

Supplementary Material for “L2RM: Low-rank Linear Regression Models for High-dimensional Matrix Responses”

Dehan Kong, Baiguo An, Jingwen Zhang, Hongtu Zhu

The Supplement Material is organized as follows. Section 1 includes the modified algorithm for our regularized low rank estimation procedure when response and covariates are not centered. Section 2 presents some additional simulation results.

1 Estimation procedure without centering

In the main paper, we assume that x_{il} has mean 0 and variance 1 for every $1 \leq l \leq s$ and $\{\mathbf{Y}_i, 1 \leq i \leq n\}$ has mean $\mathbf{0}$. If these assumptions are not satisfied, then one approach is to standardize (center and scale) our covariates \mathbf{x}_i s and center our responses \mathbf{Y}_i s. An alternative approach is to introduce another intercept matrix term \mathbf{B}_0 in our model even without centering. In this case, our model is given by

$$\mathbf{Y}_i = \mathbf{B}_0 + \sum_{l=1}^s x_{il} * \mathbf{B}_l + \mathbf{E}_i. \quad (1)$$

Define $\mathbf{B}_{int} = [\mathbf{B}_0, \mathbf{B}_l, l \in \widehat{\mathcal{M}}] = [\mathbf{B}_0, \mathbf{B}_{l_1}, \dots, \mathbf{B}_{l_{|\widehat{\mathcal{M}}|}}] \in \mathbb{R}^{p \times q(|\widehat{\mathcal{M}}|+1)}$ and $\mathbf{X}_{int} = [\mathbf{1}, \mathbf{X}_{\widehat{\mathcal{M}}}]$. For our estimation procedure, we will calculate the regularized least square estimator of \mathbf{B} by minimizing

$$Q^*(\mathbf{B}_{int}) = \frac{1}{2n} \sum_{i=1}^n \|\mathbf{Y}_i - \mathbf{B}_0 - \sum_{l \in \widehat{\mathcal{M}}} x_{il} * \mathbf{B}_l\|_F^2 + \lambda \sum_{l \in \widehat{\mathcal{M}}} \|\mathbf{B}_l\|_*. \quad (2)$$

Denote $R(\mathbf{B}_{int}) = (2n)^{-1} \sum_{i=1}^n \|\mathbf{Y}_i - \mathbf{B}_0 - \sum_{l \in \widehat{\mathcal{M}}} x_{il} * \mathbf{B}_l\|_F^2$ and $\nabla R(\mathbf{B}_{int})$ denotes the first-order gradient of $R(\mathbf{B}_{int})$ with respect to \mathbf{B}_{int} . Specifically, let \mathbf{B}_{l_d} , $\mathbf{S}_{l_d}^{(t)}$, and $\nabla R(\mathbf{S}^{(t)})_{l_d}$

be the $(dq + 1)$ th to the $(dq + d)$ th columns of the corresponding $p \times q(|\widehat{\mathcal{M}}| + 1)$ matrices \mathbf{B}_{int} , $\mathbf{S}^{(t)}$, and $\nabla R(\mathbf{S}^{(t)})$, respectively, and \mathbf{B}_0 , $\mathbf{S}_0^{(t)}$, and $\nabla R(\mathbf{S}^{(t)})_0$ be the first to the d th columns of the corresponding $p \times q(|\widehat{\mathcal{M}}| + 1)$ matrices \mathbf{B}_{int} , $\mathbf{S}^{(t)}$, and $\nabla R(\mathbf{S}^{(t)})$.

To solve the minimization problem (2), we can still apply the Nesterov gradient method and modify our original algorithm a little bit. Our algorithm can be stated as follows:

1. Initialize $\mathbf{B}^{(0)} = \mathbf{B}^{(1)}$, $\alpha^{(0)} = 0$ and $\alpha^{(1)} = 1$, $t = 1$, and $\delta = n / \{\lambda_{\max}(\mathbf{X}_{int}^T \mathbf{X}_{int})\}$.

