Communications for Statistical Applications and Methods
2015, Vol. 22, No. 1, 41–54

DOI: http://dx.doi.org/10.5351/CSAM.2015.22.1.041
Print ISSN 2287-7843 / Online ISSN 2383-4757

Variable Selection with Nonconcave Penalty Function on Reduced-Rank Regression

Sang Yong Jung[a], Chongsun Park[1,a]

[a]Department of Statistics, Sungkyunkwan University, Korea

Abstract

In this article, we propose nonconcave penalties on a reduced-rank regression model to select variables and estimate coefficients simultaneously. We apply HARD (hard thresholding) and SCAD (smoothly clipped absolute deviation) symmetric penalty functions with singularities at the origin, and bounded by a constant to reduce bias. In our simulation study and real data analysis, the new method is compared with an existing variable selection method using L_1 penalty that exhibits competitive performance in prediction and variable selection. Instead of using only one type of penalty function, we use two or three penalty functions simultaneously and take advantages of various types of penalty functions together to select relevant predictors and estimation to improve the overall performance of model fitting.

Keywords: Group penalty, multivariate linear model, nonconcave penalty, reduced-rank regression, variable selection.

1. Introduction

Multivariate linear regression is an extension of multiple linear regression to model multiple responses; consequently, it is natural to assume that predictors as well as multiple responses are interrelated. Reduced-rank regression (RRR; Izenman, 1975) is an enhancement of the classical multivariate linear regression model that brings true multivariate features into the model by imposing a rank constraint on regression coefficients and recently received attention in various areas such as statistics, econometrics, and genetics.

In practice, a large number of predictors are introduced in the initial stage of regression modeling and one of the most difficult aspects in this case is the interpretation of linear combination of predictors normally with non-zero coefficient estimates. One possible solution to this problem would be using stepwise deletion and subset selection; however, the stepwise deletion and subset selection tend to ignore stochastic errors inherited in the stages of variable selections and suffer from several drawbacks such as a lack of stability as analyzed by Breiman (1995).

Another possible solution would be introducing penalty function as in usual regression analysis. The rank constraint implies that the coefficient matrix can be expressed as the product of two lower rank matrices. The variable selection in this case is achieved by adding a penalty to the least squares fitting criterion for sparse solutions of the reduced-rank coefficient matrix. Chen and Huang (2012) proposed a group-lasso type penalty that treats each row of the regression coefficient as a group and developed two numerical algorithms for computation and tuning parameter selection methods. A

[1] Corresponding author: Department of Statistics, Sungkyunkwan University, 25-2 Sungkyunkwan-Ro, Jongno-Gu, Seoul 110-745, Korea. E-mail: cspark@skku.edu

Published 31 January 2015 / journal homepage: http://csam.or.kr

unified estimation strategy that combines regression-type formulation of reduced-rank regression and grouped lasso penalties was effective but fails to deliver a robust estimator since it adopts L_1 penalties.

Fan and Li (2001) proposed a variable selection method using penalized likelihood functions and argued that it is possible to retain good features of both subset selection and ridge regression using a unified approach via penalized least squares. They further showed with proper choice of regularization parameters that the proposed estimators perform as well as the oracle procedure in variable selection. This penalized least squared idea can be extended to reduced-rank regression because it belongs to likelihood-based models.

In Section 2, we discuss adding nonconcave penalty functions like HARD (Antoniadis, 1997) and SCAD (Fan and Li, 2001) to reduced-rank regression setup. A unified iterative algorithm extending variational method of Chen and Huang (2012) to solve penalized reduced-rank regression is introduced in Section 3. Section 4 includes numerical comparisons through simulation studies. Results of analyzing well known Yeast data with several penalty functions are provided in Section 5. Some discussion is given in Section 6.

2. Nonconcave Penalized Reduced-Rank Regression

2.1. Reduced-Rank regression model (Reinsel and Velu, 1998)

Reduced-rank regression is a multivariate linear regression method (with more than one response variables and a set of predictors), where the estimated matrix of regression coefficients is of reduced rank. Suppose we have q response variables and p predictors and assuming a linear relationship between each response variable and the predictors. Further assume that we have n multivariate observations of the response and predictor variables, organized in matrices $\mathbf{Y}$ and $\mathbf{X}$ of dimensions $n \times q$ and $n \times p$. We can write the model in matrix notation as

$$\mathbf{Y} = \mathbf{XC} + \mathcal{E}, \tag{2.1}$$

where $\mathbf{C}$ is the $p \times q$ matrix of regression coefficients, and $\mathcal{E}$ is the $n \times q$ error matrix. Residual sum of squares for the multiple response case are defined as

$$\mathrm{Q}\left(\hat{\mathbf{C}}\right) = \mathrm{tr}\left[\left(\mathbf{Y} - \mathbf{X}\hat{\mathbf{C}}\right)^T \left(\mathbf{Y} - \mathbf{X}\hat{\mathbf{C}}\right)\right] = \left\|\mathbf{Y} - \mathbf{X}\hat{\mathbf{C}}\right\|^2, \tag{2.2}$$

where $\|\cdot\|$ denotes the Frobenius norm for a matrix. In the method of ordinary least squares (OLS) (2.2) is minimized without any restrictions on $\hat{\mathbf{C}}$. This yields the ordinary full-rank estimate

$$\hat{\mathbf{C}}^{\mathrm{OLS}} = \mathbf{S}_{xx}^{-1}\mathbf{S}_{xy}, \tag{2.3}$$

where $\mathbf{S}_{xx} = (1/n)\mathbf{X}^T\mathbf{X}$, $\mathbf{S}_{xy} = (1/n)\mathbf{X}^T\mathbf{Y}$ and the rank of $\hat{\mathbf{C}}$ is $m = \min(p, q)$ are of full rank.

