Yokohama Publishers
ISSN 1349-8169 ONLINE JOURNAL

Pacific Journal of Optimization
Vol. 13, No. 2, pp. 185–196, April 2017

HIGH-DIMENSIONAL LEAST SQUARE MATRIX REGRESSION VIA ELASTIC NET PENALTY*

Bingzhen Chen and Lingchen Kong

Abstract: Matrix regression model was recently developed by Zhou and Li [Journal of the Royal Statistical Society, Series B, 2014, 76: 463-483] and Negahban and Wainwright [Annals of Statistics, 2011, 39(2): 1069-1097] and Obozinski, Wainwright and Jordan [Annals of Statistics, 2011, 39(1): 1-47]. In this paper, we focus on high-dimensional least square matrix regression via elastic net penalty, which can deal with group variables. We show that elastic net-penalized matrix regression has grouping effect property. Moreover, we give a upper bound of deviation between two correlated prediction vectors. Finally we propose the VNS-EN method to compute the estimator of elastic net-penalized matrix regression. We give its convergence results and iteration-complexity. The numerical experiments are reported.

Key words: *high-dimensional matrix regression, matrix least square, matrix elastic net, grouping effect*

Mathematics Subject Classification: *62H12, 65K10, 90C46*

1 Introduction

With the increasing prominence of large-scale data in modern science, data of interest is more complex, which may be in the form of a matrix, not a vector. For instance, two-dimensional digital imaging data record the quantized brightness value of a colour at rows and columns of pixels, and flow cytometric data contain the fluorescence intensity of multiple cells at multiple channels. Both of them are not in the form of vectors but matrices. In particular, as in [16], a well known example is the study of an electroencephalography data set of alcoholism. The study consists of 122 subjects with two groups, an alcoholic group and a normal control group, and each subject was exposed to a stimulus. Voltage values were measured from 64 channels of electrodes placed on the subject's scalp for 256 time points, so each sampling unit is a 256×64 matrix. To address scientific questions arising from those data, sparsity or other forms of regularization are crucial owing to the ultrahigh dimensionality and complex structure of the matrix data. Often, a variety of models in statistics lead to the estimation of matrices with rank constraints. The true signal often has low rank, which can be well approximated by a low rank matrix. Zhou and Li [16] propose a class of regularized matrix regression methods based on spectral regularization. In [16], the authors proposed many regularized matrix regression models with different penalties, such as Power family [5], elastic net [17], log-penalty [1, 2], SCAD [4], MC+-penalty [15]. In fact, matrix regression model can be traced back to earlier literatures by Jordan, Wainwright et

*The work was supported in part by National Natural Science Foundation of China (11431002, 11671029).

© 2017 Yokohama Publishers

al., see, e.g., [8, 11]. They study the same model with block-structured regularization but not consider penalty terms related to singular values of the matrix variable. It happens that Lu [7] explore convex optimization methods for solving matrix regression model with nuclear norm penalty, but not in high-dimensional case.

On the other hand, Zou and Hastie [17] propose the elastic net model in the case of vector. Similar to the lasso, the elastic net simultaneously does automatic variable selection and continuous shrinkage. And more important, the elastic net model can select groups of correlated variables. Note that, Zhou and Li [16] does not consider the grouping effect and other property as in Zou and Hastie [17]. An interesting question occurs: Is there also a similar grouping effect property for elastic net-penalized matrix regression ? If so, can we use the method VNS [7] to give the optimal solution of our model?

This paper focuses on the above questions and gives affirm answer. Based on the results in Zou and Hastie [17], we get the grouping effect property of elastic net-penalized matrix regression. Moreover, we give a upper bound of deviation between two correlated prediction vectors. By reforming the elastic net-penalized matrix regression as nuclear norm matrix regression, we apply the approach in [7] to solve it. Following their terminology, we call our method as VNS-EN algorithm. We give the convergence results and iteration-complexity of VNS-EN algorithm. We also report the numerical experiment results.

The rest of the paper is organized as follows. We formulate elastic net-penalized matrix regression in Section 2 and show its grouping effect property in Section 3. In Section 4, we derive the VNS-EN method and show its convergence and iteration-complexity. Then we do simulation study using the proposed method. We conclude the paper with a discussion of potential future research in Section 5.

2 Elastic Net-Penalized Matrix Regression

We will introduce the elastic net-penalized matrix regression for high-dimensional linear model.

We begin with the following high-dimensional matrix regression model. Consider the matrix regression model

$$B = AX + W, \tag{2.1}$$

where $A \in \mathbb{R}^{n\times m}$ is the predictor matrix, $B \in \mathbb{R}^{n\times q}$ is the response matrix, $X \in \mathbb{R}^{m\times q}$ is the unknown regression matrix, and $W \in \mathbb{R}^{n\times q}$ is the measurement error/noise matrix. When $q = 1$, it is the linear multivariate regression which has been attracted much more research for long times. Throughout the paper, we assume A with each column vector $A_{\cdot j}$ being normalized such that $\|A_{\cdot j}\|_2 = 1$ for $j = 1, 2, \cdots, m$, and $W = (\varepsilon_{ij})$ with all components $\varepsilon_{ij}(i = 1, 2, \cdots, n, j = 1, 2, \cdots, q)$ being independently identically distributed from normal distribution with mean zero and finite variance σ^2.

