Statistica Sinica **24** (2014), 799-814
doi:http://dx.doi.org/10.5705/ss.2012.033

STRUCTURED LASSO FOR REGRESSION WITH MATRIX COVARIATES

Junlong Zhao[1] and Chenlei Leng[2,3]

[1] *Beihang University,* [2] *University of Warwick*
and [3] *National University of Singapore*

Abstract: High-dimensional matrix data are common in modern data analysis. Simply applying Lasso after vectorizing the observations ignores essential row and column information inherent in such data, rendering variable selection results less useful. In this paper, we propose a new approach that takes advantage of the structural information. The estimate is easy to compute and possesses favorable theoretical properties. Compared with Lasso, the new estimate can recover the sparse structure in both rows and columns under weaker assumptions. Simulations demonstrate its better performance in variable selection and convergence rate, compared to methods that ignore such information. An application to a dataset in medical science shows the usefulness of the proposal.

Key words and phrases: High-dimensional data, Lasso, model selection, non-asymptotic bounds, restricted eigenvalues, structured Lasso.

1. Introduction

In many modern applications, the number of the covariates p is much larger than the number of the observations n. In this high-dimensional setting, while the number of the variables can be large, the number of the important variables truly related to the response is often small, sometimes much smaller than p. Motivated by this, many approaches have been developed to recover a sparse signal vector from the linear regression model

$$Y_i = X_i^T \beta + \varepsilon_i, \quad i = 1, \ldots, n,$$

where $X_i \in R^p$ is the covariate, $Y_i \in R$ is the response, ε_i is the random noise, and β is an unknown sparse coefficient. Estimation accuracy and parsimony are the two fundamental considerations for handling high-dimensional data. A number of useful methods in the penalized likelihood framework have been developed during the past fifteen years. These include, for example, Lasso (Tibshirani (1996)) and SCAD (Fan and Li (2001)), among others. Meinshausen and Bühlmann (2006) is one of the first papers to study high-dimensional problems with $p \gg n$.

With the rapid development of new technology, it is increasingly common that data are collected with structural information. A structural information exists when matrix observations, rather than vector observations, are collected, and variables in each row and column share common characteristics. We wish to utilize this information. These data arise in, for example, food sciences and medical diagnosis via imaging, economics, and finance. Here the i-th observation is denoted by (X_i, Y_i), where X_i is a high-dimensional matrix. A simple way to deal with this type of data in the linear model framework is to vectorize X_i and then build a linear regression model based on the data $\{(\mathrm{vec}(X_i), Y_i), i = 1, \ldots, n\}$. There are two potential difficulties associated with this over-simplified approach: vectorization gives a regression problem of dimension $p \times q$ that may affect estimation accuracy; this approach inevitably loses meaningful information in the rows and columns of X_i, making interpretation less natural.

As a concrete example, Zhong and Suslick (2012) considered a data set from a colorimetric sensor array measuring multiple chemical interactions by using chemo-responsive dyes. The experimenters recorded the color changes of these dyes before and after exposure to some toxicants. These changes, digitalized in the form of a matrix with rows representing dye effects and columns representing the spectrum of colors, are the matrix covariates that are suitable for data analysis. In particular, Zhong and Suslick (2012) pointed out that by preserving the matrix nature of these variables, better predictive models could be developed.

As another example, Leng and Tang (2012) investigated the US agricultural exports in the last few decades. These agricultural exports are cross-classified according to the categories of the products, such as wheat, corn and others, and the regions to which they were exported, for instance, Asia, Europe Union, ect.. A natural question to ask among others is how the US exports are affected by different economic factors, with agricultural export being an important indicator.

Of course, data recorded in matrix form do no mean that the structure has to be meaningful. There are many examples where data are recorded in this way for efficient storage and better processing. Even for data sets where matrix variates make physical sense, we can always use some kind of cross validation to compare the predictive performance of the approach that vectorizes the data, and the one that preserves the matrix nature of the covariates. We have done this in our later data analysis.

We propose to use the following model for dealing with matrix covariates,

$$Y_i = \alpha^T X_i \beta + \varepsilon_i, \quad i = 1, \ldots, n, \tag{1.1}$$

where X_i is a $p \times q$ covariate matrix, $\alpha = (\alpha_1, \ldots, \alpha_p)^T$, $\beta = (\beta_1, \ldots, \beta_q)^T$, and the ε_i are *i.i.d.* $N(0, \sigma^2)$. A similar model was proposed by Li, Kim, and Altman (2010) in the setting of sufficient dimension reduction. Here we use this model with sparse vectors α and β to describe the contribution to Y of rows and columns of X, and to make variable selection for rows and columns.

If we use linear combinations of the column (or the row) variables in the form $X\beta$ (or $X\alpha$), our model is a standard linear regression model. This interpretation has a connection to the usual factor model if we view $X\beta$ (or $X\alpha$) as an unknown rank-one factor of the matrix data and view α (or β) as the unknown regression coefficient. Thus, our model specifies that the response variable is a linear function of the predictors made of linear combinations of either the row variables or the column variables that play symmetric roles. We always assume that $\min\{p, q\} > 1$. To avoid identification problems, we assume that $\|\alpha\|_1 = 1$ and that the element of β with the largest absolute value is positive. An immediate appeal of this model is that the number of the parameters is $p+q-1$ in stead of $p \times q$ were we to vectorize X_i. If a component of α is zero, the variables in the corresponding row of X then play no role in determining the response. Similarly, a zero in β states that the corresponding column in X does not influence Y.

Note that this model can also be written as $Y_i = \text{vec}(X_i)^T(\beta \otimes \alpha) + \epsilon_i$. Then, if $\beta \otimes \alpha$ is treated as a single parameter, say θ, we can apply the Lasso (Tibshirani (1996)) to obtain an estimate $\hat{\theta}_{las}$. Here we propose an approach termed structured Lasso for parameter estimation and variable selection by accounting for the structure. We show that the estimator, denoted as $\hat{\beta} \otimes \hat{\alpha}$, that uses the structure information is easy to compute and has favorable theoretical properties compared with $\hat{\theta}_{las}$. We discuss sufficient conditions inspired by, but weaker than, those in Bickel, Ritov, and Tsybakov (2009) that guarantee the success of the method. The finite sample performance of the method is compared to the usual Lasso via simulations and a data analysis.