Reduced-rank regression utilizes interrelationships between response variables by imposing a constraint on the rank of $\mathbf{C}$. Now, let us assume

$$\mathrm{rank}\,(\mathbf{C}) = r \leq m. \tag{2.4}$$

Equations (2.1) and (2.4) together define the reduced-rank regression model. When $\mathbf{C}$ has reduced-rank r, there exist two full-rank matrices $\mathbf{A}$ and $\mathbf{B}$ such that

$$\mathbf{C} = \mathbf{BA}^T,$$

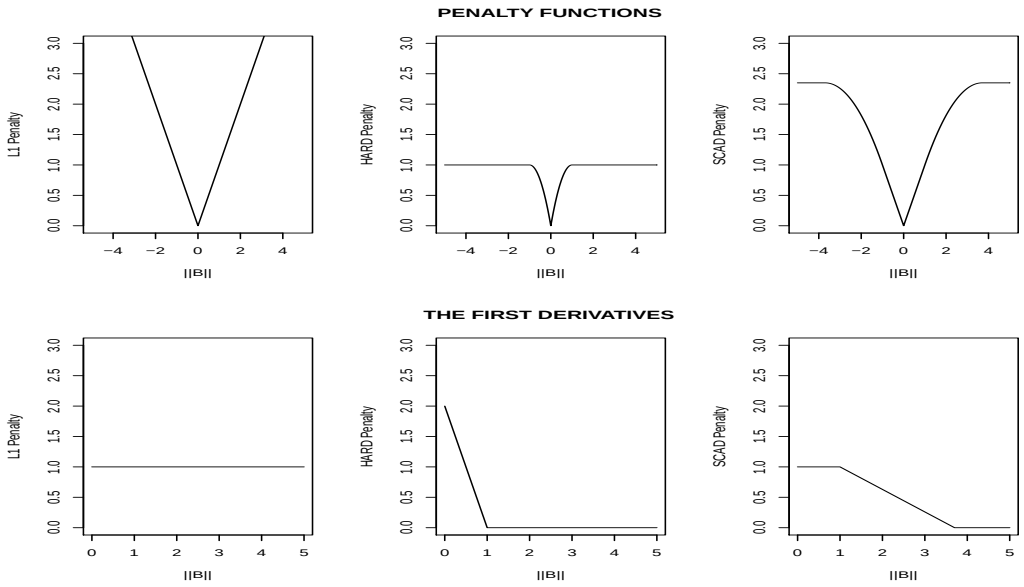

Figure 1: *Plot of penalty functions and the first derivatives*

where $\mathbf{B}$ has dimension $p \times r$, and $\mathbf{A}$ is of dimension $q \times r$. The model (2.1) can be rewritten as

$$\mathbf{Y} = \mathbf{XBA}^T + \mathcal{E}, \tag{2.5}$$

which makes it possible to model q response variables with $r (r \leq q)$ linear combinations of the predictor variables.

Estimates of $\mathbf{A}$ and $\mathbf{B}$ can be achieved by solving the optimization problem

$$\min_{\mathbf{A},\mathbf{B}} \|\mathbf{Y} - \mathbf{XBA}\|^2 \tag{2.6}$$

with $\text{rank}(\mathbf{A}) = \text{rank}(\mathbf{B}) = r$. A set of solution for matrices $\mathbf{A}$ and $\mathbf{B}$ is provided by

$$\hat{\mathbf{A}} = \mathbf{V}, \qquad \hat{\mathbf{B}} = \mathbf{S}_{xx}^{-1}\mathbf{S}_{xy}\mathbf{V},$$

where $\mathbf{V} = [\mathbf{v}_1, \ldots, \mathbf{v}_r]$ and $\mathbf{v}_i$ is the eigenvectors of $\mathbf{S}_{yx}\mathbf{S}_{xx}^{-1}\mathbf{S}_{xy}$ corresponding to the i^{th} largest eigenvalue γ_i with $\mathbf{S}_{yx} = (1/n)\mathbf{Y}^T\mathbf{X}$.

2.2. Penalized reduced-rank regression

We consider applying nonconcave penalty function idea to reduced-rank regression in this section. Using penalized least squares (or likelihood idea) in variable selection is better because they retain good features of both subset selection and ridge regression (Fan and Li, 2001). By including penalty functions in reduced-rank regression problems, it is possible to force estimates related to unnecessary predictors to zero safely so making interpretations easier.

Chen and Huang (2012) argued that using grouped lasso penalty (Yuan and Lin, 2006) instead of applying element-wise lasso penalty (Tibshirani, 1996) guarantees the identifiability of variable

selection. Thus, the group lasso penalty encourages row-wise sparsity on the **B** matrix. Applying row-wise group penalty function to the optimization problem (2.6) becomes

$$\min_{\mathbf{A},\mathbf{B}} \left\|\mathbf{Y} - \mathbf{X}\mathbf{B}\mathbf{A}^T\right\|^2 + \sum_{i=1}^{p} p_{\lambda_i}\left(\left\|\mathbf{B}^i\right\|\right), \quad \text{such that } \mathbf{A}^T\mathbf{A} = \mathbf{I}, \tag{2.7}$$

where $\mathbf{B}^i$ is the i^{th} row of **B** and $\lambda_i > 0$ are penalty parameters. Instead of using $\lambda p(\|\cdot\|)$ we uses $p_\lambda(\|\cdot\|)$ so $p(\|\cdot\|)$ may be allowed to depend on λ and we assume that the penalty functions for all coefficients are the same and delete subscript i of tuning parameter λ_i. Fan and Li (2001) argued that unbiasedness, sparsity, and continuity are three properties possessed by a good penalty function, and suggested SCAD penalty function as best for regression problems. Several well-known penalty functions including SCAD penalty function are as follows.

1. L_p: $p_\lambda(\|\mathbf{B}^i\|) = \lambda\|\mathbf{B}^i\|^p$ and it becomes LASSO with $p = 1$ for least squares case.

2. Hard Thresholding (HARD) Penalty:

$$p_\lambda\left(\|\mathbf{B}\|^i\right) = \lambda^2 - \left(\|\mathbf{B}\|^i - \lambda\right)^2 I\left(\left\|\mathbf{B}^i\right\| < \lambda\right).$$

3. Smoothly Clipped Absolute Deviation (SCAD) Penalty:

$$p_\lambda\left(\left\|\mathbf{B}^i\right\|\right) = \begin{cases} \lambda\left\|\mathbf{B}^i\right\|, & \text{if } \left\|\mathbf{B}^i\right\| < \lambda, \\ -\dfrac{\left\|\mathbf{B}^i\right\|^2 - 2a\left\|\mathbf{B}^i\right\| + \lambda^2}{2(a-1)}, & \text{if } \lambda \le \left\|\mathbf{B}^i\right\| < a\lambda, \\ \dfrac{(a+1)\lambda^2}{2}, & \text{if } \left\|\mathbf{B}^i\right\| \ge a\lambda. \end{cases}$$

Fan and Li (2001) noted that consistency and oracle properties cannot be satisfied simultaneously for the L_1 penalty. The hard thresholding (HARD) penalty function is unbiased and is sparse (but not continuous). SCAD behaves somewhat between L_1 and HARD and need two dimensional burdensome Generalized Cross-Validation (GCV) or usual Cross-Validation (CV) to find optimal values for two parameters, a and λ. Figure 1 provides shapes of penalty functions and 1st derivatives for L_1, HARD, and SCAD penalties with $\lambda = 1$ and $a = 3.7$ for the SCAD penalty function.