2. repeat

$$\mathbf{S}^{(t)} = \mathbf{B}^{(t)} + \left(\frac{\alpha^{(t)} - 1}{\alpha^{(t)}}\right)(\mathbf{B}^{(t)} - \mathbf{B}^{(t-1)});$$

$$(\mathbf{B}_{temp})_0 = (\mathbf{S}^{(t)})_0;$$

for $d = 1 : |\widehat{\mathcal{M}}|$,

i. $(\mathbf{A}_{temp})_{l_d} = \mathbf{S}_{l_d}^{(t)} - \delta \nabla R(\mathbf{S}^{(t)})_{l_d}$;

ii. Compute singular value decomposition (SVD) $(\mathbf{A}_{temp})_{l_d} = \mathbf{U}_{l_d} \text{diag}(\mathbf{a}_{l_d}) \mathbf{V}_{l_d}^T$;

iii. $\mathbf{b}_{l_d} = \mathbf{a}_{l_d} - \lambda \delta * \mathbf{1}$;

iv. $(\mathbf{B}_{temp})_{l_d} = \mathbf{U}_{l_d} \text{diag}(\mathbf{b}_{l_d}) \mathbf{V}_{l_d}^T$;

end

Combine $(\mathbf{B}_{temp})_0$ and $\{(\mathbf{B}_{temp})_{l_d}, 1 \leq d \leq |\widehat{\mathcal{M}}|\}$ sub matrices and get the entire matrix \mathbf{B}_{temp} ;

$$\mathbf{B}^{(t+1)} = \mathbf{B}_{temp};$$

$$\alpha^{(t+1)} = \{1 + \sqrt{1 + (2\alpha^{(t)})^2}\} / 2; \quad t = t + 1;$$

3. until objective function $Q^*(\mathbf{B}^{(t)})$ converges.

For the above $p \times (|\widehat{\mathcal{M}}| + 1)q$ matrices \mathbf{A}_{temp} and \mathbf{B}_{temp} , $(\mathbf{A}_{temp})_{l_d}$ and $(\mathbf{B}_{temp})_{l_d}$ denote the $(dq + 1)$ -th to the $(dq + q)$ -th columns of the corresponding matrices, respectively.

2 Additional Simulation Results

Following the simulation study in Section 4.2 of the main paper, we first present additional simulation results for the case $(\sigma_e^2, s_n) = (1, 5000)$ in Figure S1 and those for the case $(\sigma_e^2, s_n) = (25, 5000)$ in Figure S2. The findings are similar to those given in Section 4.2 of the main paper.

Next, we consider the same simulation setting as Section 4.1 of the main paper. Specifically, we set the first four coefficient matrices as the four shapes of images including the cross (\mathbf{B}_{10}), square (\mathbf{B}_{20}), triangle (\mathbf{B}_{30}), and butterfly (\mathbf{B}_{40}). For all remaining coefficients, they were set as zero. We set the number of covariates to be $s = 2,000$ and $5,000$. We consider the same autoregressive type of covariance matrix for \mathbf{x}_i with $\rho_1 = 0.5$ and the same covariance matrix for \mathbf{E}_i . We still consider two different signal to noise ratios with $\sigma_e^2 = 1$ and $\sigma_e^2 = 25$. For ρ_2 , we use $\rho_2 = 0.5$. We consider different threshold values τ from 1 to 200. We run 100 replicates of Monte Carlo Studies to evaluate the finite sample performance of our screening procedure. We present the curves of percentage of the average true nonzero coverage proportion for different threshold values in Figure S3. The simulation results reveal that all methods perform very similarly, since the sizes of all effective regions of interest are quite large.

We further compare our estimates with centering and those without centering. We simulate 64×64 matrix responses according to model (1) with $s = 4$ covariates. All the settings are the same as those in Section 4.1 of the main paper except that we independently generate all scalar covariates \mathbf{x}_i from $N(\mathbf{1}, \Sigma_x)$ and include an intercept matrix \mathbf{B}_{00} , where each element of \mathbf{B}_{00} is set as 2. We only consider the case as $\rho_2 = 0.5$. To evaluate the estimation accuracy, we compute the mean squared errors of \mathbf{B}_l and calculate the prediction errors of $\widehat{\mathbf{B}}_l$ by generating $n^{test} = 500$ independent testing observations. Table S1 presents the estimation results from both methods. The results reveal that both methods perform similarly.