The contents of this paper are arranged as follows. In Section 2, we introduce the structured Lasso and its computational algorithm. The theoretical properties of the estimate are presented in Sections 3 and 4. Simulation results and a data analysis are presented in Section 5, while some discussion is in Section 6. All the proofs involved are are available in the web-appendix in the Statistica Sinica web-page.

2. Structured Lasso

Let $\{(Y_i, X_i), i = 1, \ldots, n\}$ be *i.i.d.* observations. Consider the model

$$Y_i = \alpha^T X_i \beta + \varepsilon_i, \tag{2.1}$$

where $\alpha \in R^p, \beta \in R^q$, $X_i \in R^{p\times q}$ with $V_i = \text{vec}(X_i) \sim N(0, \Sigma)$, diagonal elements of Σ are 1 and the $\varepsilon_i \sim N(0, \sigma^2)$. For any $\beta = (\beta_1, \ldots, \beta_q)^T$, denote $|\beta_{(1)}| \geq \cdots \geq |\beta_{(q)}|$ as the decreasing order of $\{|\beta_j|, j = 1, \ldots, q\}$. For identifiability, and without loss of generality, we assume that $\|\alpha\|_1 = 1$ and $\text{sign}(\beta_{(1)}) = 1$.

Denote $S_\alpha = \{j : \alpha_j \neq 0\}$ and $S_\beta = \{j : \beta_j \neq 0\}$ as the set for important rows and columns, respectively, and let $p_0 = |S_\alpha|$ and $q_0 = |S_\beta|$ be the cardinalities of these two sets. Letting $\theta = \beta \otimes \alpha$, it is easy to see that $\alpha^T X_i \beta = V_i^T \theta$. For brevity, we write

$$P_\theta = \|\alpha\|_1 \|\beta\|_1 = \|\theta\|_1,$$

where $\|\cdot\|_1$ denotes the ℓ_1 norm. We propose the structured Lasso estimator as the solution to the optimization problem

$$(\hat{\alpha}, \hat{\beta}) = \arg\min_{(\alpha,\beta)\in\mathcal{E}} \frac{1}{n}\sum_{i=1}^{n}(Y_i - \alpha^T X_i \beta)^2 + \lambda_n \|\alpha\|_1 \|\beta\|_1, \tag{2.2}$$

where $\mathcal{E} = \{(\alpha, \beta) : \alpha \in R^p, \|\alpha\|_1 = 1, \beta \in R^q, \text{sign}(\beta_{(1)}) = 1\}$.

It is not difficult to see that (2.2) is a nonconvex function in $(\alpha^T, \beta^T)^T$. However, it is conditionally convex for one set of parameters given the other. This observation motivates an iterative algorithm for computation. Specifically, we can optimize α for fixed β, and vice versa.

If one set of parameters is fixed, the optimization problem reduces to the Lasso formulation and can be solved by the Lasso algorithm. In particular, the fast coordinate descent algorithm (Friedman, Hastie, and Tibshirani (2010)) designed especially for the Lasso can be used for our optimization. Let $\rho_{ij} = \text{cov}(Y, X_{ij})$, where X_{ij} is the (i,j)th element of X, and $\hat{\rho}_{ij}$ be the corresponding sample version. Write $(i_0, j_0) = \arg\max \hat{\rho}_{ij}$. We now state the algorithm.

(1) Take $\alpha^{(1)}$ as the vector with i_0-th element $\sum_j \hat{\rho}_{i_0 j}$ and 0 otherwise, and standardize $\alpha^{(1)}$, so that $\|\alpha^{(1)}\|_1 = 1$.

(2) Fix $\alpha^{(m)}$ and apply the Lasso algorithm with $(X_i^T \alpha^{(m)}, Y_i)$ as the i-th observation to optimize β and obtain $\beta^{(m)}$. Adjust $\beta^{(m)}$ such that $\text{sign}(\beta_{(1)}^{(m)}) = 1$. Then fix $\beta^{(m)}$ and apply the Lasso algorithm with $(X_i \beta^{(m+1)}, Y_i)$ as the i-th observation to obtain $\alpha^{(m+1)}$.

(3) Let $m \leftarrow m + 1$. Repeat Step (2) until convergence.

We scale the estimator so that the solution of (2.2) has the property satisfying $\|\hat{\alpha}\|_1 = 1$ and $\text{sign}(\hat{\beta}_{(1)}) = 1$. The algorithm converges very quickly in our simulation studies. We also initialized the algorithm using the best fitting α and β to minimize the ℓ_2 norm between $\alpha \otimes \beta$ and the Lasso solution. The results for the simulations were similar and thus are omitted.

Although this alternating approach is not guaranteed to find the global minimizer, it converges to a stationary point due to the bi-convexity of the objective function. The proposed algorithm performs well in comparison to the usual Lasso that vectorizes X and is convex. If finding the global minimizer is a concern, multiple randomized initial values can be tried to alleviate the problem.

3. Theoretical Properties of The Global Minimizer

Let $\mathbb{Y} = (Y_1, \ldots, Y_n)$, $\mathbb{V} = (V_1, \ldots, V_n)^T$, and $\varepsilon = (\varepsilon_1, \ldots, \varepsilon_n)^T$. Then (2.2) can be rewritten as

$$(\hat{\alpha}, \hat{\beta}) = \arg \min_{(\alpha,\beta)\in\mathcal{E}} \frac{1}{n}\|\mathbb{Y} - \mathbb{V}^T\theta\|_2^2 + \lambda_n P_\theta,$$

where $\|\cdot\|_2$ is the ℓ_2 norm and $P_\theta = \|\theta\|_1$. Let $S_\theta = \{i : \theta_i \neq 0\}$ and $s_0 = |S_\theta|$. Here $s_0 = p_0 q_0$.

At the first glance, this is quite similar to a standard Lasso problem with parameter $\theta \in R^{pq}$ as the coefficient. The difference is that θ here has the special structure $\theta = \beta \otimes \alpha$. Therefore $\hat{\theta} = \hat{\beta} \otimes \hat{\alpha}$ is the optimal solution on a subspace of R^{pq} denoted as $\{v_1 \otimes v_2 : (v_1, v_2) \in \mathcal{E}\}$. We introduce a structured restricted eigenvalue (SRE) condition tailored for our matrix data analysis.