We will focus on the high dimensional case where the number of observations n is less than the size m and q of unknown coefficients matrix X, i.e., $n < \min\{m, q\}$. Without loss of generality, let $m \le q$. In this setting of $n < m \le q$, the goal is to propose a good estimator of the true coefficient matrix. To do so, a key and common assumption is that the true coefficient matrix X^* is a low-rank matrix, which guarantees the model identifiability and enhances the model fitting accuracy and interpretability, see, e.g, [4, 12]. We assume $r = \mathrm{rank}(X^*)$ with $r < n$. Let the singular value decomposition (SVD) of coefficient matrix X^* be given as

$$X^* = UDV^{\mathrm{T}},$$

where $U \in \mathbb{R}^{m\times m}$, $V \in \mathbb{R}^{q\times q}$ are orthogonal matrices, and D is a diagonal block matrix with $D_{ii} = \lambda_i(X^*)$ and $D_{ij} = 0$ for $i \neq j (i = 1, \cdots, m, j = 1, \cdots, q)$, with the singular values $\lambda_1(X^*) \geq \lambda_2(X^*) \geq \cdots \geq \lambda_r(X^*) > \lambda_{r+1}(X^*) = \cdots = \lambda_m(X^*) = 0$. Note that the above SVD is not unique. Clearly, the parameters X^* and r are implicitly dependent on the sample size n, but we omit the index n in notation whenever there is no confusion.

In order to reconstruct the high-dimensional matrix regression model and estimate the low-rank matrix X^*, we consider the following elastic net-penalized matrix regression problem

$$\min_{X\in\mathbb{R}^{m\times q}} \frac{1}{2}\|B - AX\|_{\mathrm{F}}^2 + \mu_1\|X\|_* + \frac{1}{2}\mu_2\|X\|_{\mathrm{F}}^2, \tag{2.2}$$

where $\|B - AX\|_{\mathrm{F}}^2$ is the matrix least square error loss function, $\|X\|_* = \sum_{i=1}^m \lambda_i(X)$ is the nuclear norm of X, and $\|X\|_{\mathrm{F}} = \sqrt{\mathrm{tr}(X^{\mathrm{T}}X)} = \sqrt{\sum_{i=1}^m \sum_{j=1}^q X_{ij}^2}$ is the Frobenius norm of X, and $\mu_1 \geq 0$ and $\mu_2 \geq 0$ are the penalized/regularization parameters. We call $\mu_1\|X\|_* + \frac{1}{2}\mu_2\|X\|_{\mathrm{F}}^2$ the matrix elastic net penalty term. It is easy to see that [17] deals with the matrix regular via elastic net penalty in case of $q = 1$. When $\mu_2 \equiv 0$, it is the nuclear norm-penalized matrix regression which has been solved by the variant of Nesterov's smooth method when A has full column rank [7]. We define the elastic net-penalized matrix regression estimator as

$$\hat{X} \in \operatorname*{argmin}_{X\in\mathbb{R}^{m\times q}} \frac{1}{2}\|B - AX\|_{\mathrm{F}}^2 + \mu_1\|X\|_* + \frac{1}{2}\mu_2\|X\|_{\mathrm{F}}^2. \tag{2.3}$$

Similarly, it is the nuclear norm-penalized matrix regression estimator when $\mu_2 \equiv 0$.

3 Grouping Effect Property

In this section, we will discuss the grouping effect property of (2.2). In the high-dimensional setting problem, for a long time, there are much attention on the 'grouped variables' situation, which is a particularly important concern for the single linear multivariate regression model, see, e.g., [6, 17] and references therein. For instance, Zou and Hastie [17] showed that in the situation where some variables exhibits the grouping effect, the following generic penalization least-square regression assigns identical coefficients to the identical variables,

$$\hat{x} = \operatorname*{argmin}_{x\in\mathbb{R}^p} \|b - Ax\|_2^2 + \mu h(x),$$

where $b \in \mathbb{R}^n$, and $h(x)$ is positive for $x \neq 0$. This important property is called grouping effect by Zou [17]. As mentioned in the introduction, can we get the grouping effect property for the matrix case?

Consider the following matrix regression estimator counterpart above

$$\hat{X} = \operatorname*{argmin}_{X\in\mathbb{R}^{m\times q}} \|B - AX\|_{\mathrm{F}}^2 + \mu J(X), \tag{3.1}$$

where $J(\cdot)$ is positive valued for $X \neq 0$.

The following result shows a clear distinction between strictly convex penalty functions and the matrix lasso-type penalty, where strict convexity guarantees the grouping effect in the extreme situation with identical predictors. Here we define J is symmetric if $J(PX) = J(X)$ where P is a permutation operator.

Proposition 3.1. *Assume that $A_{\cdot i} = A_{\cdot j}, i, j \in \{1, \cdots, m\}$. Let $\hat{X}$ be the penalized matrix regression estimator of* (3.1).

(a) *If $J(\cdot)$ is strictly convex and symmetric, then $X_{i\cdot} = X_{j\cdot}$ for any $\mu > 0$.*

(b) *If $J(X) = \|X\|_*$, then $\hat{X}^*$ is another minimizer of the penalized matrix regression estimator of* (3.1)*, where*

$$\hat{X}^*_{k\cdot} = \begin{cases} \hat{X}_{k\cdot}, & if\ k \neq i \text{ and } k \neq j, \\ t\hat{X}_{i\cdot} + (1-t)\hat{X}_{j\cdot}, & if\ k = i, \\ (1-t)\hat{X}_{i\cdot} + t\hat{X}_{j\cdot}, & if\ k = j. \end{cases}$$

for any $t \in [0, 1]$.