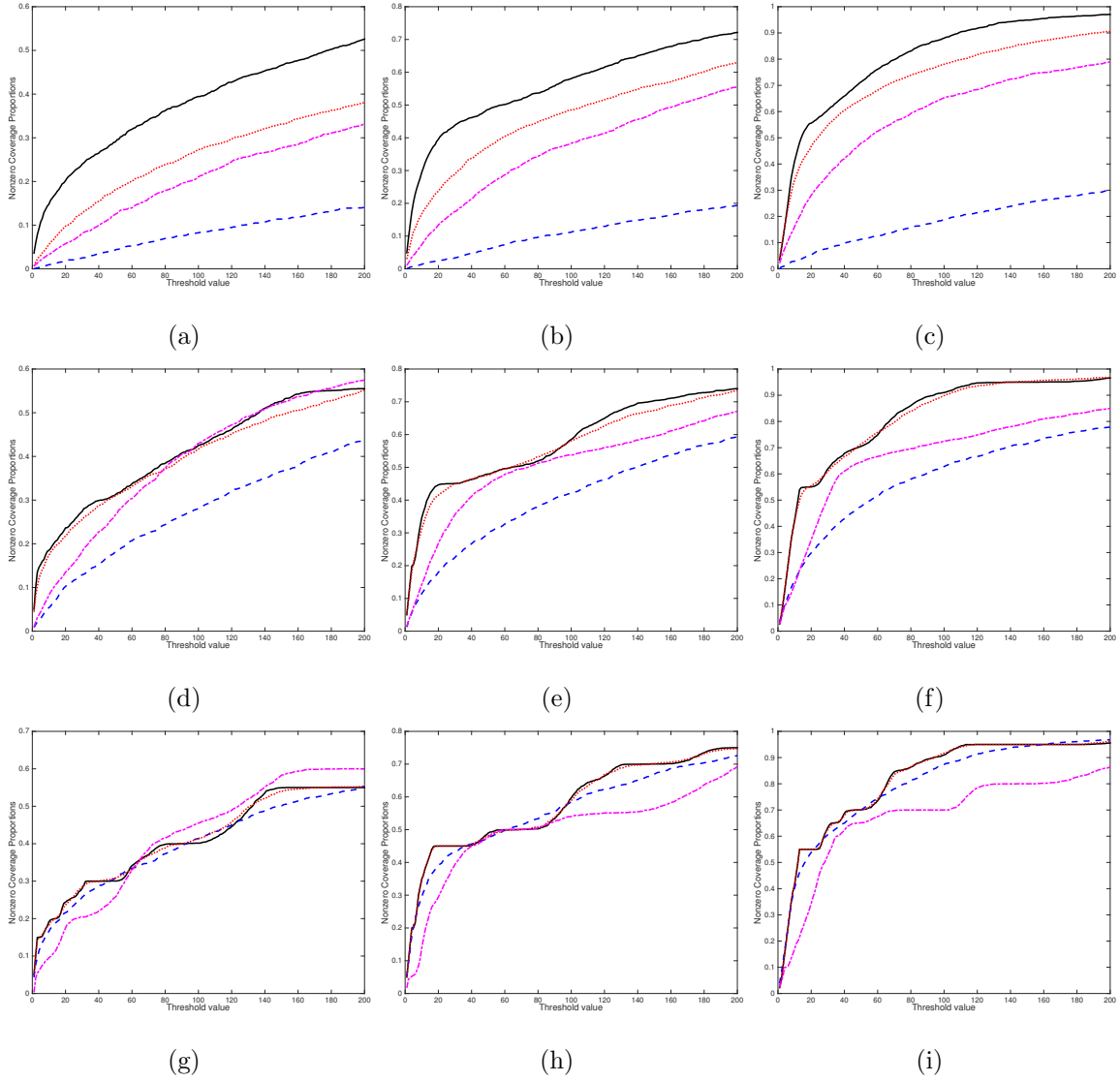


Figure S1. Screening results for the case $(\sigma_e^2, s_n) = (1, 5000)$: the curves of percentage of the average true nonzero coverage proportion. The black solid, blue dashed, red dotted and purple dashed dotted lines correspond to the rank-one screening, the L1 entrywise norm screening, the Frobenius norm screening and the global Wald test screening, respectively.

Panels (a)-(i) correspond to $(n, p_s, q_s) = (100, 4, 4), (200, 4, 4), (500, 4, 4), (100, 8, 8), (200, 8, 8), (500, 8, 8), (100, 16, 16), (200, 16, 16),$ and $(500, 16, 16)$, respectively.