3. Numerical Algorithm

In this section, we presents a unified algorithm to solve the optimization problem (2.7) and also discussed methods to select appropriate tuning parameters.

3.1. Iterative optimization

We can solve the optimization problem (2.7) iteratively by optimizing over **A** for fixed **B** and then optimizing over **B** for fixed **A** as follows.

For fixed **B**, (2.7) becomes

$$\min_{\mathbf{A}} \left\|\mathbf{Y} - \mathbf{X}\mathbf{B}\mathbf{A}^T\right\|^2, \quad \text{such that } \mathbf{A}^T\mathbf{A} = \mathbf{I}. \tag{3.1}$$

This optimization is known as an orthogonal Procrustes problem (Gower and Dijksterhuis, 2004). The solution is $\hat{\mathbf{A}} = \mathbf{U}\mathbf{V}^T$, where $\mathbf{U}$ and $\mathbf{V}$ are obtained from singular value decomposition $\mathbf{Y}^T\mathbf{X}\mathbf{B} = \mathbf{U}\mathbf{D}\mathbf{V}^T$ with $\mathbf{U}_{q\times r}$ and $\mathbf{V}_{r\times r}$.

The optimization over $\mathbf{B}$ for fixed $\mathbf{A}$ reduces to

$$\min_{\mathbf{B}} \left\|\mathbf{Y}\mathbf{A} - \mathbf{X}\mathbf{B}^T\right\|^2 + \sum_{i=1}^{p} p_\lambda\left(\left\|\mathbf{B}^i\right\|\right) \tag{3.2}$$

since there exists a matrix $\mathbf{A}^\perp$ with orthogonal columns such that $(\mathbf{A}, \mathbf{A}^\perp)$ is an orthogonal matrix. We may apply local quadratic approximation approach of Fan and Li (Section 3.3, 2001) and subgradient method (Friedman *et al.*, 2007) to solve optimization problem of (3.2) with respect to $\mathbf{B}^l$, the i^{th} row of $\mathbf{B}$. When $\|\mathbf{B}^l\| \neq 0$, the first derivative of penalty function can be locally approximated by a quadratic function as

$$\left[p_\lambda\left(\left\|\mathbf{B}^l\right\|\right)\right]' = p'_\lambda\left(\left\|\mathbf{B}^l\right\|\right)\operatorname{sgn}\left(\left\|\mathbf{B}^l\right\|\right) \approx \left\{p'_\lambda\left(\left\|\mathbf{B}^l\right\|\right) / \left\|\mathbf{B}^l\right\|\right\}\mathbf{B}^l \tag{3.3}$$

and (3.2) also can be locally approximated by

$$\mathrm{Q}(\mathbf{B}_0) + \frac{\partial \mathrm{Q}(\mathbf{B}_0)}{\partial \mathbf{B}^l}^T\left(\mathbf{B}^l - \mathbf{B}_0^l\right) + \frac{1}{2}\left(\mathbf{B}^l - \mathbf{B}_0^l\right)^T \frac{\partial^2 \mathrm{Q}(\mathbf{B}_0)}{\partial \mathbf{B}^l \partial \mathbf{B}^{lT}}\left(\mathbf{B}^l - \mathbf{B}_0^l\right) + \frac{1}{2}\mathbf{B}^{lT}\Sigma_\lambda(\mathbf{B}_0)\mathbf{B}^l, \tag{3.4}$$

where we are given an initial value $\mathbf{B}_0$ that is close to the minimizer of (3.2) with $\mathrm{Q}(\mathbf{B}_0) = \|\mathbf{Y}\mathbf{A} - \mathbf{X}\mathbf{B}_0^T\|^2$, and $\Sigma_\lambda(\mathbf{B}_0) = \operatorname{diag}(p'_\lambda(\|\mathbf{B}_0^1\|/\|\mathbf{B}_0^1\|), \ldots, p'_\lambda(\|\mathbf{B}_0^p\|/\|\mathbf{B}_0^p\|))$.

Taking the first-order condition for $\mathbf{B}^i$ $(i = 1, \ldots, p)$ for optimization problem (3.2), and collecting all these conditions result as

$$-\mathbf{X}^T(\mathbf{Y}\mathbf{A} - \mathbf{X}\mathbf{B}) + \frac{1}{2}\Sigma_\lambda(\mathbf{B}_0)\mathbf{B} = 0.$$

Solving for $\mathbf{B}$ yields

$$\mathbf{B} = \left\{\mathbf{X}^T\mathbf{X} + \frac{1}{2}\Sigma_\lambda(\mathbf{B}_0)\right\}^{-1}\mathbf{X}^T\mathbf{Y}\mathbf{A}. \tag{3.5}$$

The solution can be found by iteratively computing the above ridge regression (3.5). One iterates between (3.4) and $\hat{\mathbf{A}}$ until convergence and needs to set a threshold for $\|\mathbf{B}^i\|$ to get sparse solutions if necessary.

3.2. Parameter tuning

In the simulation studies, we fix the number of rank (r) to true value 3 instead of estimating it from cross-validation for each data set generated to focus on the effect of applying different penalty functions. We also fix the number of rank for the Yeast data set as 4 suggested by Chen and Huang (2012) in their analysis. Penalty parameter λ is selected from 5-fold cross-validation (CV) for each data set simulated and Yeast data set. We have one more parameter a in the SCAD penalty function in addition to usual λ. Fan and Li (2001) argued that the choice of $a = 3.7$ is very reasonable for wide range of data sets, so we also fixed a as 3.7 throughout all simulated and real data analysis.