Proof. (a) Fix $\mu > 0$. Assume $X_{i\cdot} \neq X_{j\cdot}$. Let us define $\tilde{X}$ as follows

$$\tilde{X}_{k\cdot} = \begin{cases} \hat{X}_{k\cdot}, & if\ k \neq i\ and\ k \neq j, \\ \hat{X}_{j\cdot}, & if\ k = i, \\ \hat{X}_{i\cdot}, & if\ k = j. \end{cases} \tag{3.2}$$

This together with the assumption $A_{\cdot i} = A_{\cdot j}$ derives $A\hat{X} = A\tilde{X}$. Notice that $\tilde{X}$ is just a permutation of $\hat{X}$ and J is symmetric, $J(\hat{X}) = J(\tilde{X})$. Thus we obtain that

$$\|B - A\tilde{X}\|_{\mathrm{F}}^2 + \mu J(\tilde{X}) = \|B - A\hat{X}\|_{\mathrm{F}}^2 + \mu J(\hat{X}).$$

Take the convex combination of $\tilde{X}$ and $\hat{X}$, $\bar{X}_t = t\tilde{X} + (1-t)\hat{X}$ for $t \in (0, 1)$. Clearly, from the strictly convexity of $\|B - AX\|_{\mathrm{F}}^2 + \mu J(X)$, we easily obtain that

$$\|B - A\bar{X}_t\|_{\mathrm{F}}^2 + \mu J(\bar{X}_t) < t\left[\|B - A\tilde{X}\|_{\mathrm{F}}^2 + \mu J(\tilde{X})\right] + (1-t)\left[\|B - A\hat{X}\|_{\mathrm{F}}^2 + \mu J(\hat{X})\right].$$

This together with above equality yields that

$$\|B - A\bar{X}_t\|_{\mathrm{F}}^2 + \mu J(\bar{X}_t) < \|B - A\hat{X}\|_{\mathrm{F}}^2 + \mu J(\hat{X}).$$

This is a contradiction of $\hat{X}$ is the solution of (3.1). Hence $X_{i\cdot} = X_{j\cdot}$.

(b) From the proof of (a), and the fact that $J(X) = \|X\|_*$ is not strictly convex, we get that $\|B - AX\|_{\mathrm{F}}^2 + \mu\|X\|_*$ is convex but not strictly convex. For the same $\tilde{X}$ in (3.2) and $\bar{X}_t = t\tilde{X} + (1-t)\hat{X}$ with $t \in (0, 1)$, we have

$$\|B - A\bar{X}_t\|_{\mathrm{F}}^2 + \mu J(\bar{X}_t) \leq t\left[\|B - A\tilde{X}\|_{\mathrm{F}}^2 + \mu J(\tilde{X})\right] + (1-t)\left[\|B - A\hat{X}\|_{\mathrm{F}}^2 + \mu J(\hat{X})\right].$$

Observing that the fact of $\tilde{X}$ and $\hat{X}$ being minimizers of the penalized matrix regression estimator $\|B - AX\|_{\mathrm{F}}^2 + \mu\|X\|_*$, we obtain

$$\|B - A\bar{X}_t\|_{\mathrm{F}}^2 + \mu J(\bar{X}_t) \leq \|B - A\hat{X}\|_{\mathrm{F}}^2 + \mu J(\hat{X}).$$

Hence the desired conclusion follows immediately. □

Remark 3.2. As we can see $\|X\|_* + \frac{1}{2}\|X\|_{\mathrm{F}}^2$ is strictly convex and symmetric, elastic net-penalized matrix regression (2.2) has grouping effect.

It is easy to note that the proposition 3.1 and its proof are different from those in the vector setting in [17]. However, the above results have the similar implication as in the vector case, see, e.g., [3, 12, 17] for further explanations. That is, for the matrix regression, the matrix lasso-type penalty does not even have a unique solution, while the matrix elastic net penalty with $\mu_2 > 0$ is strictly convex and possesses the above grouping effect property.

Furthermore, the elastic net-penalized matrix regression method can provide a quantitative description for the grouping effect as in vector case. We below show that the difference between the i-th and j-th row coefficient paths of predictors is almost 0 if $A_{\cdot i}$ and $A_{\cdot j}$ are highly correlated, i.e. $\rho \doteq 1$ (if $\rho \doteq -1$ then consider $-A_{\cdot j}$), where $\rho = A_{\cdot i}^{\mathrm{T}} A_{\cdot j}$ is the sample correlation.

Theorem 3.3. *Given data (B, A) and parameters (μ_1, μ_2), let $\hat{X}$ be the elastic net-penalized matrix regression estimator,*

$$\hat{X} = \underset{X \in \mathbb{R}^{m \times q}}{\operatorname{argmin}} \frac{1}{2}\|B - AX\|_{\mathrm{F}}^2 + \mu_1 \|X\|_* + \frac{1}{2}\mu_2 \|X\|_{\mathrm{F}}^2. \tag{3.3}$$

Suppose that $\|B\|_{\mathrm{F}} \neq 0$ and $\mu_2 \neq 0$. Then

$$\frac{1}{\|B\|_{\mathrm{F}}}\|\hat{X}_{i\cdot} - \hat{X}_{j\cdot}\|_2 \leq \frac{1}{\mu_2}\sqrt{2(1-\rho)}.$$

