance matrix $\mathrm{I}_{N\times N}$. Then, by the definition of $Q(\beta)$ in (6.1), we have

$$\begin{aligned}\sqrt{N}||\nabla Q^{1/2}(\beta)||_\infty &= \frac{||\sum_{i=1}^N \tilde{\mathbb{X}}_i'(\tilde{\mathbb{Y}}_i - \tilde{\mathbb{X}}_i'\beta)||_\infty}{\sqrt{\sum_{i=1}^N(\tilde{\mathbb{Y}}_i - \tilde{\mathbb{X}}_i'\beta)^2}} \\ &= \frac{||\sum_{i=1}^N \tilde{\mathbb{X}}_i'\tilde{\mathcal{E}}_i||_\infty}{\sqrt{\sum_{i=1}^N \tilde{\mathcal{E}}_i^2}}. \end{aligned} \tag{6.2}$$

Note that $\sum_{i=1}^N \tilde{\mathbb{X}}_{ij}\tilde{\mathcal{E}}_i \sim N(0,N)$ and $\sum_{i=1}^N \tilde{\mathcal{E}}_i^2 \sim \chi^2_N$, where $j = 1, \cdots, d$. Then it follows from Lemma 3 that

$$P\left(\sum_{i=1}^N \tilde{\mathcal{E}}_i^2 \le N(1-r_N)\right) \le \exp\left(-\frac{Nr_N^2}{4}\right),$$

where $0 \le r_N \le 1/2$. Moreover, we can derive the following inequality

$$\begin{aligned}&P\left(\frac{||\sum_{i=1}^N \tilde{\mathbb{X}}_i'\tilde{\mathcal{E}}_i||_\infty}{\sqrt{\sum_{i=1}^N \tilde{\mathcal{E}}_i^2}} > \sqrt{2a\log d}\right) \\ &\le P\left(||\sum_{i=1}^N \tilde{\mathbb{X}}_i'\tilde{\mathcal{E}}_i||_\infty > \sqrt{1-r_N}\cdot\sqrt{2Na\log d}\right) + P\left(\sum_{i=1}^N \tilde{\mathcal{E}}_i^2 \le N(1-r_N)\right) \\ &\le \sum_{j=1}^d P\left(|\sum_{i=1}^N \tilde{\mathbb{X}}_{ij}\tilde{\mathcal{E}}_i| > \sqrt{1-r_N}\cdot\sqrt{2Na\log d}\right) + \exp\left(-\frac{Nr_N^2}{4}\right) \\ &\le 2d\{1-\Phi(\sqrt{1-r_N}\cdot\sqrt{2a\log d})\} + \exp\left(-\frac{Nr_N^2}{4}\right) \\ &\le 2d\cdot\frac{d^{-a(1-r_N)}}{\sqrt{2\pi}\cdot\sqrt{1-r_N}\cdot\sqrt{2a\log d}} + \exp\left(-\frac{Nr_N^2}{4}\right) \\ &= \frac{d^{-a(1-r_N)}}{\sqrt{\pi(1-r_N)a\log d}} + \exp\left(-\frac{Nr_N^2}{4}\right),\end{aligned}$$

where the last inequality follows from $1-\Phi(t) \le \frac{1}{\sqrt{2\pi}t}\exp(-\frac{t^2}{2})$.

Let $r_N = 2\sqrt{\frac{(a-1)\log d}{N}}$, when n is large enough, we have

$$P\left(\sqrt{N}||\nabla Q^{1/2}(\beta)||_\infty \leq \sqrt{2a\log d}\right) \geq 1 - \sqrt{\frac{2}{\pi a \log d}} \cdot d^{1-a\left(1-2\sqrt{\frac{(a-1)\log d}{N}}\right)} - d^{1-a}.$$

Q.E.D. $\square$

Proof of Lemma 2. First, from the definition of $\hat{\beta}$ in (2.4), we notice that

$$\begin{aligned} Q^{1/2}(\hat{\beta}) - Q^{1/2}(\beta) &\leq \lambda \sum_{j=1}^{d} w_j|\beta_j| - \lambda \sum_{j=1}^{d} w_j|\hat{\beta}_j| \\ &\leq \lambda W_{\max}||(\hat{\beta}-\beta)_{\mathcal{S}}||_1 - \lambda W_{\min}||(\hat{\beta}-\beta)_{\mathcal{S}^c}||_1. \end{aligned} \tag{6.3}$$