Structured RE condition ($SRE(s_0, k_0)$). There exists $\kappa(s_0, k_0) > 0$, such that

$$\min_{\substack{S_0\subseteq\{1,\ldots,pq\}\\ |S_0|\le s_0}} \min_{\substack{u_{S_0^c}<k_0u_{S_0}\\ u\in\mathcal{J}_0}} \frac{\|\mathbb{V}u\|_2}{\sqrt{n}\|u_{S_0}\|_2} \geq \kappa(s_0, k_0),$$

where $\mathcal{J}_0 = \{u : u = v_1 \otimes v_2 - \beta \otimes \alpha; v_1 \in R^q, v_2 \in R^p\}$ is a subset of R^{pq}.

This contrasts with what we call the (unstructured) RE condition introduced by Bickel, Ritov, and Tsybakov (2009) that ignores the structure of the parameter.

RE condition (Unstructured) ($RE(s_0, k_0)$). There exists $\tilde{\kappa}(s_0, k_0) > 0$, such that

$$\min_{\substack{S_0\subseteq\{1,\ldots,pq\}\\ |S_0|\le s_0}} \min_{\substack{u_{S_0^c}<k_0u_{S_0}\\ u\in\widetilde{\mathcal{J}}_0}} \frac{\|\mathbb{V}u\|_2}{\sqrt{n}\|u_{S_0}\|_2} \geq \tilde{\kappa}(s_0, k_0),$$

where $\widetilde{\mathcal{J}}_0 = \{u : u = v - \beta \otimes \alpha; v \in R^{pq}\}$ is a subset of R^{pq}.

Since $\mathcal{J}_0 \subseteq \widetilde{\mathcal{J}}_0$, $\kappa(s_0, k_0) \geq \tilde{\kappa}(s_0, k_0)$, so the unstructured RE condition implies the structured RE condition. We give bounds on the convergence rate of the new estimate, and then discuss sufficient conditions for the structured and unstructured RE conditions. In particular, we show that the minimal sample size required by the SRE is less than that by the RE.

3.1. Bounds on the convergence rate

Based on the optimality of $\hat{\theta}$ on the space $\{v_1 \otimes v_2, (v_1, v_2) \in \mathcal{E}\}$ and the structured RE condition, we obtain bound $\hat{\beta} - \beta$ and $\hat{\alpha} - \alpha$ under the normality assumption on X_i and ϵ_i. Our assumption is motivated by Raskutti, Wainwright, and Yu (2010) who provided sufficient conditions for the RE condition to hold.

For any $\delta_0 > 0$, let

$$\mathcal{A} = \left\{ \frac{\|\mathbb{V}\varepsilon\|_\infty}{n} < \lambda_n, \frac{\|\mathbb{V}_j\|_2}{\sqrt{n}} \le 1 + \delta_0, j = 1, \ldots, pq \right\}.$$

Lemma 1. *Suppose $X_i, i = 1, \ldots, n$ are i.i.d. multivariate normal. With*

$$\lambda_n = (1+\delta_0)\sigma\sqrt{2(1+a)\frac{\log(pq)}{n}}$$

for any $a > 0$, we have

$$P(\mathcal{A}) \ge (1 - [(pq)^a\sqrt{\pi\log(pq)}]^{-1})[1 - \exp(-\frac{1}{8}\min(n\delta_1, n\delta_1^2) + \log(pq))],$$

where $\delta_1 = (1+\delta_0)^2 - 1$.

Based on Lemma 1, the event $\mathcal{A}$ holds with probability tending to one under appropriate conditions.

Theorem 1. *Let $\hat{\alpha}, \hat{\beta}$ be the global estimators defined in* (2.2)*. Suppose the structured restricted eigenvalue condition holds with $\kappa(s_0, 3) > 0$. For any $\delta_0 > 0$, given $\lambda_n = (1+\delta_0)\sigma\sqrt{2(1+a)\log(pq)/n}$ for model* (2.1)*, conditioning on $\mathcal{A}$, for $B_0 = 4\kappa^2(s_0, 3)$ we have*

$$\|\hat{\alpha} - \alpha\|_1 \le \frac{2B_0\lambda_n s_0}{|\beta_{(1)}|}$$
$$\|\hat{\beta} - \beta\|_1 \le B_0\lambda_n s_0(1 + \frac{2\|\beta\|_1}{|\beta_{(1)}|}).$$

Ignoring the structure and simply treating $\beta\otimes\alpha$ as one parameter, we obtain the Lasso estimator $\hat{\theta}_{las}$ of θ. We have the following results from Bickel, Ritov, and Tsybakov (2009) and Zhou (2009).

Proposition 1 (Zhou (2009))**.** *Suppose the unstructured restricted eigenvalue condition holds with $\kappa(s_0, 3) > 0$. For any $\delta_0 > 0$, given $\lambda_n = (1+\delta_0)\sigma$ $\sqrt{2(1+a)\log(pq)/n}$ for model* (2.1)*, conditioning on $\mathcal{A}$, for $\tilde{B}_0 = 4\tilde{\kappa}^2(s_0, 3)$ we have*

$$\|u_{las}\|_1 \le \tilde{B}_0\lambda_n s_0,$$

where $u_{las} = \hat{\theta}_{las} - \alpha\otimes\beta$.

3.2. Sufficient conditions

We consider conditions under which the structured and unstructured RE condition hold. Two theorems give, respectively, the sufficient conditions for the structured RE condition and unstructured RE condition. Vershynin (2012) reviewed tools for non-asymptotic analysis of random matrices. Accordingly, we define the ψ_2 norm of a sub-Gaussian random scalar $X \in R$ as $\|X\|_{\psi_2} = \sup_{k\geq 1} k^{-1/2}(E|X|^k)^{1/k}$. If $X \sim N(0,1)$, we have $\|X\|_{\psi_2} \leq C$ where C is an absolute constant. For a sub-Gaussian random vector $\tilde{X} \in R^p$, the ψ_2 norm is defined as $\|\tilde{X}\|_{\psi_2} = \sup\limits_{\|a\|_2=1} \|a^T\tilde{X}\|_{\psi_2}$. Let α_0 denote the ψ_2 norm of $N(0, I_{pq})$.