Proof. Let $L(\mu_1, \mu_2, X) = \frac{1}{2}\|B - AX\|_{\mathrm{F}}^2 + \mu_1\|X\|_* + \frac{1}{2}\mu_2\|X\|_{\mathrm{F}}^2$, and let the singular value decomposition (SVD) of $\hat{X} = U\Xi V^{\mathrm{T}}$ with $\Xi_{ii} = \lambda_i(\hat{X}) \geq 0$ and $\Xi_{ij} = 0$ for $i \neq j$. Considering the subdifferential of $\|\cdot\|_*$ in [14], we obtain the first-order optimality condition of (3.3)

$$0 \in \frac{\partial L(\mu_1, \mu_2, \hat{X})}{\partial X} = A^{\mathrm{T}}(A\hat{X} - B) + \mu_1 U \mathrm{Sign}(\Xi) V^{\mathrm{T}} + \mu_2 \hat{X},$$

where $(\mathrm{Sign}(\Xi))_{ii} = \mathrm{sign}(\lambda_i(\hat{X}))$ and its other components are zero. Set $\mathrm{rank}(\hat{X}) = r$. Thus, there is a nonnegative matrix D with $D_{ii} = 1$ for $i \in \{1, 2, \cdots, r\}$, $D_{ii} \in [0, 1]$ for $i \in \{r+1, r+2, \cdots, m\}$, and $D_{ij} = 0$ for $i \neq j$ such that

$$A^{\mathrm{T}}(A\hat{X} - B) + \mu_1 U D V^{\mathrm{T}} + \mu_2 \hat{X} = 0.$$

Let $e_m^{(k)} = (0, \cdots, 0, 1, 0, \cdots, 0)^{\mathrm{T}} \in \mathbb{R}^m$ with all zero but the k-th component one. Then

$$[e_m^{(k)}]^{\mathrm{T}} A^{\mathrm{T}}(A\hat{X} - B) + \mu_1 [e_m^{(k)}]^{\mathrm{T}} U D V^{\mathrm{T}} + \mu_2 [e_m^{(k)}]^{\mathrm{T}} \hat{X} = 0.$$

Taking $k = i$ and $k = j$, by direct computation we obtain

$$\begin{aligned} A_{\cdot i}^{\mathrm{T}}(A\hat{X} - B) + \mu_1 U_{i\cdot} D V^{\mathrm{T}} + \mu_2 \hat{X}_{i\cdot} &= 0, \\ A_{\cdot j}^{\mathrm{T}}(A\hat{X} - B) + \mu_1 U_{j\cdot} D V^{\mathrm{T}} + \mu_2 \hat{X}_{j\cdot} &= 0. \end{aligned} \tag{3.4}$$

Thus, we have

$$(A_{\cdot i}^{\mathrm{T}} - A_{\cdot j}^{\mathrm{T}})(A\hat{X} - B) + \mu_1 (U_{i\cdot} - U_{j\cdot}) D V^{\mathrm{T}} + \mu_2 (\hat{X}_{i\cdot} - \hat{X}_{j\cdot}) = 0.$$

Then on both sides of the above equation, taking the inner product with the row vector $\hat{X}_{i\cdot} - \hat{X}_{j\cdot}$, we obtain

$$(\hat{X}_{i\cdot} - \hat{X}_{j\cdot}) \left[(A_{\cdot i}^{\mathrm{T}} - A_{\cdot j}^{\mathrm{T}})(A\hat{X} - B)\right]^{\mathrm{T}} + \mu_1 (\hat{X}_{i\cdot} - \hat{X}_{j\cdot}) \left[(U_{i\cdot} - U_{j\cdot}) D V^{\mathrm{T}}\right]^{\mathrm{T}} + \mu_2 \|\hat{X}_{i\cdot} - \hat{X}_{j\cdot}\|^2 = 0. \tag{3.5}$$

Noting that $\hat{X}_{i\cdot} - \hat{X}_{j\cdot} = \left([e_m^{(i)}]^{\mathrm{T}} - [e_m^{(j)}]^{\mathrm{T}}\right) U\Xi V^{\mathrm{T}}$ and $U_{i\cdot} - U_{j\cdot} = \left([e_m^{(i)}]^{\mathrm{T}} - [e_m^{(j)}]^{\mathrm{T}}\right) U$, by direct calculation, we obtain

$$\begin{aligned} & (\hat{X}_{i\cdot} - \hat{X}_{j\cdot}) \left((U_{i\cdot} - U_{j\cdot})DV^{\mathrm{T}}\right)^{\mathrm{T}} \\ = & \left([e_m^{(i)}]^{\mathrm{T}} - [e_m^{(j)}]^{\mathrm{T}}\right) U\Xi V^{\mathrm{T}} V D^{\mathrm{T}} U^{\mathrm{T}} (e_m^{(i)} - e_m^{(j)}) \\ = & \left([e_m^{(i)}]^{\mathrm{T}} - [e_m^{(j)}]^{\mathrm{T}}\right) U\Xi D^{\mathrm{T}} U^{\mathrm{T}} (e_m^{(i)} - e_m^{(j)}). \end{aligned} \tag{3.6}$$

This together with the fact that $U\Xi D^{\mathrm{T}} U^{\mathrm{T}}$ is positive semidefinite yields

$$(\hat{X}_{i\cdot} - \hat{X}_{j\cdot}) \left((U_{i\cdot} - U_{j\cdot})DV^{\mathrm{T}}\right)^{\mathrm{T}} \geq 0. \tag{3.7}$$

Combining the above arguments (3.5), (3.7) and $\mu_1 \geq 0$, we obtain

$$\mu_2 \|\hat{X}_{i\cdot} - \hat{X}_{j\cdot}\|^2 \leq \left| (\hat{X}_{i\cdot} - \hat{X}_{j\cdot}) \left[(A_{\cdot i}^{\mathrm{T}} - A_{\cdot j}^{\mathrm{T}})(A\hat{X} - B) \right]^{\mathrm{T}} \right|.$$