4. Small Simulation

In this section, we use simulated data to illustrate and compare the performance of L_1, HARD, and SCAD penalty functions.

4.1. Simulation setup

The model used in simulation study is (2.5), $\mathbf{Y} = \mathbf{XBA}^T + \mathcal{E}$. The $n \times p$ predictor matrix $\mathbf{X}$ was generated from multivariate normal distribution $\mathcal{N}(0, \mathbf{\Sigma}_x)$, where $\mathbf{\Sigma}_x$ having diagonal elements 1 and off-diagonal elements ρ_x. We generated first p_0 rows of the $p \times r$ matrix $\mathbf{B}$ from $\mathcal{N}(0, 1)$ and rest $p - p_0$ rows were set to be zero. The elements of $q \times r$ matrix $\mathbf{A}$ were generated from $\mathcal{N}(0, 1)$ and elements of the $n \times q$ random error matrix $\mathcal{E}$ were generated from $\mathcal{N}(0, \sigma^2\mathbf{\Sigma}_e)$ with $\mathbf{\Sigma}_e$ having diagonal elements 1 and off-diagonal elements ρ_e. We set $\sigma^2 = 0.01$ to make the signal-to-noise ratio, $\text{trace}(\mathbf{C}^T\mathbf{\Sigma}_x\mathbf{C})/\text{trace}(\mathcal{E}) \approx 1$.

After centering and standardizing $\mathbf{Y}$, we applied L_1 penalty function as well as HARD, and SCAD penalty function to each data set generated. We compared accuracy of three penalty functions in terms of the angle between true and estimated row vector of coefficient matrix $\mathbf{B}$,

$$\theta_i = \text{acos}\left(\frac{(\mathbf{B}^i)^T\hat{\mathbf{B}}^i}{\|\mathbf{B}^i\| \cdot \|\hat{\mathbf{B}}^i\|}\right), \quad i = 1, \ldots, p. \tag{4.1}$$

We also compared the three penalty functions in terms of the number of correct non-zeros (C) and incorrect zeros (IC) for each data set. Correct non-zeros C is simply the average presents restricted only to the true non-zero rows of coefficients. The incorrect zeros IC depicts the average of coefficient rows erroneously set to 0. All measures are reported as the average and standard deviation (SD) in parenthesis of 100 repetitions for each case.

We considered three sets of simulations as follows. In the first set of simulations, we fix $r = 3$, $N = 100$, $p = 30$, $p_0 = 10$, $q = 10$ and varying ρ_x and ρ_e. Moderate amount of correlation among predictors is introduced in the second cases (Case 2, 6 and 10) of each set, and stronger correlation of 0.9 for the third cases (Case 3, 7, and 11). And moderate amount of correlation among error terms or responses is introduced in the fourth cases (Case 4, 8, and 12).

Case 1: $N = 100$, $p = 30$, $p_0 = 10$, $q = 10$, $\rho_x = 0$, $\rho_e = 0$ (Table 1, and Figure 2)

Case 2: $N = 100$, $p = 30$, $p_0 = 10$, $q = 10$, $\rho_x = 0.5$, $\rho_e = 0$ (Table 2, and Figure 3)

Case 3: $N = 100$, $p = 30$, $p_0 = 10$, $q = 10$, $\rho_x = 0.9$, $\rho_e = 0$ (Table 3, and Figure 4)

Case 4: $N = 100$, $p = 30$, $p_0 = 10$, $q = 10$, $\rho_x = 0$, $\rho_e = 0.5$ (Table 4, and Figure 5)

The second set is similar to the first set except the number of observation (N) reduces to 50.

Case 5: $N = 50$, $p = 30$, $p_0 = 10$, $q = 10$, $\rho_x = 0$, $\rho_e = 0$ (Table 5, and Figure 6)

Case 6: $N = 50$, $p = 30$, $p_0 = 10$, $q = 10$, $\rho_x = 0.5$, $\rho_e = 0$ (Table 6, and Figure 7)

Case 7: $N = 50$, $p = 30$, $p_0 = 10$, $q = 10$, $\rho_x = 0.9$, $\rho_e = 0$ (Table 7, and Figure 8)

Case 8: $N = 50$, $p = 30$, $p_0 = 10$, $q = 10$, $\rho_x = 0$, $\rho_e = 0.5$ (Table 8, and Figure 9)

The third set is a higher dimensional case with $n << p$. We consider the case with $r = 3$, $N = 50$, $p = 70$, $p_0 = 20$, $q = 10$ and varying ρ_x and ρ_e similar to previous two setups.

Case 9: $N = 50$, $p = 70$, $p_0 = 20$, $q = 10$, $\rho_x = 0$, $\rho_e = 0$ (Table 9, and Figure 10)

Case 10: $N = 50$, $p = 70$, $p_0 = 20$, $q = 10$, $\rho_x = 0.5$, $\rho_e = 0$ (Table 10, and Figure 11)

Case 11: $N = 50$, $p = 70$, $p_0 = 20$, $q = 10$, $\rho_x = 0.9$, $\rho_e = 0$ (Table 11, and Figure 12)

Case 12: $N = 50$, $p = 70$, $p_0 = 20$, $q = 10$, $\rho_x = 0$, $\rho_e = 0.5$ (Table 12, and Figure 13)

Table 1: Mean of angles between true and estimated $\mathbf{B}^i$, C (No. of correct selection), and IC (No. of incorrect selection) (Case 1: $N = 100$, $p = 30$, $p_0 = 10$, $q = 10$, $\rho_x = 0$, $\rho_e = 0$)

Penalty	Mean of row Angles	C	IC
L_1	0.136 (0.062)	30.00 (0.000)	2.22 (6.777)
HARD	0.149 (0.052)	29.97 (0.300)	0.86 (2.179)
SCAD	0.144 (0.056)	29.97 (0.300)	6.68 (6.716)