Second, on the event Ω, there is $c \parallel \nabla Q^{1/2}(\beta) \parallel_\infty \leq \lambda W_{\min}$. Thus, using the fact that $Q(\beta)$ is a convex function, we have

$$\begin{aligned} Q^{1/2}(\hat{\beta}) - Q^{1/2}(\beta) &\geq -\nabla Q^{1/2}(\beta)(\beta - \hat{\beta}) \\ &\geq - \parallel \nabla Q^{1/2}(\beta) \parallel_\infty \cdot ||\hat{\beta}-\beta||_1 \\ &\geq -\frac{\lambda}{c} W_{\min}||\hat{\beta}-\beta||_1 \\ &= -\frac{\lambda}{c} W_{\min}\left(||(\hat{\beta}-\beta)_{\mathcal{S}}||_1 + ||(\hat{\beta}-\beta)_{\mathcal{S}^c}||_1\right). \end{aligned} \tag{6.4}$$

Combining (6.3) and (6.4), we can obtain

$$||(\hat{\beta}-\beta)_{\mathcal{S}^c}||_1 \leq \frac{cW_{\max} + W_{\min}}{W_{\min}(c-1)}||(\hat{\beta}-\beta)_{\mathcal{S}}||_1.$$

Q.E.D. $\square$

Proof of Theorem 1. We notice the following relation:

$$\begin{aligned} Q(\hat{\beta}) - Q(\beta) &= ||\hat{e}||_{2,N}^2 - \frac{2}{N}\sum_{i=1}^{N}(\tilde{\mathbb{Y}}_i - \tilde{\mathbb{X}}_i'\beta)\tilde{\mathbb{X}}_i'\hat{e} \\ &\geq ||\hat{e}||_{2,N}^2 - 2Q^{1/2}(\beta)||\nabla Q^{1/2}(\beta)||_\infty||\hat{e}||_1, \end{aligned} \tag{6.5}$$

where (6.5) holds by the Hölder inequality. Then it follows from the definition of $\kappa(\xi, \mathcal{S})$ that

$$\begin{aligned} ||\hat{e}||_{2,N}^2 &\leq 2Q^{1/2}(\beta)||\nabla Q^{1/2}(\beta)||_\infty||\hat{e}||_1 \\ &\quad +[Q^{1/2}(\hat{\beta}) + Q^{1/2}(\beta)] \cdot \lambda \left[W_{\max} \frac{s^{1/2}||\hat{e}||_{2,N}}{\kappa(\xi, \mathcal{S})} - W_{\min}||\hat{e}_{\mathcal{S}^c}||_1 \right]. \end{aligned} \tag{6.6}$$

Moreover, we note

$$Q^{1/2}(\hat{\beta}) \leq Q^{1/2}(\beta) + \lambda W_{\max} \left(\frac{s^{1/2}||\hat{e}||_{2,N}}{\kappa(\xi, \mathcal{S})} \right). \tag{6.7}$$

From (6.6) and (6.7), we have

$$\begin{aligned} ||\hat{e}||_{2,N}^2 &\leq 2Q^{1/2}(\beta)||\nabla Q^{1/2}(\beta)||_\infty||\hat{e}||_1 + 2Q^{1/2}(\beta)\lambda W_{\max} \left(\frac{s^{1/2}||\hat{e}||_{2,N}}{\kappa(\xi, \mathcal{S})} \right) \\ &\quad + \left\{ \lambda W_{\max} \left(\frac{s^{1/2}||\hat{e}||_{2,N}}{\kappa(\xi, \mathcal{S})} \right) \right\}^2 - 2Q^{1/2}(\beta)\lambda W_{\min}||\hat{e}_{\mathcal{S}^c}||_1. \end{aligned}$$

Since $c \parallel \nabla Q^{1/2}(\beta) \parallel_\infty \leq \lambda W_{\min}$, we have

$$\begin{aligned} ||\hat{e}||_{2,N}^2 &\leq 2Q^{1/2}(\beta)||\nabla Q^{1/2}(\beta)||_\infty||\hat{e}_{\mathcal{S}}||_1 + 2Q^{1/2}(\beta)\lambda W_{\max} \left(\frac{s^{1/2}||\hat{e}||_{2,N}}{\kappa(\xi, \mathcal{S})} \right) \\ &\quad + \left\{ \lambda W_{\max} \left(\frac{s^{1/2}||\hat{e}||_{2,N}}{\kappa(\xi, \mathcal{S})} \right) \right\}^2. \end{aligned}$$