Then,

$$\mu_2 \|\hat{X}_{i\cdot} - \hat{X}_{j\cdot}\|^2 \leq \|\hat{X}_{i\cdot} - \hat{X}_{j\cdot}\|_2 \|(A_{\cdot i}^{\mathrm{T}} - A_{\cdot j}^{\mathrm{T}})(A\hat{X} - B)\|_2.$$

On the other hand,

$$\frac{1}{2}\|B - A\hat{X}\|_{\mathrm{F}} + \mu_1 \|\hat{X}\|_* + \frac{1}{2}\mu_2 \|\hat{X}\|_{\mathrm{F}}^2 \leq \frac{1}{2}\|B\|_{\mathrm{F}}.$$

Thus, $\|B - A\hat{X}\|_{\mathrm{F}} \leq \|B\|_{\mathrm{F}}$. Therefore,

$$\begin{aligned} & \frac{1}{\|B\|_{\mathrm{F}}} \|\hat{X}_{i\cdot} - \hat{X}_{j\cdot}\|_2 \\ \leq & \frac{1}{\mu_2 \|B\|_{\mathrm{F}}} \|(A_{\cdot i}^{\mathrm{T}} - A_{\cdot j}^{\mathrm{T}})(A\hat{X} - B)\|_2 \\ \leq & \frac{1}{\mu_2 \|B\|_{\mathrm{F}}} \|A_{\cdot i}^{\mathrm{T}} - A_{\cdot j}^{\mathrm{T}}\|_2 \cdot \|A\hat{X} - B\|_{\mathrm{F}} \\ \leq & \frac{1}{\mu_2} \|A_{\cdot i} - A_{\cdot j}\|_2 \\ \leq & \frac{1}{\mu_2} \sqrt{2(1-\rho)}. \end{aligned} \tag{3.8}$$

The proof is complete. □

4 Optimal Method

In this section, we give VNS-EN method for solving (2.2) based on [7]. Following the ideas from [7], we give the convergence results and iteration-complexity of the VNS-EN method. Lu et al. [7] explore convex optimization methods VNS for solving nuclear norm-penalized matrix regression model, which is a variant due to Nesterov's smooth method [9, 10]. Note that there is a precondition that the predictor matrix must have full rank. It is not possible under high-dimensional data. We find that our elastic net-penalized matrix regression model can be transformed into nuclear norm-penalized matrix regression model. And then the new predictor matrix has full column rank. So, we can use VNS to solve elastic net-penalized matrix regression.

4.1 VNS-EN Algorithm

We give the VNS-EN method for solving (2.2) and discuss its convergence, iteration-complexity.

There is a closely related connection between the elastic net-penalized matrix regression and the nuclear norm-penalized matrix regression as in the linear multivariate regression model in [17]. In fact, letting $B^\star = \begin{pmatrix} B \\ O \end{pmatrix}, A^\star = \frac{1}{\sqrt{1+\mu_2}} \begin{pmatrix} A \\ \sqrt{\mu_2} I \end{pmatrix}$, it is easy to see that elastic net-penalized matrix regression problem (2.2) can be seen as nuclear norm-penalized matrix regression

$$\min_{X^\star \in \mathbb{R}^{m\times q}} \frac{1}{2}\|B^\star - A^\star X^\star\|_{\mathrm{F}}^2 + \gamma \|X^\star\|_*, \tag{4.1}$$

where $X^\star = \sqrt{1+\mu_2} X, \gamma = \frac{\mu_1}{\sqrt{1+\mu_2}}$. If $\hat{X}^\star \in \underset{X^\star \in \mathbb{R}^{m\times q}}{\operatorname{argmin}} L(\gamma, X^\star)$, then $\hat{X} = \frac{1}{\sqrt{1+\mu_2}} \hat{X}^\star$.

For any $\mu_2 > 0$, no matter how much the rank(A) is, the matrix $A^\star$ has full column rank, so we can use the method VNS in [7] to solve (4.1). In order to reduce the amount of computation, we first simplify the problem (4.1).

Since $A^\star$ has a full column rank, there exists an orthonormal matrix $Q \in \mathbb{R}^{m\times m}$ and a positive diagonal matrix $\Lambda \in \mathbb{R}^{m\times m}$ such that $A^{\star\mathrm{T}} A^\star = Q\Lambda^2 Q^{\mathrm{T}}$. Letting $\breve{X} = Q^{\mathrm{T}} X^\star, H = \Lambda^{-1} Q^{\mathrm{T}} A^{\star\mathrm{T}} B^\star$, then (4.1) is equivalent to

$$\min_{\breve{X} \in \mathbb{R}^{m\times q}} \frac{1}{2}\|\Lambda \breve{X} - H\|_{\mathrm{F}}^2 + \gamma \|\breve{X}\|_*. \tag{4.2}$$

Now, we present the VNS-EN method step by step. We first need to present the following results which are easy to induce from [7]. So, we omit the proofs.

Lemma 4.1. *For every $\gamma > 0$, problem (4.2) has a unique optimal solution $\breve{X}_\lambda^*$. Moreover,*

$$\|\breve{X}_\lambda^*\|_{\mathrm{F}} \le \|\breve{X}_\lambda^*\|_* \le r_{\breve{X}} := \min\left\{ \frac{\|H\|_{\mathrm{F}}^2}{2\gamma},\ \|\Lambda^{-1} H\|_* \right\}.$$