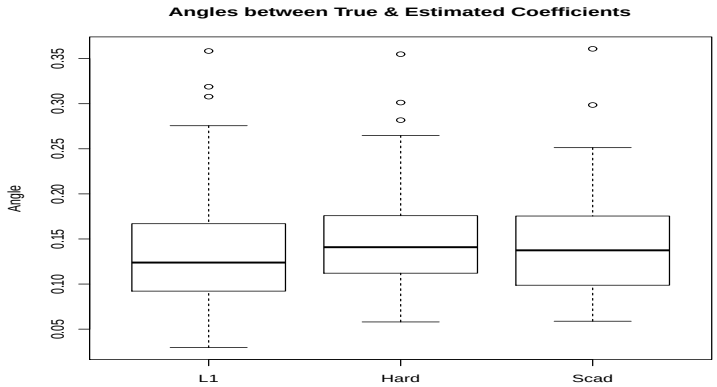

Figure 2: *Boxplot of mean angles between true and estimated* $\mathbf{B}^i$*, C (No. of correct selection), and IC (No. of incorrect selection): Case 1*

Table 2: Mean of angles between true and estimated $\mathbf{B}^i$, C (No. of correct selection), and IC (No. of incorrect selection) (Case 2: $N = 100$, $p = 30$, $p_0 = 10$, $q = 10$, $\rho_x = 0.5$, $\rho_e = 0$)

Penalty	Mean of row Angles	C	IC
L_1	0.167 (0.067)	30.00 (0.000)	0.24 (1.393)
HARD	0.201 (0.067)	29.64 (0.980)	0.60 (1,858)
SCAD	0.168 (0.074)	29.99 (0.100)	4.66 (4.835)

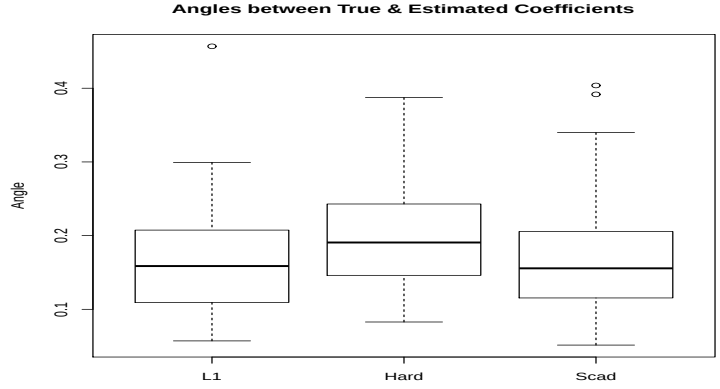

Figure 3: *Boxplot of mean angles between true and estimated* $\mathbf{B}^i$*, C (No. of correct selection), and IC (No. of incorrect selection): Case 2*

Table 3: Mean of angles between true and estimated $\mathbf{B}^i$, C (No. of correct selection), and IC (No. of incorrect selection) (Case 3: $N = 100$, $p = 30$, $p_0 = 10$, $q = 10$, $\rho_x = 0.9$, $\rho_e = 0$)

Penalty	Mean of row Angles	C	IC
L_1	0.344 (0.105)	28.95 (1.828)	0.21 (0.880)
HARD	0.449 (0.110)	24.67 (3.843)	0.15 (0.989)
SCAD	0.329 (0.114)	28.80 (1.907)	0.78 (2,321)

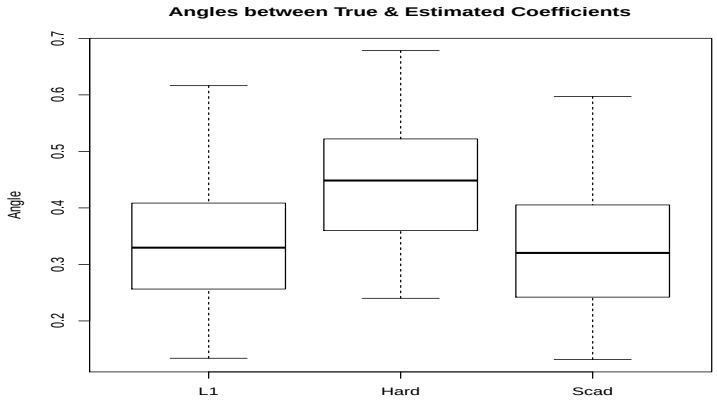

Figure 4: *Boxplot of mean angles between true and estimated* $\mathbf{B}^i$*, C (No. of correct selection), and IC (No. of incorrect selection): Case 3*

Table 4: Mean of angles between true and estimated $\mathbf{B}^i$, C (No. of correct selection), and IC (No. of incorrect selection) (Case 4: $N = 100$, $p = 30$, $p_0 = 10$, $q = 10$, $\rho_x = 0$, $\rho_e = 0.5$)

Penalty	Mean of row Angles	C	IC
L_1	0.146 (0.057)	30,00 (0.000)	1.97 (5.987)
HARD	0.139 (0.055)	29.90 (0.522)	0.45 (1.374)
SCAD	0.128 (0.050)	29.99 (0.100)	6.42 (6.671)

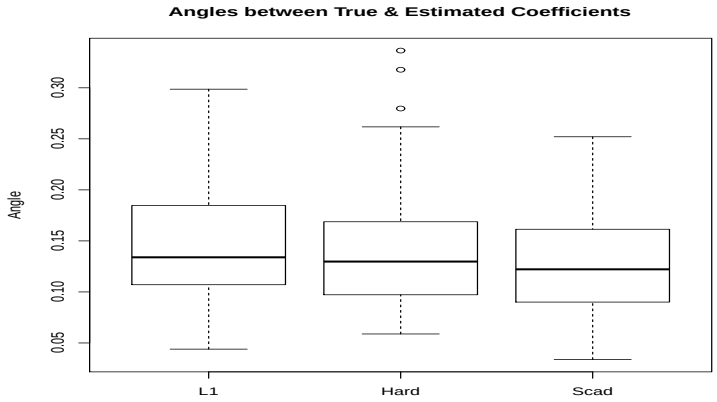

Figure 5: *Boxplot of mean angles between true and estimated* $\mathbf{B}^i$*, C (No. of correct selection), and IC (No. of incorrect selection): Case 4*

Table 5: Mean of angles between true and estimated $\mathbf{B}^i$, C (No. of correct selection), and IC (No. of incorrect selection) (Case 5: $N = 50$, $p = 30$, $p_0 = 10$, $q = 10$, $\rho_x = 0$, $\rho_e = 0$)

Penalty	Mean of row Angles	C	IC
L_1	0.175 (0.057)	29.94 (0.422)	3.53 (6.940)
HARD	0.214 (0.052)	29.19 (1.468)	1.60 (2.625)
SCAD	0.176 (0.051)	29.94 (0.422)	9.05 (6.570)