Then,

$$\left\{ 1 - \left(\frac{W_{\max}\lambda s^{1/2}}{\kappa(\xi, \mathcal{S})} \right)^2 \right\} ||\hat{e}||_{2,N}^2 \leq 2 \left(W_{\max} + \frac{W_{\min}}{c} \right) Q^{1/2}(\beta) \frac{\lambda s^{1/2}}{\kappa(\xi, \mathcal{S})} ||\hat{e}||_{2,N}. \tag{6.8}$$

Since $\frac{W_{\max}\lambda s^{1/2}}{\kappa(\xi, \mathcal{S})} \leq \zeta < 1$ and $Q(\beta; \Omega_p) \leq K^2$ hold, by solving the above inequality (6.8), we can obtain the error bound stated in the theorem. Q.E.D. □

References

Basu, S. and Michailidis, G. (2015). Regularized estimation in sparse high-dimensional time series models. *The Annals of Statistics*, **43**, 1535-1567.

Belloni, A., Chernozhukov, V. and Wang, L. (2011). Square-root LASSO: Pivotal recovery of sparse signals via conic programming. *Biometrika*, **98**, 791-806.

Belloni, A., Chernozhukov, V. and Wang, L. (2014). Pivotal estimation via square-root LASSO in nonparametric regression. *The Annals of Statistics*, **42**, 757-788.

Bickel, P. J., Ritov, Y. and Tsybakov, A. (2009). Simultaneous analysis of LASSO and Dantzig selector. *The Annals of Statistics*, **37**, 1705-1732.

Bishop, C., Spiegelhalter, D. and Winn, J. (2003). VIBES: A variational inference engine for Bayesian networks. In *Advances in Neural Information Processing Systems 15* (S. Becker, S. Thrun and K. Obermayer, eds.). MIT Press, Cambridge, MA, 777 - 784.

Bradic, J., Fan, J. and Jiang, J. (2011). Regularization for Cox's proportional hazards model with NP-Dimensionality. *The Annals of Statistics*, **39**, 3092-3120.

Bühlmann, P. and van de Geer, S. (2011). *Statistics for High-Dimensional Data: Methods, Theory and Applications*. Springer.

Cai, T., Liu, W. and Luo, X. (2011). A constrained ℓ_1 minimization approach to sparse precision matrix estimation. *Journal of the American Statistical Association*, **106**, 594-607.

Cai, T. and Yuan, M. (2012). Adaptive covariance matrix estimation through block thresholding. *The Annals of Statistics*, **40**, 2014-2042.

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

Fan, J., and Li, R. (2002). Variable selection for Cox's proportional hazards model and frailty model. *The Annals of Statistics*, **30**, 74-99.

Fan, J., Liao, Y. and Mincheva, M. (2013). Large covariance estimation by thresholding principal orthogonal complements (with discussion). *Journal of the Royal Statistical Society, Series B*, **75**, 603-680.

Fan, Y. and Lv, J. (2014). Asymptotic properties for combined L_1 and concave regularization. *Biometrika*, **101**, 57-70.

Friedman, J., Hastie, T. and Tibshirani, R. (2008). Regularization paths for generalized linear models via coordinate descent. *Journal of Statistical Software*, **33**, No.1.

Gao, X., Jia, M., Zhang, Y., Breitling, L. and Brenner, H. (2015). DNA methylation changes of whole blood cells in response to active smoking exposure in adults: a systematic review of DNA methylation studies. *Clinical Epigenetics*, **7**:113.

Huang, J., Ma, S., Li, H. and Zhang, C.-H. (2011). The sparse Laplacian shrinkage estimator for high-dimensional regression. *The Annals of Statistics*, **39**, 2021-2046.

Huang, J., Sun, T., Ying, Z., Yu, Y. and Zhang, C.-H. (2013). Oracle inequalities for the LASSO in the Cox model. *The Annals of Statistics*, **41**, 1142-1165.

Jiang, Y., He, Y. and Zhang, H. (2016). Variable selection with prior information for generalized linear models via the prior lasso method. *Journal of the American Statistical Association*, **111**, 355-376.

Laurent, B. and Massart, P. (2000). Adaptive estimation of a quadratic functional by model selection. *The Annals of Statistics*, **28**, 1302-1338.