Theorem 4.2. *For some $\epsilon \ge 0$, assume that X_ϵ is an ϵ-optimal solution of the smooth saddle point problem*

$$\min_{\breve{X} \in \mathbf{B}_{\mathrm{F}}^{m\times q}(r_{\breve{X}})} \quad \max_{W \in \Omega_1} \left\{ \frac{1}{2}\|\Lambda \breve{X} - H\|_{\mathrm{F}}^2 + \gamma m \mathrm{tr}\left(W^{\mathrm{T}} \mathbf{G}(\breve{X}) \right) \right\}, \tag{4.3}$$

where $\Omega_1 = \{W \in S^{m+q} | 0 \preceq W \preceq I/m,\ \mathrm{tr}(W) = 1\}$, $\mathbf{B}_{\mathrm{F}}^{m\times q}(r_{\breve{X}}) = \left\{ \breve{X} \in \mathbb{R}^{m\times q} | \|\breve{X}\|_{\mathrm{F}} \le r_{\breve{X}} \right\}$, $\mathbf{G}(\breve{X}) = \begin{pmatrix} O_q & \breve{X}^{\mathrm{T}} \\ \breve{X} & O_m \end{pmatrix}$. *Then X_ϵ is an $\epsilon-$optimal solution of* (4.2),

According to Theorem 4.2, we only need to solve (4.3) rather than (4.2). But as discussed in [7], owing to the objective function has Lipschitz continuous gradient, we would like to solve the dual problem of (4.3), i.e.

$$\max_{W \in \Omega_1} \quad \min_{\breve{X} \in \mathbf{B}_{\mathrm{F}}^{m\times q}(r_{\breve{X}})} \left\{ \frac{1}{2}\|\Lambda \breve{X} - H\|_{\mathrm{F}}^2 + \gamma m \mathrm{tr}\left(W^{\mathrm{T}} \mathbf{G}(\breve{X}) \right) \right\}. \tag{4.4}$$

By scaling the variable $\breve{X}$ of (4.4) as $\breve{X} \leftarrow \breve{X}/r_{\breve{X}}$, and multiplying the resulting formulation by -1, then we can reformulate (4.4) as the problem

$$\min_{W \in \Omega_1} \quad \max_{\breve{X} \in \mathbf{B}_{\mathrm{F}}^{m\times q}(1)} \left\{ -\gamma m r_{\breve{X}} \mathrm{tr}\left(W^{\mathrm{T}} \mathbf{G}(\breve{X}) \right) - \frac{1}{2}\|r_{\breve{X}} \Lambda \breve{X} - H\|_{\mathrm{F}}^2 \right\}. \tag{4.5}$$

Let $\phi(W,\breve{X}) := -\gamma m r_{\breve{X}}\mathrm{tr}\left(W^{\mathrm{T}}\mathbf{G}(\breve{X})\right) - \frac{1}{2}\|r_{\breve{X}}\Lambda\breve{X} - H\|_{\mathrm{F}}^2$, $N := m+q$, for every $W \in \Omega_1$, the function $\phi(W,\cdot) : \mathbf{B}_{\mathrm{F}}^{m\times q}(1) \to \mathbb{R}$ is strictly concave. Thus

$$\max_{\breve{X}\in\mathbf{B}_{\mathrm{F}}^{m\times q}(1)} \left\{-\gamma m r_{\breve{X}}\mathrm{tr}\left(W^{\mathrm{T}}\mathbf{G}(\breve{X})\right) - \frac{1}{2}\|r_{\breve{X}}\Lambda\breve{X} - H\|_{\mathrm{F}}^2\right\}$$

has the unique solution $\breve{X}(W)$. Selecting the prox-function $p_{\Omega_1}(\cdot)$ for the set Ω_1 as $p_{\Omega_1}(W) = \mathrm{tr}\,(W\log W) + \log(N)$, then

$$W_0 := \operatorname*{argmin}_{W\in\Omega_1} \; p_{\Omega_1}(W) = I/N,\; D_{\Omega_1} := \max\{p_{\Omega_1}(W) : W \subseteq \Omega_1\} = \log(N/m)$$

and $p_{\Omega_1}(W)$ is a differentiable strongly convex function with modulus $\sigma_{\Omega_1} = m$.

To deal with $p_{\Omega_1}(W)$, we define

$$d_{p_{\Omega_1}}(W;\tilde{W}) = p_{\Omega_1}(W) - l_{p_{\Omega_1}}(W;\tilde{W}), \forall W,\tilde{W}\in\Omega_1, \tag{4.6}$$

where $l_{p_{\Omega_1}}(W;\tilde{W}) = p_{\Omega_1}(\tilde{W}) + \left\langle \nabla p_{\Omega_1}(\tilde{W}), W-\tilde{W}\right\rangle$. Similarly, we can define the function $l_f(\cdot,\cdot)$ for $f(W) := \max\limits_{\breve{X}\in\mathbf{B}_{\mathrm{F}}^{m\times q}(1)} \phi(W,\breve{X}), \forall W\in\Omega_1$ as

$$l_f(W;\tilde{W}) = f(\tilde{W}) + \left\langle \nabla f(\tilde{W}), W-\tilde{W}\right\rangle, \tag{4.7}$$

where $\nabla f(\tilde{W})$ can be computed as $\nabla f(\tilde{W}) = \nabla_W\phi\left(W,\breve{X}(W)\right)$. One can verify that $f(W)$ is L-Lipschitz-differentiable on Ω_1 with $L = 2\gamma^2m^2\|\Lambda^{-1}\|_{\mathrm{F}}^2$.

Before stating our algorithm, we need the following assumption

$$0 < \alpha_k \le \left(\sum_{i=0}^{k}\alpha_i\right)^{1/2}, \forall k \ge 0, \tag{4.8}$$

where $\{\alpha_k\}_{k\ge0}$ is a sequence of scalars.

Now, we present the method VNS-EN for our problem (4.5).