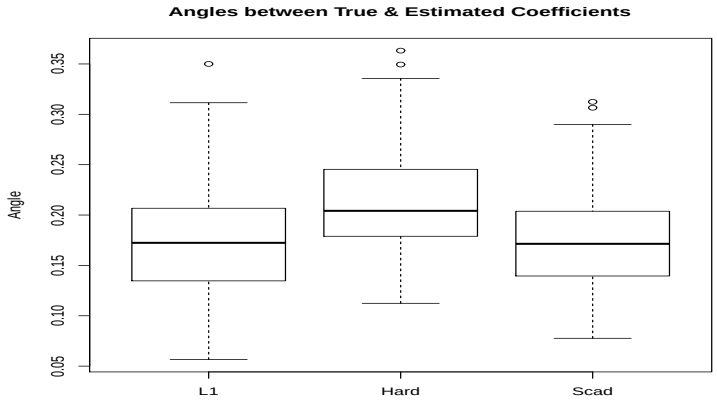

Figure 6: *Boxplot of mean angles between true and estimated* $\mathbf{B}^i$*, C (No. of correct selection), and IC (No. of incorrect selection): Case 5*

Table 6: Mean of angles between true and estimated $\mathbf{B}^i$, C (No. of correct selection), and IC (No. of incorrect selection) (Case 6: $N = 50$, $p = 30$, $p_0 = 10$, $q = 10$, $\rho_x = 0.5$, $\rho_e = 0$)

Penalty	Mean of row Angles	C	IC
L_1	0.237 (0.071)	29.69 (0.907)	2.52 (4.279)
HARD	0.307 (0.064)	28.25 (2.017)	1.49 (2.002)
SCAD	0.219 (0.068)	29.64 (0.980)	6.30 (5.846)

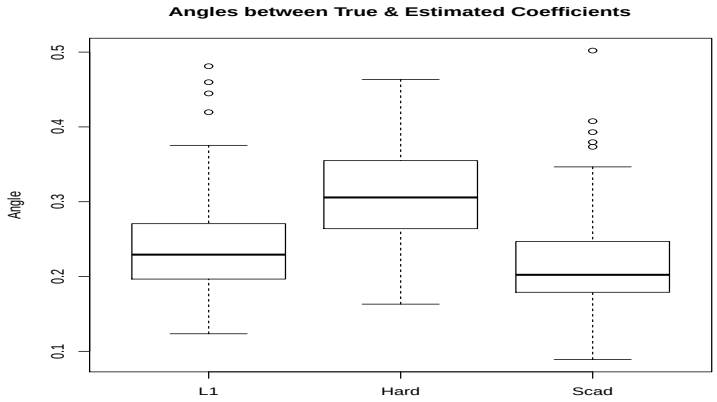

Figure 7: *Boxplot of mean angles between true and estimated* $\mathbf{B}^i$*, C (No. of correct selection), and IC (No. of incorrect selection): Case 6*

Table 7: Mean of angles between true and estimated $\mathbf{B}^i$, C (No. of correct selection), and IC (No. of incorrect selection) (Case 6: $N = 50$, $p = 30$, $p_0 = 10$, $q = 10$, $\rho_x = 0.9$, $\rho_e = 0$)

Penalty	Mean of row Angles	C	IC
L_1	0.541 (0.113)	25.39 (3.035)	1.95 (3.701)
HARD	0.737 (0.136)	17.34 (3.542)	1.26 (2.144)
SCAD	0.524 (0.118)	24.40 (3.300)	2.22 (2.801)

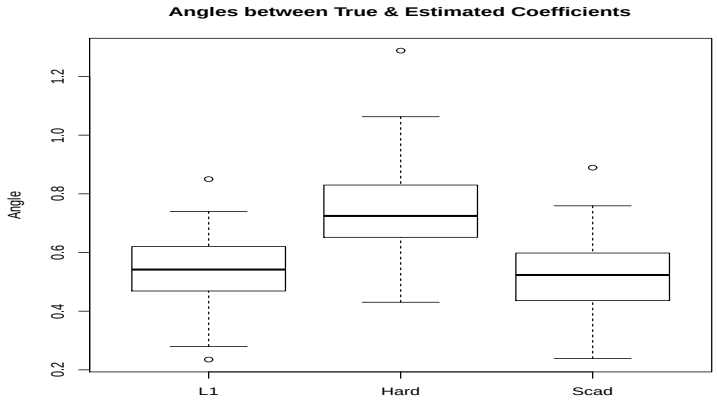

Figure 8: *Boxplot of mean angles between true and estimated* $\mathbf{B}^i$*, C (No. of correct selection), and IC (No. of incorrect selection): Case 7*

Table 8: Mean of angles between true and estimated $\mathbf{B}^i$, C (No. of correct selection), and IC (No. of incorrect selection) (Case 8: $N = 50$, $p = 30$, $p_0 = 10$, $q = 10$, $\rho_x = 0$, $\rho_e = 0.5$)

Penalty	Mean of row Angles	C	IC
L_1	0.180 (0.055)	29.88 (0.591)	3.06 (6.310)
HARD	0.203 (0.048)	29.64 (1.069)	1.37 (2.533)
SCAD	0.170 (0.053)	29.86 (0.603)	7.65 (7.006)

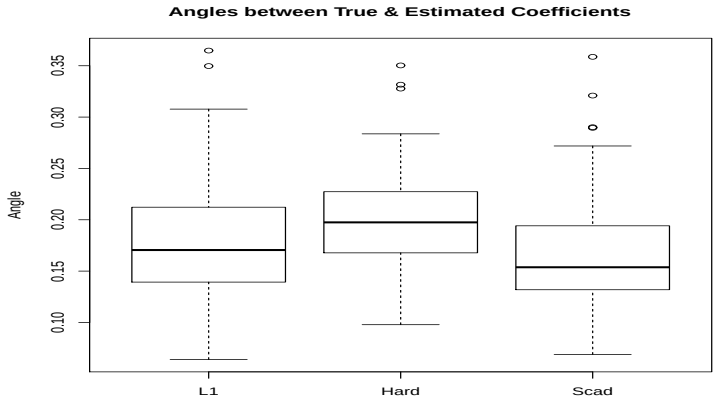

Figure 9: *Boxplot of mean angles between true and estimated* $\mathbf{B}^i$*, C (No. of correct selection), and IC (No. of incorrect selection): Case 8*