Theorem 2. *Suppose V is normal random vector on R^{pq}, and that $V_i, i = 1,\ldots,n$ are i.i.d. copies of V. Under* (4.1)*, for any $0\leq\epsilon\leq 1/2$ and $0\leq\gamma\leq 1$, if*

$$n > \frac{c'\alpha_0^4 C_{s_0,k_0}^2}{\gamma^2}\log[c(\epsilon,s_0)\max(p^{2s_0}q, pq^{2s_0}], \tag{3.1}$$

then with probability at least $1-\exp\{-\bar{c}\gamma^2 n/\alpha_0^4\}$ the structured RE condition holds, where $c', \bar{c} > 0$ and $c(\epsilon,s_0), C_{s_0,k_0}$ are defined in Lemma 3*.*

The following is a variant of a result of Zhou (2009).

Theorem 3. *Under the conditions of Theorem* 2*, for any $0\leq\epsilon\leq 1/2$, if*

$$n > \frac{c'\alpha_0^4 C_{s_0,k_0}^2}{\gamma^2}\Big(\log[c_1(\epsilon,s_0)(pq)^{2s_0}]\Big), \tag{3.2}$$

then with probability at least $1-\exp\{-\bar{c}\gamma^2 n/\alpha_0^4\}$ the unstructured RE condition holds, where $c', \bar{c} > 0$, $c_1(s_0,\epsilon) = 2s_0\left(5d_0 e/4s_0\epsilon\right)^{2s_0}$ and d_0, C_{s_0,k_0} are defined in Lemma 3*.*

The difference between the lower bound of n in (3.2) and that in (3.1) tends to $+\infty$, as $\min(p,q)\longrightarrow\infty$. Therefore, the sample size required for the structured RE condition is smaller than that of unstructured RE condition. This supports our empirical findings that structured Lasso has better performance on prediction error than Lasso.

4. Structured RE Condition

For any $u = (u_1,\ldots,u_{pq})^T \in R^{pq}$, sort $|u_1|,\cdots,|u_{pq}|$ in decreasing order, and let T_0 be index of the first largest s_0 elements, T_1 be the next largest s_0 elements. Similarly we can define $T_k, k = 2,3,\cdots$, and we take u_{T_k} to be the sub-vector of u consisting of elements with index in T_k. Let

$$\mathcal{J}_1 = \{u \in \mathcal{J}_0, \text{ such that } \|\Sigma^{1/2}u\|_2 = 1, \|u_{T_0^c}\|_1 \leq \|k_0 u_{T_0}\|_1\},$$
$$\mathcal{I}_1 = \{v = \Sigma^{1/2}u : u \in \mathcal{J}_1\}.$$

Zhou (2009) considered the RE condition on a sub-gaussian random matrix by applying the results of Mendelson, Pajor, and Tomczak-Jaegermann (2007, 2008) that are based on the complexity measure of the set considered. Since we consider the RE condition with a special structure, these general approaches are also suitable for our problem. For $\mathcal{I}_1 \subset R^{pq}$, define the complexity measure (Zhou (2009)) as

$$l_*(\mathcal{I}_1) = \mathbb{E} \sup_{v \in \mathcal{I}_1} |g^T v| = \mathbb{E} \sup_{u \in \mathcal{J}_1} |g^T \Sigma^{1/2} u|,$$

where $u = (u_1, \ldots, u_{pq})^T \in R^{pq}$ and $g = (g_1, \ldots, g_{pq})^T \in R^{pq}$, with the g_i's are independent normal $N(0,1)$.

RE condition on Σ. Suppose the diagonal elements of Σ are 1 and that for some $1 \le s \le pq$ and a positive number k,

$$K(s,k,\Sigma) := \min_{\substack{J_0 \subset \{1,\ldots,pq\} \\ |J_0| \le s}} \min_{\substack{u \neq 0 \\ \|u_{J_0^c}\|_1 \le k \|u_{J_0}\|_1}} \frac{\|\Sigma^{1/2} u\|_2}{\|u_{J_0}\|_2} > 0. \tag{4.1}$$

Let

$$\sqrt{\rho_{\max}(m)} = \max_{\substack{\|u\|_2 = 1 \\ |\mathrm{supp}(u)| \le m}} \|\Sigma^{1/2} u\|_2,$$

where $\mathrm{supp}(u) = \{i : u_i \neq 0\}$. To simplify the computation of $l_*(\mathcal{I}_1)$, we first give a lemma; its proof is that of Proposition 1.4 of Zhou (2009), and is omitted.

Lemma 2. *For any vector $u \in R^{pq}$, if $\|u_{S_\theta}\|_1 \le k_0 \|u_{S_\theta^c}\|_1$, then $\|u_{T_0^c}\|_1 \le k_0 \|u_{T_0}\|_1$.*

Lemma 3. *Suppose* (4.1) *holds. For $0 < \epsilon \le 1/2$,*

$$l_*(\mathcal{I}_1) \le C_{s_0,k_0} \sqrt{\log[c(\epsilon, s_0) \max(p^{2s_0} q, pq^{2s_0})]},$$

where $C_{s_0,k_0} = 6(k_0+2)\sqrt{\rho_{\max}(s_0)}/K(s_0,k_0,\Sigma)$, $c(\epsilon,s_0) = 2s_0\,(15d_0/2\epsilon)^{2s_0+1}$ $(e/2s_0)^{2s_0}$, and $d_0 = \|\beta\|_2 \cdot \|\alpha\|_2 + 1$.

The proof of Lemma 3 is closely related to the covering number of U_{s_0} defined as

$$U_{s_0} = \{u : u \in \mathcal{J}_0, \|u\|_2 = 1, |\mathrm{supp}(u)| = s_0\},$$

where $s_0 = p_0 q_0$.