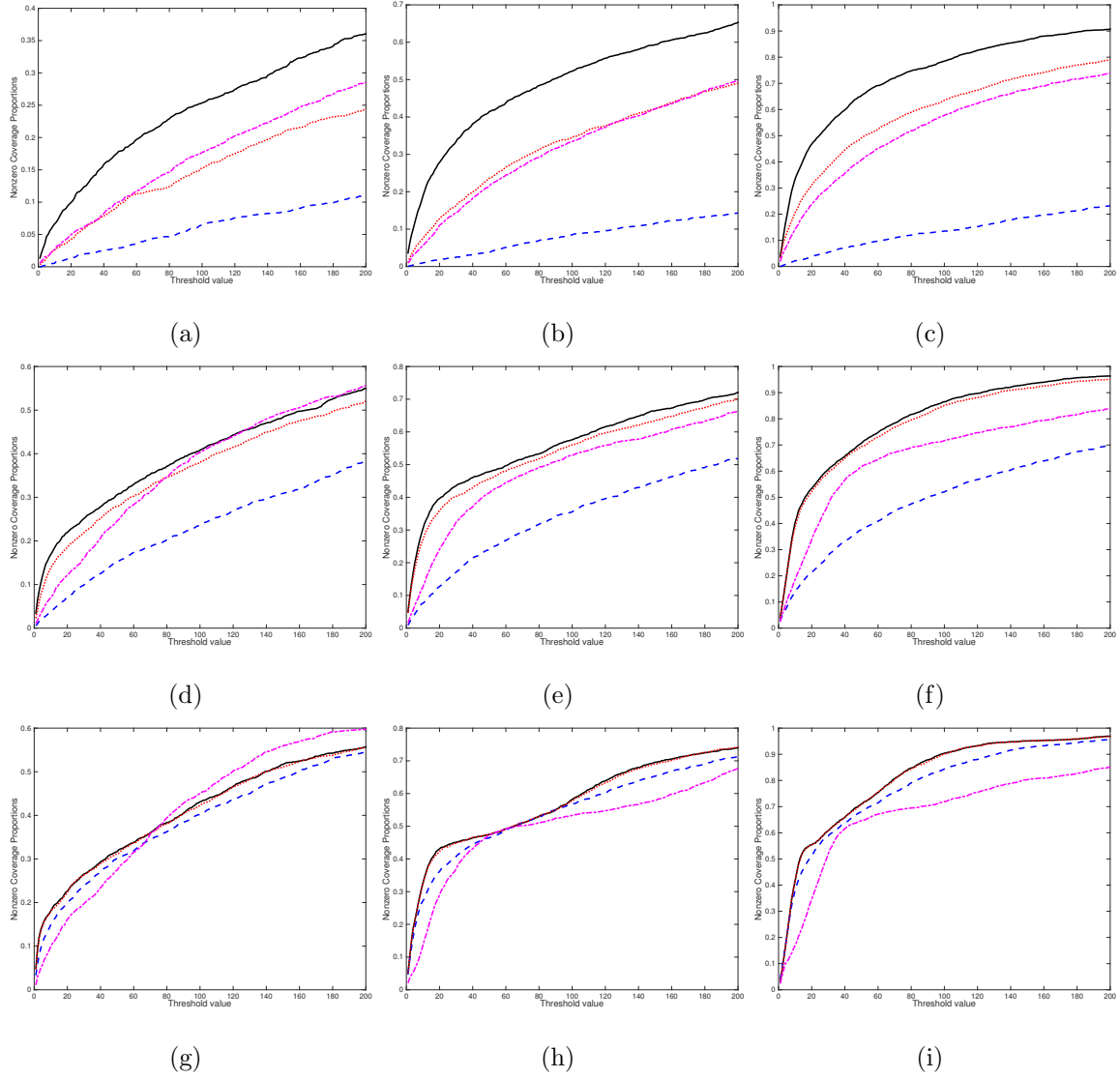


Figure S2. Screening results for the case $(\sigma_e^2, s_n) = (25, 5000)$: the curves of percentage of the average true nonzero coverage proportion. The black solid, blue dashed, red dotted, and purple dashed dotted lines correspond to the rank-one screening, the L1 entrywise norm screening, the Frobenius norm screening, and the global Wald test screening, respectively. Panels (a)-(i) correspond to

$$(n, p_s, q_s) = (100, 4, 4), (200, 4, 4), (500, 4, 4), (100, 8, 8), (200, 8, 8), (500, 8, 8), (100, 16, 16), (200, 16, 16), \text{ and } (500, 16, 16), \text{ respectively.}$$