Table 9: Mean of angles between true and estimated $\mathbf{B}^i$, C (No. of correct selection), and IC (No. of incorrect selection) (Case 9: $N = 50$, $p = 70$, $p_0 = 20$, $q = 10$, $\rho_x = 0$, $\rho_e = 0$)

Penalty	Mean of row Angles	C	IC
L_1	0.275 (0.059)	58.61 (2.188)	26.19 (16.765)
HARD	0.380 (0.060)	54.69 (3.317)	16.93 (7.966)
SCAD	0.283 (0.051)	58.34 (2.171)	36.09 (13.238)

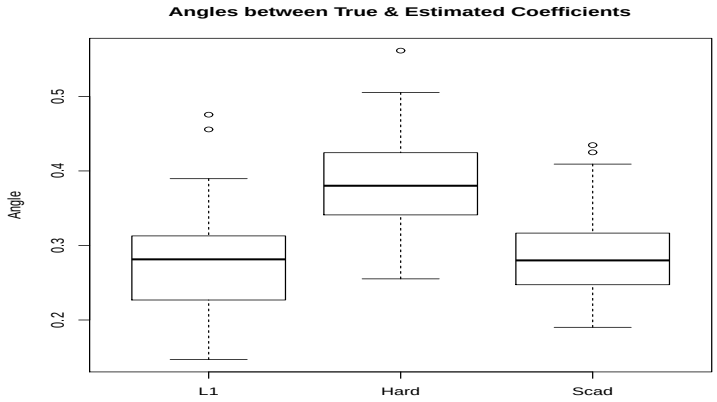

Figure 10: *Boxplot of mean angles between true and estimated* $\mathbf{B}^i$*, C (No. of correct selection), and IC (No. of incorrect selection): Case 9*

Table 10: Mean of angles between true and estimated $\mathbf{B}^i$, C (No. of correct selection), and IC (No. of incorrect selection) (Case 10: $N = 50$, $p = 70$, $p_0 = 20$, $q = 10$, $\rho_x = 0.5$, $\rho_e = 0$)

Penalty	Mean of row Angles	C	IC
L_1	0.380 (0.074)	56.60 (3.094)	18.06 (10.756)
HARD	0.545 (0.086)	47.73 (4.481)	11.18 (6.743)
SCAD	0.384 (0.082)	56.19 (3.071)	25.84 (10.445)

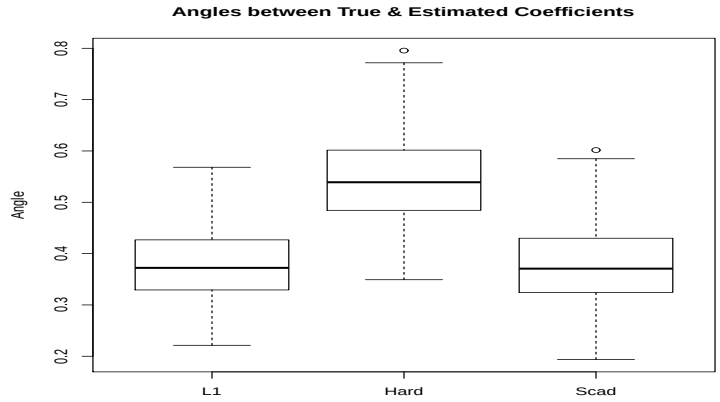

Figure 11: *Boxplot of mean angles between true and estimated* $\mathbf{B}^i$*, C (No. of correct selection), and IC (No. of incorrect selection): Case 10*

Table 11: Mean of angles between true and estimated $\mathbf{B}^i$, C (No. of correct selection), and IC (No. of incorrect selection) (Case 11: $N = 50$, $p = 70$, $p_0 = 20$, $q = 10$, $\rho_x = 0.9$, $\rho_e = 0$)

Penalty	Mean of row Angles	C	IC
L_1	0.770 (0.116)	38.01 (6.091)	7.16 (5.361)
HARD	1.009 (0.116)	22.33 (5.067)	4.63 (4.691)
SCAD	0.758 (0.126)	36.77 (5.894)	9.85 (6.611)

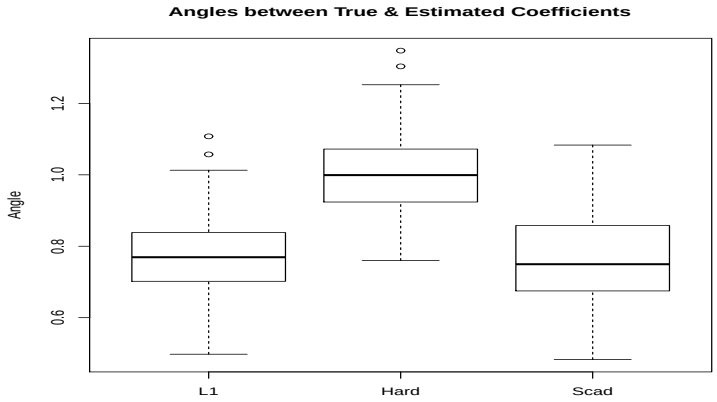

Figure 12: *Boxplot of mean angles between true and estimated* $\mathbf{B}^i$*, C (No. of correct selection), and IC (No. of incorrect selection): Case 11*

Table 12: Mean of angles between true and estimated $\mathbf{B}^i$, C (No. of correct selection), and IC (No. of incorrect selection) (Case 12: $N = 50$, $p = 70$, $p_0 = 20$, $q = 10$, $\rho_x = 0$, $\rho_e = 0.5$)

Penalty	Mean of row Angles	C	IC
L_1	0.283 (0.054)	58.80 (1.809)	25.06 (16.227)
HARD	0.395 (0.068)	55.22 (3.230)	15.32 (7.542)
SCAD	0.290 (0.063)	58.53 (2.100)	35.06 (12.520)