4.1. The covering number of U_{s_0}

Consider the ϵ-cover $\Pi_{U_{s_0}}$ of U_{s_0}. Here we compute the corresponding covering number. Any $u \in U_{s_0}$ has the form $(v_1 \otimes v_2 - \beta \otimes \alpha)$. Due to the fact $|\mathrm{supp}(\alpha)| = p_0$, $|\mathrm{supp}(\beta)| = q_0$, and $|\mathrm{supp}(u)| = s_0$, it follows that $|\mathrm{supp}(v_1 \otimes v_2)| = |\mathrm{supp}(v_2)| \cdot |\mathrm{supp}(v_1)| \le 2s_0$. Since $\|u\|_2 = 1$ for any $u \in U_{s_0}$, the triangular inequality gives

$$\|v_1 \otimes v_2\|_2 \leq \|\beta \otimes \alpha\|_2 + 1 = \|\alpha\|_2 \|\beta\|_2 + 1 := d_0.$$

Take

$$W_0 = \{w = v_1 \otimes v_2;\ v_1 \in R^q,\ v_2 \in R^p, \|v_1 \otimes v_2\|_2 \leq d_0,\ |\mathrm{supp}(v_1)| \cdot |\mathrm{supp}(v_2)| \leq 2s_0\}.$$

It is obvious that $U_{s_0} \subseteq W_0 - \beta \otimes \alpha$. If Π_{W_0} is the ϵ-cover of W_0, then $\Pi_{W_0} - \beta \otimes \alpha$ is the ϵ-cover of U_{s_0}, so

$$\Pi_{U_{s_0}} \leq \Pi_{W_0}. \tag{4.2}$$

Therefore, it is sufficient to consider the covering number of W_0.

Lemma 4. *There exists an ϵ-cover Π_{W_0} of W_0 with*

$$|\Pi_{W_0}| \leq \sum_{0 < k_1, k_2 \in \mathbb{Z}, k_1 k_2 \leq 2s_0} \left(\frac{15 d_0}{2\epsilon}\right)^{k_1 + k_2} \binom{q}{k_1} \binom{p}{k_2}. \tag{4.3}$$

Ignoring the structure of $u = v_1 \otimes v_2$, we define the counterpart $\widetilde{W_0}$ of W_0 as

$$\widetilde{W_0} = \{u \in R^{pq} : \|u\|_2 \leq d_0, \mathrm{supp}(u) \leq 2s_0\},$$

and compare the covering number of W_0 and $\widetilde{W_0}$. It is easy to see that

$$\widetilde{W_0} = \bigcup_{m=1}^{2s_0} \widetilde{U}_m,$$

where $\widetilde{U}_m = \{u \in R^{pq} : \|u\|_2 \leq d_0, \mathrm{supp}(u) = m\}$. By Lemma 2.3 of Mendelson, Pajor, and Tomczak-Jaegermann (2008), for any $0 < \epsilon \leq 1/2$, there exists Π_m an ϵ-cover of $\widetilde{U}_m$ with

$$|\Pi_m| = \left(\frac{5 d_0}{2\epsilon}\right)^m \binom{pq}{m}.$$

Consequently, $\Pi_{\widetilde{W_0}} = \bigcup_{m=1}^{2s_0} \Pi_m$ is the ϵ-cover of $\widetilde{W_0}$ with

$$|\Pi_{\widetilde{W_0}}| \leq \sum_{m=1}^{2s_0} \left(\frac{5 d_0}{2\epsilon}\right)^m \binom{pq}{m}. \tag{4.4}$$

To compare the right side of (4.4) and (4.3), we first compare the behavior of

$$\binom{q}{k_1}\binom{p}{k_2} := a_1 \quad \text{and} \quad \binom{pq}{m} := a_2.$$

By Stirling's formula,

$$a_1 \approx \left[\frac{pq}{(q - k_1)(p - k_2)}\right]^{1/2} \frac{e^{k_1 + k_2} p^p q^q}{(q - k_1)^{q - k_1} (p - k_2)^{p - k_2}} \approx \frac{e^{k_1 + k_2} p^p q^q}{(q - k_1)^{q - k_1} (p - k_2)^{p - k_2}}$$

and

$$a_2 \approx \left[\frac{pq}{pq-m}\right]^{1/2} \frac{e^m(pq)^{pq}}{(pq-m)^{pq-m}} \approx \frac{e^m(pq)^{pq}}{(pq-m)^{pq-m}}.$$

Since $\log(1-x) = x + o(1)$ as $x \to 0$, we have $\log(p-k_1) = \log p + k_1/p$. Consequently, $a_1 \approx q^{k_1}p^{k_2}$ and $a_2 \approx (pq)^m$. The right sides of both (4.4) and (4.3) are finite sums. In addition, k_1, k_2, m are all finite with only $p, q \to \infty$. We need only compare the largest terms of (4.4) and (4.3). The leading term on the right side of (4.4) is

$$\left(\frac{5d_0}{2\epsilon}\right)^{2s_0} \binom{pq}{2s_0}(1+o(1)) \approx \left(\frac{5}{2\epsilon}\right)^{2s_0}(pq)^{2s_0}(1+o(1)).$$

The largest term of the right side of (4.3) is achieved at $(k_1,k_2) = (1, 2s_0)$ or $(k_1,k_2) = (2s_0, 1)$, so the right side of (4.3) is approximately $(15d_0/2\epsilon)\max(p^{2s_0}q, pq^{2s_0})(1+o(1))$. If $\min(p,q) \to \infty$, then $\max(p^{2s_0}q, pq^{2s_0})/(pq)^{2s_0} \to 0$. Therefore, the covering number of the structured W_0 is much smaller than that of $\widetilde{W_0}$ which ignores the structure. Moreover, the difference between $\log|\Pi_{\widetilde{W_0}}|$ and $\log|\Pi_{W_0}|$ tends to $+\infty$, as $\min(p,q) \to \infty$.

5. Simulations

Simulation studies were conducted to compare the proposed structured Lasso method (Struc) and the usual Lasso method in terms of parameter estimation and model selection.

Suppose $(X_i, Y_i), i = 1, \ldots, n$ are *i.i.d.* observations from the model

$$Y_i = \alpha^T X_i \beta + 0.5\epsilon_i,$$

where $\alpha \in R^p$, $\beta \in R^q$, and $\epsilon_i \sim N(0,1)$. We randomly chose n_1 observations as the training data. For performance, we recorded the predicted mean square error (MSE) on the remaining $n - n_1$ testing observations. Let $\hat{\alpha}$ and $\hat{\beta}$ be the estimate of α and β based on the training data. We computed the distance between the true and estimated parameters as

$$\mathrm{ERR} = \|\hat{\beta} \otimes \hat{\alpha} - \beta \otimes \alpha\|_1.$$

For every simulation setup, we generated 100 datasets and computed the mean and standard error of the MSE and ERR. For model selection, we computed the average number of the variables selected, missed, and false positives in 100 replicates, denoted as V_{all}, V_{miss} and V_{fp}, respectively. The optimal tuning parameter λ was selected by 5-fold cross validation on the training data. We took $n = 150$, and set the training sample size $n_1 = 120$, with 30 observations used as the testing data. We considered $(p,q) = (20, 50)$, $(p,q) = (40, 50)$, $(p,q) = (50, 100)$, and $(p,q) = (100, 100)$.