VNS-EN Algorithm

(0) Set $W_0^{sd} = W_0, \breve{X}_0 = 0, \tau_0 = 1$ and $k = 1$;
(1) Compute $\breve{X}(W_{k-1})$ and $\nabla f(W_{k-1})$;
(2) Compute $\left(W_k^{sd}, W_k^{ag}\right) \subseteq \Omega_1\times\Omega_1$ and $\breve{X}_k \subseteq \mathbf{B}_{\mathrm{F}}^{m\times q}(1)$ as

$$\breve{X}_k = (1-\tau_{k-1})\breve{X}_{k-1} + \tau_{k-1}\breve{X}(W_{k-1}),$$
$$W_k^{ag} = \operatorname*{argmin}_{W\in\Omega_1}\left\{\frac{L}{\sigma_{\Omega_1}}d_{p_{\Omega_1}}(W;W_0) + \sum_{i=0}^{k-1}\alpha_i l_f(W;W_i)\right\}, \left(d_{p_{\Omega_1}}\text{in}(4.6), l_f\text{in}(4.7)\right)$$
$$W_k^{sd} = (1-\tau_{k-1})W_{k-1}^{sd} + \tau_{k-1}W_k^{ag};$$

(3) Set $\tau_k = \frac{\alpha_k}{\sum_{i=0}^{k}\alpha_i}$ and $W_k = (1-\tau_k)W_k^{sd} + \tau_k W_k^{ag}$;
(4) Set $k \leftarrow k+1$ and go to step (1).

We below give the convergence result for the above VNS-EN algorithm. Its proof is easy from Corollary 3 of Tseng [13].

Theorem 4.3. *The sequence* $\{(W_k^{sd}, \breve{X}_k)\} \subseteq \Omega_1 \times \mathbf{B}_{\mathrm{F}}^{m\times q}(1)$ *generated by VNS-EN algorithm satisfies*

$$0 \leq f(W_k^{sd}) - g(\breve{X}_k) \leq \frac{LD_{\Omega_1}}{\sigma_{\Omega_1}(\sum_{i=0}^{k-1}\alpha_i)} = \frac{2\gamma^2 m^2 \|\Lambda^{-1}\|_{\mathrm{F}}^2 \log(N/m)}{m(\sum_{i=0}^{k-1}\alpha_i)}, \forall k \leq 1,$$

where $g(\breve{X}) := \min_{W\in\Omega_1} \phi(W, \breve{X})$, for all $\breve{X}_k \subseteq \mathbf{B}_{\mathrm{F}}^{m\times q}(1)$. A typical sequence α_k satisfying (4.8) is the one in which $\alpha_k = (k+1)/2$ for all $k \geq 0$. With this choice for α_k, we have the following specialization of Theorem 4.3.

Corollary 4.4. *If* $\alpha_k = (k+1)/2$ *for every* $k \geq 0$, *then the sequence* $\{(W_k^{sd}, \breve{X}_k)\} \subseteq \Omega_1 \times \mathbf{B}_{\mathrm{F}}^{m\times q}(1)$ *generated by VNS-EN algorithm satisfies*

$$0 \leq f(W_k^{sd}) - g(\breve{X}_k) \leq \frac{4LD_{\Omega_1}}{\sigma_{\Omega_1}k(k+1)} = \frac{8\gamma^2 m^2 \|\Lambda^{-1}\|_{\mathrm{F}}^2 \log(N/m)}{mk(k+1)}, \forall k \leq 1.$$

From corollary 4.4, we obtain the following iteration-complexity theorem.

Theorem 4.5. *For a given* $\epsilon > 0$, *the iteration-complexity of finding an* ϵ*-optimal solution to* (4.3) *and its dual* (4.4) *by VNS-EN algorithm does not exceed*

$$\left\lceil \frac{2\sqrt{2}\gamma\|\Lambda^{-1}\|_{\mathrm{F}}}{\sqrt{\epsilon}}\sqrt{m\log(N/m)} \right\rceil = \left\lceil \frac{2\sqrt{2}\gamma\|(A^{\star\mathrm{T}}A^{\star})^{-1/2}\|_{\mathrm{F}}}{\sqrt{\epsilon}}\sqrt{m\log(N/m)} \right\rceil.$$

We observe that the iteration-complexity given in Theorem 4.5 is in terms of the transformed data of problem (4.2) or (4.1). We next relate it to the original data of problem (2.2).

Corollary 4.6. *For a given* $\epsilon > 0$, *the iteration-complexity of finding an* ϵ*-optimal solution to* (4.3) *and its dual* (4.4) *by VNS-EN algorithm does not exceed*

$$\left\lceil \frac{2\sqrt{2}\mu_1\|(A^{\mathrm{T}}A + \mu_2 I)^{-1/2}\|_{\mathrm{F}}}{\sqrt{\epsilon}}\sqrt{m\log(N/m)} \right\rceil.$$

4.2 Numerical Experiment

We now report the results of our computational experiment using the VNS-EN algorithm for solving elastic net-penalized matrix regression (2.2) on a set of randomly generated instances.

The random instances of (2.2) used in our experiments were generated as follows. We first randomly generated matrices $A \in \mathbb{R}^{n\times m}, m = 2n$ with entries uniformly distributed in $[0, 1]$ and $B \in \mathbb{R}^{n\times q}, q = 5n$ with entries from the standard normal distribution. We then computed H for (4.2) according to the procedures described in Sec. 4.1. In addition, all computations were performed on an Intel Core(TM)i7-2640M CPU (2.80 GHz) and 8 GB RAM. The code for VNS is written in MATLAB, and the initial point for this method is set to be $W_0 = I/(m+q)$. The method VNS-EN terminates once the duality gap is less than $\epsilon = 10^{-8}$.