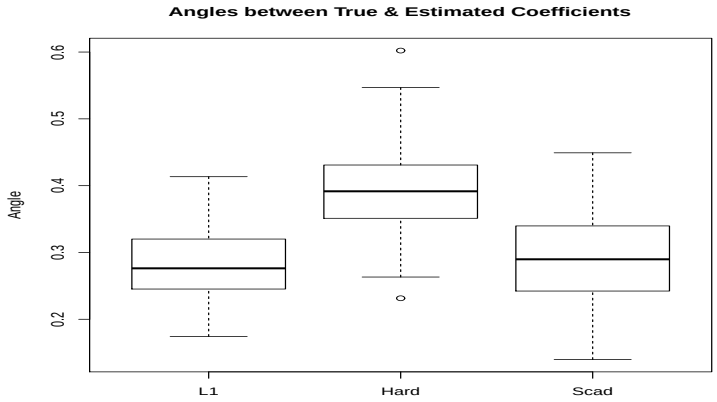

Figure 13: *Boxplot of mean angles between true and estimated* $\mathbf{B}^i$*, C (No. of correct selection), and IC (No. of incorrect selection): Case 12*

Table 13: Mean and SE of prediction errors for Yeast Dataset

Penalty	Mean	SE
L_1	0.187	0.00141
HARD	0.192	0.00170
SCAD	0.188	0.00135

4.2. Simulation results

We applied three penalty functions to each of twelve cases. For each case, 5-fold cross-validation is used to select optimal tuning parameter λ. Table 1 through Table 12 reports average of row angles between true and estimated row vector of coefficients, the number of correctly selected rows, and the number of incorrectly selected rows erroneously set to 0 and their standard deviations from 100 test samples. Figure 2 to Figure 13 includes boxplots for average angles of three penalties for each case.

Regardless of differences in parameter settings, behavior of all three measures between cases are similar. Overall, angles between true and estimated coefficients of L_1 and SCAD penalty functions are smaller than HARD penalty, but SCAD penalty look better with smaller SD in more cases. Even though HARD penalty looks more biased than other two penalties, it seems to be more efficient in that it avoids erroneously forcing coefficients of relevant predictors to 0.

All three penalties show almost perfect performance to select irrelevant predictors in all cases with $n > p$. In the higher dimensional case with $n << p$, measure C (number of correct selection) deteriorates when there is strong correlation between predictors. All three penalties fail to select true non-zero coefficients properly.

Angles between true and estimated coefficients increases as correlation between predictors becomes stronger. Unless there exists strong correlation between predictors, it has little effect on all three measures regardless of penalty type. However, it is interesting to note that SCAD penalty is always a winner when there exists a strong correlation between predictors.

5. Real Data Example

Transcription factors (TF) play an important role to regulate RNA transcript levels of yeast genes within the eukaryotic cell cycle. We use a data set analyzed by Chun and Keles (2010), that includes 524 genes for a total of 106 TFs and RNA levels measured every 7 min for 119 min with a total of 18 time points. Data set consists of expression levels for 542 cell cycle related genes at 18 time points using the chip-chip data of 106 TFs and identifies cell cycle related TFs as well as to infer TF activities. Then $\mathbf{Y}$ becomes 524×18 matrix and $\mathbf{X}$ is a 524×106 matrix.

We set the number of rank, $r = 4$ as recommended by Chen and Huang (2012) and 5-fold cross-validation was used for selecting tuning parameter λ. Additional tuning parameter a in SCAD is set to be 3.7. Prediction accuracy was calculated 100 times as follows. We split the data set into 50% training and 50% test set. The training selects the best penalty parameters and model used in the corresponding model to predict test data. Table 13 reports the means and standard errors of squared prediction errors from random 100 splits.

All three penalties perform similar to angles in the simulation studies from the previous section in terms of prediction errors. The larger prediction error of HARD seems to be due to the relatively larger biases than the other two penalties. In terms of standard error, HARD penalty is the worst and SCAD is the best among the three.

6. Concluding Remarks

We propose a group-type nonconcave penalties for a reduced-rank regression model and compared their performances with group lasso or L_1 penalty function. We show that the HARD penalty function is efficient to avoid incorrect selection and that the SCAD penalty function is comparable with L_1 penalty in terms of angles between true and estimated coefficients, and is robust with respect to correlation between predictors. However, SCAD penalty fails more often in selecting irrelevant predictors overall.

We may use two or three penalty functions simultaneously instead of only one type of penalty function. We can then take advantage of various features of penalty function together to select relevant predictors and estimation to improve overall modeling performance. Results from applying the HARD penalty can identify a subset of relevant predictors; consequently, results from SCAD or L_1 help obtain accurate estimation and prediction.

It is also possible to extend above idea to RVGLM (Reduced-rank Vector Generalized Linear Models; Yee and Hastie, 2003) simply by replacing the optimization part of (2.7) with a negative of appropriate likelihood function defendant on type of response variable.

References

Antoniadis, A. (1997). Wavelets in statistics: A review (with discussion), *Journal of the Italian Statistical Association*, **6**, 97–144.

Breiman, L. (1995). Better subset regression using the nonnegative garrote, *Technometrics*, **37**, 373–384.

Chen, L. and Huang, J. Z. (2012). Sparse reduced-rand regression for simultaneous dimension reduction and variable selection, *Journal of the American Statistical Association*, **107**, 1533–1545.

Chun, H. and Keles, S. (2010). Sparse partial least squares regression for simultaneous dimension reduction and variable selection, *Journal of the Royal Statistical Society, Series B*, **72**, 3–25.

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.

Friedman, J., Hastie, T., Hofling, H. and Tibshirani, R. (2007). Pathwise coordinate optimization, *The Annals of Applied Statistics*, **1**, 302–332.

Gower, J. C. and Dijksterhuis, G. B. (2004). *Procrustes Problems*, Oxford University Press, New York.

Izenman, A. J. (1975). Reduced-rank regression for the multivariate linear model, *Journal of Multivariate Analysis*, **5**, 248–264.

Reinsel, G. C. and Velu, R. P. (1998). *Multivariate Reduced-Rank Regression: Theory and Applications*, Springer, New York.

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

Yee, T. W. and Hastie, T. J. (2003). Reduced-rank vector generailized linear models, *Statistical Modeliing*, **3**, 15–41.

Yuan, M. and Lin, Y. (2006). Model selection and estimation in regression with grouped variables, *Journal of the Royal Statistical Society, Series B*, **68**, 49–67.

Received August 30, 2014; Revised December 9, 2014; Accepted December 21, 2014