Model 1: We set $\alpha = (4,4,3,2,0\cdots,0)^T \in R^p$ and $\beta = (3,2,0,\ldots,0)^T \in R^q$, and generated $\text{vec}(X_i)$ from $N(0, I_{pq})$.

Model 2: For the model $Y_i = \alpha^T X_i \beta + 0.5\epsilon_i$, we let $X_i = A\tilde{X}_i B$ where $\text{vec}(\tilde{X}_i) \sim N(0, I_{pq})$; $A = (a_{ij})$ with $a_{ij} = 0.2^{|i-j|}$ and $B = (b_{ij})$ with $b_{ij} = 0.6^{|i-j|}$. The parameters α and β were those of Model 1.

Model 3: We took $Y_i = \alpha^T X_i \beta + \beta_{err}^T \text{vec}(X_i) + 0.5\epsilon_i$. Here X_i, α, β were the same as in Model 1; $\beta_{err} = (\mathbf{0}_5^T, v^T, v^T, v^T, v^T, \mathbf{0}_{pq-29}^T)^T \in R^{pq}$, where $v = (c,0,0,0,0,0)^T \in R^6$ and $\mathbf{0}_5 = (0,0,0,0,0)^T$ with $\mathbf{0}_{pq-29}$ defined similarly. We took $c = 0, 0.3, 0.6, 0.9, 1.2, 1.5$. This is Model 1 when $c = 0$.

Model 4: Here Y_i, X_i, α, β were the same as in Model 3, but $\beta_{err} = (\mathbf{0}_5^T, \mathbf{1}_d^T \otimes w^T, \mathbf{0}_{pq-5-6d})^T \in R^{pq}$, $\mathbf{1}_d \in R^d$ is the vector with all elements 1, and $w = (0.5,0,0,0,0,0)^T$. We took $d = 2,4,6,8$; $d = 0$ gives Model 1.

In Table 1, the proposed method shows much better performance than Lasso in both prediction and variable selection. For Model 1, the structured Lasso improves the prediction and the estimation accuracy substantially: for the MSE, improvement ranges from 20% to 30% while the estimation accuracy in terms of ERR improves by about 30% to 50%. In terms of variable selection, the proposed structured Lasso also outperforms Lasso, especially for Model 1. This is reflected in the average number of the variables selected on the average, that of the variables being missed, and the average number of the false positives. Clearly here, taking into account of the matrix structure of the covariates gives estimates and selects variables more accurately.

For Model 3, it is reasonable that for small c, structured Lasso still has good performance and that for large c, Lasso performs better. Simulation results for $p = 20$ and $q = 50$ and for $p = 50$ and $q = 100$ are presented in Table 2, results for other cases were quite similar and are omitted here. Table 2 shows that, when $(p,q) = (20,50)$, structured Lasso outperforms Lasso in both prediction and variable selection for $c \leq 0.3$; and when $c \geq 0.6$, Lasso outperforms structured Lasso. With $(p,q) = (50,100)$, structured Lasso performs better than Lasso for $c \leq 0.9$ and worse than Lasso for $c \geq 1.2$. In addition, when $c = 0.6$ or 0.9, structured Lasso performs better than Lasso for $p = 50$ and $q = 100$, and worse than Lasso for $p = 20$ and $q = 50$; as dimension increases, β_{err} is better approximated by a Keronecker product. This indicates that structured Lasso can have advantages over Lasso as dimensionality increases.

For Model 4, it is clear from Table 3 that when $p = 20$ and $q = 50$, structured Lasso is slightly better than Lasso for $d = 2$, similar to Lasso for $d = 4$, and worse than Lasso for $d \geq 6$. Recall that Model 1 is a special case of Model 4 with $d = 0$, where structured Lasso outperforms Lasso. This shows structured Lasso superior to Lasso when d is small. When $p = 50$ and $q = 100$, we find that for all values

Table 1. Simulation results for Models 1 and 2. For variable selection part, the three rows for the structured Lasso (Struc) method are results on $\beta\otimes\alpha$, α, and β, respectively. In the columns of MSE and ERR, the quantities outside and inside the bracket are the means and standard deviations, respectively.

model	p, q		MSE	ERR	Variable Selection		
					V_{all}	V_{miss}	V_{fp}
1		Struc	3.452 (0.518)	1.120 (0.134)	8.086	0.000	0.086
	$p = 20$				4.000	0.000	0.000
	$q = 50$				2.021	0.000	0.021
		Lasso	4.262(0.697)	1.777 (0.240)	12.586	0.000	4.586
		Struc	3.587 (0.556)	1.190 (0.142)	8.088	0.000	0.088
	$p = 40$				4.014	0.000	0.014
	$q = 50$				2.014	0.000	0.014
		Lasso	4.516 (0.717)	1.935 (0.279)	14.470	0.000	6.470
		Struc	3.648(0.553)	1.191(0.267)	8.301	0.000	0.301
	$p = 50$				4.048	0.000	0.048
	$q = 100$				2.048	0.000	0.048
		Lasso	4.932(0.901)	2.278(0.549)	21.084	0.000	13.084
		Struc	3.748(0.569)	1.215(0.364)	8.500	0.000	0.500
	$p = 100$				4.083	0.000	0.083
	$q = 100$				2.083	0.000	0.083
		Lasso	5.244(0.923)	2.406(0.558)	29.208	0.000	21.208
2		Struc	5.074(0.645)	1.570(0.113)	8.000	0.000	0.000
	$p = 20$				4.000	0.000	0.000
	$q = 50$				2.000	0.000	0.000
		Lasso	5.788(0.697)	1.857(0.127)	8.020	0.000	0.020
		Struc	5.254(0.772)	1.571(0.118)	8.000	0.000	0.000
	$p = 40$				4.000	0.000	0.000
	$q = 50$				2.000	0.000	0.000
		Lasso	5.996(0.852)	1.858(0.130)	8.020	0.000	0.020
		Stru	5.106(0.693)	1.581(0.108)	8.000	0.000	0.000
	$p = 50$				4.000	0.000	0.000
	$q = 100$				2.000	0.000	0.000
		Lasso	5.838(0.846)	1.850(0.140)	8.040	0.000	0.040
		Struc	5.108(0.696)	1.580(0.094)	8.000	0.000	0.000
	$p = 100$				4.000	0.000	0.000
	$q = 100$				2.000	0.000	0.000
		Lasso	5.935(0.963)	1.869(0.146)	8.033	0.000	0.033

of d, structured Lasso is superior to Lasso. This suggests that structured Lasso is more competitive as dimensionality becomes large.