The performance of VNS-EN for our randomly generated instances is presented in Table 1 with different (μ_1, μ_2). The problem size (n, m, q) is given in column one. The numbers of iterations of VNS-EN are given in column two. CPU times (in seconds) are given in column three, and the amount of memory (in mega bytes) used are given in column four. In the last column, the rank of optimization are presented.

In Table 2, we present the performance of problem $(10, 20, 50)$ with different (μ_1, μ_2). The first column is the different values of (μ_1, μ_2). The rest columns are the same as in Table 1.

Table 1 Performance of VNS-EN with $\mu_1 = 0.02, \mu_2 = 0.44$

Problem (n, m, q)	**Iteration**	**Time**	**Rank of optimization solution**
$(10, 20, 50)$	119	2.06	10
$(50, 100, 250)$	258	13.29	50
$(100, 200, 500)$	360	40.93	100
$(150, 300, 750)$	448	127.20	150
$(200, 400, 1000)$	505	300.08	200

Table 2 Performance of VNS-EN Algorithm for solving $(10, 20, 50)$

(μ_1, μ_2)	**Iteration**	**Time**	**Rank of optimization solution**
$(0.025, 0.44)$	215	2.51	10
$(0.05, 0.44)$	1042	13.87	10
$(0.1, 0.44)$	5798	53.98	10
$(0.2, 0.44)$	16326	75.99	10
$(8, 63)$	50136	461.70	16
$(32, 1023)$	20390	141.04	10
$(128, 16383)$	88	1.17	10

5 Concluding Remarks

In this paper, we studied the elastic net-penalized matrix regression in high-dimensional case. We show the grouping effect property of this model. Following the ideas from [7], we gave a VNS-EN method to solve the elastic net-penalized matrix regression (2.2). Here we assume $n < m \leq q$. As we can see, if we solve the elastic net-penalized matrix regression (2.2) directly, we only deal with $n \times q$ matrices. But when we use VNS-EN algorithm to solve the elastic net-penalized matrix regression (2.2), we should make a reformation of data set. In this case, we must manage $(n + m) \times q$ matrices. Therefore, when we use VSN-EN method, we must cope with matrices of higher dimension. And we must handle heavier computational task. So, in the future one can develop an effective algorithm of solving elastic net-penalized matrix regression directly.

Acknowledgments

We would like to thank two anonymous referees and the associate editor for their insightful comments and suggestions, which improved the presentation of the paper.

References

[1] A. Armagan, D. Dunson and J. Lee, Generalized double Pareto shrinkage, *Statist. Sinica* 23 (2013) 119–143.

[2] E. J. Candés, M. B. Wakin and S. P. Boyd, Enhancing sparsity by reweighted l_1 minimization, *J. Fourier Anal. Appl.* 14 (2008) 877–905.

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

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

[5] I. E. Frank and J. H. Friedman, A statistical view of some chemometrics regression tools, *Technometrics* 35 (1993) 109–135.

[6] T. Hastie, R. Tibshirani, M. Eisen, P. Brown, D. Ross, U. Scherf, J. Weinstein, A. Alizadeh, L. Staudt and D. Botstein, 'Gene shaving' as a method for identifying distinct sets of genes with similar expression patterns, *Genome Biol.* 1(2) (2000) 1–21.

[7] Z. Lu, R. D. C. Monteiro and M. Yuan, Convex optimization methods for dimension reduction and coefficient estimation in multivariate linear regression, *Math. Program.* 131 (2012) 163–194.

[8] S. Negahban and M. J. Wainwright, Estimation of (near) low-rank matrices with noise and high-dimensional scaling, *Ann. Statist.* 39(2) (2011) 1069–1097.

[9] Y. E. Nestero'v, A method for unconstrained convex minimization problem with the rate of convergence $O(1/k^2)$, *Soviet Math. Doklady* 269 (1983) 543–547.

[10] Y. E. Nestero'v, Smooth minimization of nonsmooth functions, *Math. Program.* 103 (2005) 127–152.

[11] G. Obozinski, M. J. Wainwright and M. I. Jordan, Support union recovery in high-dimensional multivariate regression, *Ann. Statist.* 39(1) (2011) 1–47.

[12] R. Tibshirani, Regression shrinkage and selection via the Lasso, *J. R. Stat. Soc. Ser. B Stat. Methodol.* 58 (1996) 267–288.

[13] P. Tseng, On accelerated proximal gradient methods for convex-concave optimization, submitted to *SIAM J. Optim.* (2008) .

[14] G. A. Watson, Characterization of the subdifferential of some matrix norms, *Linear Algebra Appl.* 170 (6) (1992) 33-45.

[15] C. H. Zhang, Nearly unbiased variable selection under minimax concave penalty, *Ann. Statist.* 38 (2010) 894-942.

[16] H. Zhou and L. Li, Regularized matrix regression, *J. R. Stat. Soc. Ser. B Stat. Methodol.* 76 (2014) 463-483.

[17] H. Zou, and T. Hastie, Regularization and variable selection via the elastic net, *J. R. Stat. Soc. Ser. B Stat. Methodol.* 67 (2005) 301-320.

Manuscript received 30 September 2016
revised February 2017
accepted for publication 16 March 2017

Bingzhen Chen
Department of Applied Mathematics
Beijing Jiaotong University
Beijing 100044, P. R. China
E-mail address: chenbingzhen6026@163.com

Lingchen Kong
Department of Applied Mathematics
Beijing Jiaotong University
Beijing 100044, P. R. China
E-mail address: lchkong@bjtu.edu.cn