Li, X., Zhao, T., Yuan, X. and Liu, H. (2015). The flare package for high dimensional linear regression and precision matrix estimation in R. *Journal of Machine Learning Research*, **16**, 553-557.

Li, Y., Nan, B. and Zhu, J. (2015). Multivariate sparse group lasso for the multivariate multiple linear regression with an arbitrary group structure. *Biometrics*, **71**, 354-363.

Lin, W. and Lv, J. (2013). High-dimensional sparse additive hazards regression. *Journal of the American Statistical Association*, **108**, 247-264.

Lin, W., Feng, R. and Li, H. (2015). Regularization methods for high-dimensional instrumental variables regression with an application to genetical genomics. *Journal of the American Statistical Association*, **110**, 270-288.

Liu, H., Wang, L. and Zhao, T. (2015). Calibrated multivariate regression with application to neural semantic basis discovery. *Journal of Machine Learning Research*, **16**, 1579 - 1606.

Liu, H. and Wang, L. (2017). TIGER: A tuning-insensitive approach for optimally estimating Gaussian graphical models. *Electronic Journal of Statistics*, **11**, 241 - 294.

Moen, E., Zhang, X., Mu, W., Delaney, S., Wing, C., McQuade, J., Myers, J., Godley, L., Dolan, M. and Zhang, W. (2013). Genome-wide variation of cytosine modifications between European and African populations and the implications for complex traits. *Genetics*, **194**, 987-996.

Mukherjee, R., Pillai, N. and Lin, X. (2015). Hypothesis testing for high-dimensional sparse binary regression. *The Annals of Statistics*, **43**, 352-381.

Rothman, A., Levina, E., and Zhu, J. (2010). Sparse multivariate regression with covariance estimation. *Journal of Computational and Graphical Statistics*, **19**, 947-962.

Sofer, T., Dicker, L. and Lin, X. (2014). Variable selection for high dimensional multivariate outcomes. *Statistica Sinica*, **24**, 1633-1654.

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

van de Geer, S. (2008). High-dimensional generalized linear models and the lasso. *The Annals of Statistics*, **36**, 614-645.

Wang, H. and Leng, C. (2007). Unified LASSO estimation by least squares approximation. *Journal of the American Statistical Association*, **102**, 1039-1048.

Wilms, I. and Croux, C. (2017). An algorithm for the multivariate group lasso with covariance estimation. *Journal of Applied Statistics*, accepted.

Yuan, M. and Lin, Y. (2007). Model selection and estimation in the Gaussian graphical model. *Biometrika*, **95**, 19-35.

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

Zhang, H., Zheng, Y., Zhang, Z., Gao, T., Joyce, B., Yoon, G., Zhang, W., Schwartz, J., Just, A., Colicino, E., Vokonas, P., Zhao, L., Lv, J., Baccarelli, A., Hou, L. and Liu, L. (2016). Estimating and testing high-dimensional mediation effects in epigenetic studies. *Bioinformatics*, **32**, 3150 - 3154.

Zou, H. and Hastie, T. (2005). Regularization and vriable selection via the elastic net. *Journal of the Royal Statistical Society, Series B*, **67**, 301-320.

Zou, H. (2006). The adaptive LASSO and its oracle properties. *Journal of the American Statistical Association*, **101**, 1418-1429.

Table 1.

Simulation results on model selection with AR(1) error covariance.

		$n = 100$				$n = 200$			
p	Methods	CMR	FPR	FNR	ME	CMR	FPR	FNR	ME
200	LASSO	0.02	0.6121	0.0006	0.4203	0.10	0.5806	$< 10^{-4}$	0.2435
	SR-LASSO	0.02	0.5924	0.0005	0.4282	0.07	0.5973	$< 10^{-4}$	0.2405
	WSR-LASSO	0.41	0.1437	0.0008	0.2213	0.69	0.0708	0.0004	0.1337
300	LASSO	0	0.6072	0.0004	0.4747	0.02	0.6205	0.0001	0.2534
	SR-LASSO	0.01	0.5857	0.0005	0.4795	0.01	0.6468	0.0001	0.2476
	WSR-LASSO	0.38	0.2956	0.0002	0.2389	0.68	0.0634	0.0003	0.1178

CMR: the correct model is selected $I\{\hat{\mathcal{S}} = \mathcal{S}\}$; FPR: the false positive rate $|\hat{\mathcal{S}}\backslash\mathcal{S}|/|\hat{\mathcal{S}}|$; FNR: the false negative rate $|\mathcal{S}\backslash\hat{\mathcal{S}}|/(d - |\hat{\mathcal{S}}|)$; and ME: the model error $\mathrm{tr}[(\hat{B} - B)\Sigma_X(\hat{B} - B)']$.