Table 2. Simulation results for Model 3. For variable selection part, V_α and V_β denote the average number of variables selected for α and β, respectively, by the structured Lasso (Struc) method.

p, q	c		MSE	ERR	Variable Selection				
					V_{all}	V_{miss}	V_{fp}	V_α	V_β
p=20 q=50	1.5	Struc	12.142(1.759)	10.344(1.765)	76.201	0.000	65.201	12.108	6.356
		Lasso	5.008(0.643)	2.769(0.438)	21.250	0.000	10.250		
	1.2	Struc	9.729(1.120)	7.004(0.842)	50.109	0.000	39.109	10.554	4.955
		Lasso	4.972(0.752)	2.668(0.386)	20.227	0.000	9.227		
	0.9	Struc	7.137(1.156)	4.627(0.464)	30.206	0.000	19.206	9.650	3.026
		Lasso	4.861(0.535)	2.615(0.325)	18.302	0.000	7.302		
	0.6	Struc	5.862(0.618)	3.249(0.257)	22.805	0.000	11.805	7.159	3.102
		Lasso	5.062(0.668)	2.738(0.538)	20.809	0.000	9.809		
	0.3	Struc	4.369(0.699)	1.945(0.182)	14.052	0.475	3.527	5.156	2.401
		Lasso	4.817(0.677)	2.603(0.415)	17.752	0.675	7.427		
	0	Struc	3.547(0.439)	1.155(0.183)	8.002	0.000	0.002	4.000	2.030
		Lasso	4.383(0.577)	1.819(0.286)	11.956	0.000	3.956		
p=50 q=100	1.5	Struc	11.240(1.710)	10.377(1.401)	81.916	0.000	69.916	17.582	4.504
		Lasso	9.309(7.889)	7.294(7.754)	42.416	0.107	30.523		
	1.2	Struc	9.537(1.898)	7.824(1.061)	61.324	0.000	49.324	15.130	4.043
		Lasso	9.197(4.773)	7.398(4.765)	46.563	0.000	34.563		
	0.9	Struc	7.744(1.273)	5.413(0.942)	40.285	0.000	28.285	12.405	3.220
		Lasso	8.831(4.139)	7.128(4.571)	48.190	0.115	36.305		
	0.6	Struc	5.813(0.785)	3.476(0.323)	25.285	0.000	13.285	10.152	2.502
		Lasso	6.769(1.883)	4.860(1.543)	38.476	0.419	26.895		
	0.3	Struc	4.424(0.703)	2.322(0.258)	16.384	1.000	5.384	7.714	2.095
		Lasso	6.569(1.172)	4.628(0.990)	34.277	1.846	24.123		
	0	Struc	3.537(0.399)	1.134(0.155)	8.262	0.000	0.262	4.042	2.035
		Lasso	4.722(0.740)	2.346(0.509)	20.761	0.000	12.761		

5.1. An analysis of a medical data

It is known that high concentrations of plasma cholesterol and triglyceride are associated with an increased risk of coronary heart disease. The risk of this disease has been shown to be closely related to the distribution of cholesterol and triglyceride in different types of lipoproteins. Therefore, it is important to measure the lipoprotein profile for assessing the risk of coronary heart disease. Ultracentrifugation is the established standard reference method for the separation and analysis of lipoproteins.

A popular approach in medical diagnosis uses the so-called nuclear magnetic resonance (NMR) spectra and the near-infrared spectroscopy (NIR) technology. We take a dataset from a study that represents 2D diffusion-edited 1H NMR spectra obtained from the website `http://www.models.kvl.dk/dosylipo`. The

Table 3. Simulation results for Model 4. For variable selection, V_α and V_β denote the number of selected variable for α and β, respectively, by the structured Lasso (Struc) method.

p, q	d		MSE	ERR	Variable Selection				
					V_{all}	V_{miss}	V_{fp}	V_α	V_β
p=20 q=50	2	Struc	4.364(0.602)	1.946(0.184)	13.400	0.000	3.400	6.333	2.115
		Lasso	4.764(0.704)	2.366(0.376)	17.566	0.000	7.566		
	4	Struc	5.176(0.846)	2.761(0.294)	19.800	0.000	8.800	7.533	2.548
		Lasso	4.932(1.105)	2.788(0.765)	21.133	0.000	10.300		
	6	Struc	6.231(0.841)	4.209(0.394)	34.100	0.300	21.400	8.033	4.324
		Lasso	5.393(0.870)	3.524(0.920)	28.033	0.000	15.033		
	8	Struc	7.140(0.730)	5.707(0.572)	41.666	0.525	27.191	9.166	4.725
		Lasso	5.625(1.380)	4.251(0.996)	32.733	0.105	17.838		
p=50 q=100	2	Struc	4.281(0.591)	1.987(0.185)	13.727	0.000	3.727	6.272	2.181
		Lasso	5.369(0.927)	3.185(0.809)	27.136	0.031	17.167		
	4	Struc	5.493(0.334)	3.047(0.475)	22.882	0.038	10.920	9.052	2.427
		Lasso	7.220(1.910)	5.297(2.066)	38.411	0.423	26.834		
	6	Struc	6.030(0.861)	4.270(0.602)	34.045	0.166	20.211	11.272	3.120
		Lasso	8.294(2.033)	7.200(1.770)	53.000	1.416	40.416		
	8	Struc	6.504(0.974)	5.400(0.514)	47.105	0.444	31.549	13.105	3.509
		Lasso	10.825(2.143)	10.877(2.298)	66.789	3.185	53.974		