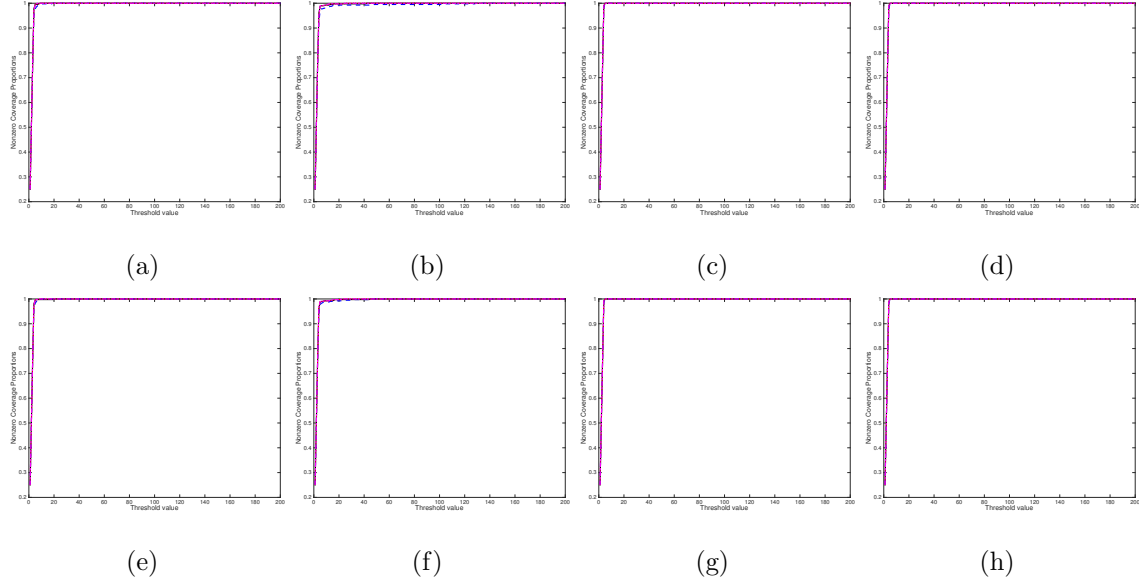


Figure S3. Screening results for true coefficients the same as Section 4.1 in the main paper: the curves of percentage of the average true nonzero coverage proportion. The black solid, blue dashed, red dotted, and purple dashed dotted lines correspond to the rank-one screening, the L1 entrywise norm screening, the Frobenius norm screening, and the global Wald test screening, respectively. Panels (a)-(d) correspond to cases when $(n, s_n, \sigma^2) = (100, 2000, 1), (100, 5000, 1), (200, 2000, 1), (200, 5000, 1)$ respectively. Panels (e)-(h) correspond to cases when $(n, s_n, \sigma^2) = (100, 2000, 25), (100, 5000, 25), (200, 2000, 25), (200, 5000, 25)$ respectively.

Table S1: The means of PEs and MSEs for our estimates with centering and without centering, and their associated standard errors in the parentheses. For each case, 100 simulated datasets are used.

(n, σ_ϵ^2)	Method	MSE(\mathbf{B}_1)	MSE(\mathbf{B}_2)	MSE(\mathbf{B}_3)	MSE(\mathbf{B}_4)	PE
(100, 1)	No Centering	15.25(0.40)	12.56(0.37)	43.93(0.61)	46.11(0.86)	1.03(0.0005)
	Centering	11.67(0.21)	9.96(0.22)	43.21(0.43)	44.88(0.52)	1.03(0.0002)
(200, 1)	No Centering	10.28(0.51)	7.62(0.23)	25.86(0.37)	26.48(0.48)	1.02(0.0004)
	Centering	7.27(0.09)	6.73(0.10)	23.77(0.20)	23.08(0.23)	1.02(0.0001)
(500, 1)	No Centering	5.80(0.32)	4.13(0.15)	11.97(0.20)	12.96(0.38)	1.01(0.0003)
	Centering	3.46(0.03)	3.53(0.03)	10.54(0.06)	9.75(0.06)	1.01(0.00003)
(100, 25)	No Centering	191.47(10.72)	137.76(3.95)	252.38(3.59)	318.34(9.54)	25.44(0.0098)
	Centering	121.61(1.69)	119.58(2.37)	227.58(2.01)	263.90(2.77)	25.37(0.0027)
(200, 25)	No Centering	84.38(1.13)	68.46(1.26)	175.97(1.43)	200.36(1.83)	25.27(0.0014)
	Centering	79.44(1.01)	71.27(1.25)	171.12(1.21)	201.43(1.63)	25.26(0.0013)
(500, 25)	No Centering	43.68(0.57)	36.87(0.55)	111.45(0.90)	124.90(0.84)	25.10(0.0005)
	Centering	42.17(0.50)	39.7(0.59)	110.16(0.79)	125.08(0.75)	25.10(0.0005)