Table 2.

Simulation results on model selection with FGN error covariance.

		$n = 100$				$n = 200$			
p	Methods	CMR	FPR	FNR	ME	CMR	FPR	FNR	ME
200	LASSO	0.02	0.5360	0.0003	0.6467	0.06	0.4504	0.0002	0.2891
	SR-LASSO	0.03	0.5696	0.0004	0.6287	0.05	0.4508	0.0001	0.2869
	WSR-LASSO	0.33	0.1055	0.0014	0.3242	0.60	0.0692	0.0007	0.1769
300	LASSO	0.03	0.5436	0.0004	0.6887	0.09	0.4922	$< 10^{-4}$	0.3762
	SR-LASSO	0.02	0.5239	0.0003	0.6741	0.07	0.4980	$< 10^{-4}$	0.3640
	WSR-LASSO	0.33	0.2634	0.0006	0.2828	0.55	0.0917	0.0006	0.2356

CMR: the correct model is selected $I\{\hat{\mathcal{S}} = \mathcal{S}\}$; FPR: the false positive rate $|\hat{\mathcal{S}}\backslash\mathcal{S}|/|\hat{\mathcal{S}}|$; FNR: the false negative rate $|\mathcal{S}\backslash\hat{\mathcal{S}}|/(d - |\hat{\mathcal{S}}|)$; and ME: the model error $\mathrm{tr}[(\hat{B} - B)\Sigma_X(\hat{B} - B)']$.

Table 3.

Model selection comparison with Gao et al. (2015)[‡].

Frequency CpGs reported	≥ 7	5-6	3-4	2
Total No. CpGs	8	12	42	89
N(%) identified in NAS	5 (62.5%)	4 (33.3%)	10 (23.8%)	15 (16.9%)

[‡] # CpG: The frequency of CpGs identified in literature, per the review by Gao et al. (2015); Total No. CpGs: Total number of CpGs reported in literature; N (%): Number (percentage) of CpGs selected by our method in the NAS data.

Table 4.

Variable selection and estimation results for CpGs[‡].

CpG	Gene name	CHR	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_3$
cg03636183	F2RL3	19	-0.0021	0.0029	0.0067
cg05575921	AHRR	5	-0.0056	0.0111	0.0083
cg06126421	*	6	-0.0014	0.0006	0.0094
cg21566642	*	2	-0.0019	0	-0.0058
cg06644428	*	2	-0.0037	-0.0085	-0.0306
cg03991871	AHRR	5	-0.0021	0.0146	0.0237
cg23576855	AHRR	5	-0.0013	0	0.0145
cg25189904	GNG12	1	-0.0027	0	-0.0092
cg08709672	AVPR1B	1	0.0011	0.0008	0
cg12803068	MYO1G	7	0.00164	0	0.0267
cg01692968	*	9	-3×10^{-5}	-0.0104	-0.0051
cg06060868	SDHA	5	4×10^{-6}	0.0116	0.0078
cg11207515	CNTNAP2	7	0.0004	-0.0049	0
cg11231349	NOS1AP	1	0.0005	0.0080	0.0149
cg22851561	C14orf43	14	0.0003	0	0
cg23771366	PRSS23	11	-0.0007	-0.0015	-0.0111
cg23916896	AHRR	5	-0.0010	-0.0092	-0.0135
cg26963277	KCNQ1	11	-0.0003	0.0133	0.0141

[‡] $\hat{\beta}_1$: packyr; $\hat{\beta}_2$: age; $\hat{\beta}_3$: BMI.

Table 4. Continued[‡].