detailed background of this dataset is found in Dyrby et al. (2005). The aim of the analysis was to evaluate the potential for quantification of lipoprotein main- and subfractions in human plasma samples. An important role of NMR in this kind of study is to provide complementary information on the classification of lipoprotein fractions compared to ultracentrifugation. The data file contains the NMR data in $\mathbb{X}$ and lipoproteins in $\mathbb{Y}$. Two index files are also provided for selecting specific parts of the data. Since $\mathbb{Y}$ provided is a 25×32 matrix, for illustration purposes, we took the fourth column $\mathbb{Y}_4$ of $\mathbb{Y}$ as our response, which is an important indicator for the quantification. The sample size was $n = 25$ and, for each subject $i (1 \le i \le 25)$, X_i is the NMR spectra, a $24\times$ 1,600 matrix. Thus, the data set is $(\mathbb{X}, \mathbb{Y}_4)$ where $\mathbb{X}$, the $25 \times 24\times$ 1,600 array, consists of $X_i's$. We standardized the data such that the empirical variance of $\mathbb{X}(\cdot, i, j)$ was 1 for $1 \le i \le 24, 1 \le j \le$ 1,600, and the empirical variance of $\mathbb{Y}_4$ was 1.

For comparison purposes, we took 20 random data points as the training data and left the other 5 observations as the testing data. We then fit the structured Lasso and the usual Lasso to the training dataset and recorded the MSE on the testing data, as well as the number of the variables selected. This process was repeated 100 times to compute the means and standard deviations of MSE and the average number of the variables selected in terms of V_{all}, V_α, and V_β, as defined in the simulation studies. Again, the optimal tuning parameter λ was

Table 4. Performance comparison between the structured Lasso (Struc) and the usual Lasso for the data set over 100 random partitioning of the data.

	MSE (SD)	V_{all}	V_α	V_β
Struc	1.951 (0.637)	5.45	2.45	1.95
Lasso	2.282 (0.698)	7.20	–	–

selected via 5-fold cross validation. The results are summarized in Table 4. We see that in terms of prediction, the proposed method outperformed the Lasso; in terms of the variable selection, the structured Lasso gave a smaller model on the average.

6. Discussion

We propose a structured Lasso for matrix covariates and obtain the theoretical property of the estimator. Simulation results confirm the usefulness of the proposed method. The optimization problem in section 2 is non-convex, raising the question of obtaining the global optimal solution. Although our numerical experience indicates satisfactory performance. It is of interest to develop a convex optimization formulation and to derive a faster rate of convergence than the one in this paper. A possible trick is via the relaxing used by d'Aspremont et al. (2007), and to apply the theoretical arguments in Amini and Wainwright (2009). However, the fact that we aim for sparsity in both α and β where they are presented in a product form means their technique may not be directly applicable. Further research to address this non-convexity is needed.

The proofs of the main results can be found in the web appendix available on the Statistica Sinica web site. This web appendix also contains the code and the data used for the data analysis in section 5.1.

Acknowledgement

Zhao's research is supported by National Natural Science Foundation of China (No. 11101022 and No. 11026049) and the Ministry of Education Humanities and Social Science Foundation Youth project (No. 10YJC910013) and the Fundamental Research Funds for the Central Universities. Leng's research is supported by National University of Singapore research grants. We thank Prof. Naisyin Wang, an associate editor, and two referees for their constructive comments.

References

d'Aspremont, A., Ghaoui, L. E., Jordan, M. I. and Lanckriet, G. R. G. (2007). A direct formulation for sparse PCA using semidefinite programming. *SIAM Rev.* **49**, 434-448.

Amini, A. A. and Wainwright, M. J. (2009). High-dimensional analysis of semidefinite relaxations for sparse principal components. *Ann. Statist.* **37**, 2877-2921.

Bickel, P., Ritov, Y. and Tsybakov, A. (2009). Simultaneous analysis of Lasso and Dantzig selector. *Ann. Statist.* **37**, 1705-1732.

Dyrby, M. Peteresen, M. Whittaker, A. D., Lambert, L., Nørgaard, L., Bro, R. and Engelsen, S. B. (2005). Analysis of lipoproteins using 2D diffusion-edited NMR spectroscopy and multi-way chemometrics. *Anal. Chimica Acta* **531**, 209-216.

Fan, J. and Li, R. (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. *J. Amer. Statist. Assoc.* **96**, 1348-1360.

Friedman, J., Hastie, T. and Tibshirani, R. (2010). Regularized paths for generalized linear models via coordinate descent. *J. Statist. Software* **33**.

Ledoux, M. and Talagrand, M. (1991). *Probability in Banach spaces: Isoperimetry and Precesses.* Springer-Verlag, Berlin.

Leng, C. and Tang, C. Y. (2012). Sparse matrix graphical models. *J. Amer. Statist. Assoc.* **107**, 1287-1300.

Li, B., Kim, M. K. and Altman, N. (2010). On dimension folding of matrix or array-valued statistical objects. *Ann. Statist.* **38**, 1094-1121.

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

Mendelson, S., Pajor, A. and Tomczak-Jaegermann, N. (2007). Reconstruction and subgaussian operators in asymptotic geometric analysis. *Geom. Funct. Anal.* **17**, 1248-1282.

Mendelson, S., Pajor, A. and Tomczak-Jaegermann, N. (2008). Uniform uncertainty principle for Bernoulli and subgaussian ensembles. *Constr. Approx.* **28**, 277-289.

Raskutti, G., Wainwright, M. J. and Yu, B. (2010). Restricted eigenvalue properties for correlated gaussian designs. J. Mach. Learn. Res. **11**, 2241-2259.

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

Vershynin, R. (2012). Introduction to the non-asymptotic analysis of random matrices. Chapter 5 in *Compressed sensing, theory and applications* (Edited by Eldar and Kutyniok). Cambridge University Press, Cambridge.

Zhong, W. and Suslick, K. (2012). Penalized classification for matrix predictors with application to colorimetric sensor arrays. *Technometrics.* Accepted.

Zhou, S. (2009). Restricted eigenvalue conditions on subgaussian random matrices. arXiv:0912.4045v2.

LMIB of the Ministry of Education, Beihang University, Beijing 100191, China.

E-mail: zhaojunlong928@126.com

Department of Statistics, University of Warwick, Coventry CV4 7AL, UK.

E-mail: chenlei.leng@nus.edu.sg

(Received January 2012; accepted March 2013)