CpG	Gene name	CHR	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_3$
cg01500140	LIM2	19	0.0008	0.0072	0.0113
cg03274391	*	3	0.0027	0	0.0001
cg03604011	AHRR	5	0.0035	-0.0202	-0.0143
cg04716530	ITGAL	16	3×10^{-5}	0.0111	0.0056
cg07465627	STXBP4	17	0.0002	-0.0052	-0.0092
cg11902777	AHRR	5	-0.0011	-0.0209	-0.0204
cg13039251	PDZD2	5	0.0031	0	0.0185
cg15187398	MOBKL2A	19	-0.0001	-0.0038	0
cg16201146	*	20	3×10^{-5}	0.0032	0.0129
cg17619755	VARS	6	0.0006	0	0.0082
cg17924476	AHRR	5	0.0011	0	-0.0102
cg23480021	*	3	0.0012	0	0.0157
cg23667432	ALPP	2	0.0002	0.0009	0.0092
cg23973524	CRTC1	19	0.0009	0	0.0057
cg26764244	GNG12	1	-0.0007	-0.0101	-0.0183

[‡] $\hat{\beta}_1$: packyr; $\hat{\beta}_2$: age; $\hat{\beta}_3$: BMI; $*$ denotes CpGs in the intergenic region;

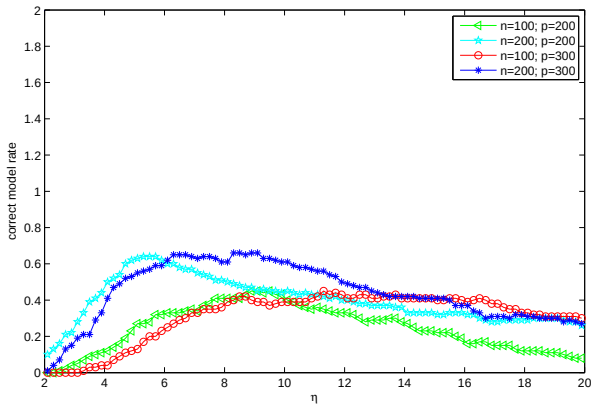

Figure 1. Correct model rate for WSR-LASSO with AR(1) error covariance.

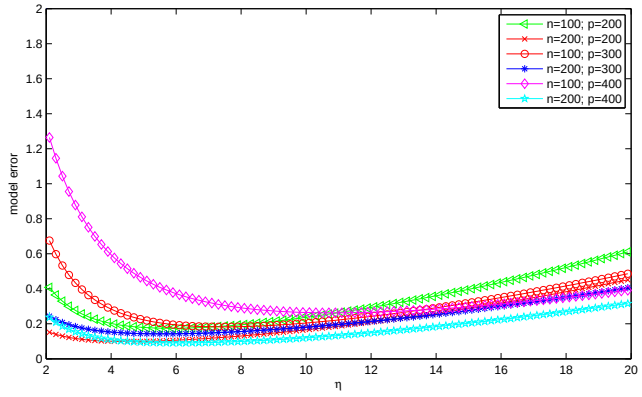

Figure 2. Model error for WSR-LASSO with AR(1) error covariance.

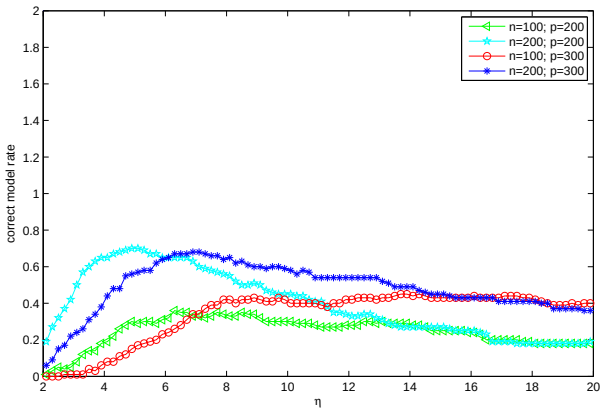

Figure 3. Correct model rate for WSR-LASSO with FGN error covariance.

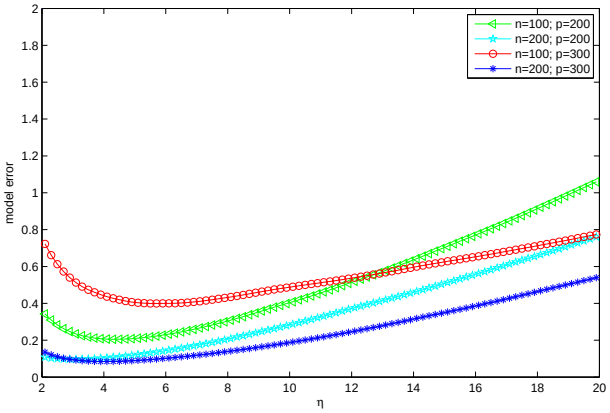

Figure 4. Model error for WSR-LASSO with FGN error covariance.