Minimax Estimation of Simultaneously Low Rank and Sparse Matrices

Yuanyuan Shen[1], Zi Yin[2] and Feng Ruan[3]

[1,3] Department of Statistics and [2] Department of Electrical Engineering, Stanford University

Introduction

In this project, we study the problem of estimating matrices with a simultaneous structure: *low rank and sparsity*, under gaussian sequence and regression models. In both models, we use $\Theta^* \in \mathbf{R}^{p\times p}$ to denote the underlying sparse and low rank matrix. In gaussian sequence model, we observe some noisy matrix denoted by $X \in \mathbf{R}^{p\times p}$:

$$\textbf{Gaussian Sequence Model} \quad X = \Theta^* + E \quad E_{ij} \sim \mathcal{N}(0, \sigma^2),$$

while in the regression model, we observe the responses $y_i \in \mathbf{R}$ for $i = 1, 2, \ldots, N$ along with their covariates $x_i \in \mathbf{R}^p$:

$$\textbf{Regression Model} \quad y_i = x_i^T \Theta^* x_i + \epsilon_i, \epsilon_i \sim \mathcal{N}(0, \sigma^2)$$

Our aim is to analyze the fundamental limits of estimation of such low-rank and sparse matrix under the above models, using techniques borrowed from theoretical computer science and information theory.

The two models can have many potential applications. For example, in genetic studies, biologists are interested in understanding the interactions between genes, where they often model the underlying interaction matrices to have certain bi-clustering structure. Such bicluster structure can be well-captured by simultaneous low-rank and sparse matrices if additional sparsity assumptions on groups of interactive genes are enforced. Similarly, in social network studies, a problem of major interest is the 'community detection' problem, where sociologists need to construct the underlying adjacent matrix determined by hidden communities. One can characterize such matrices as simultaneously low rank and sparse matrices, as long as there are only a few different communities.

Background and Related Work

Here, we formalize the concept of low rank and sparse matrix. A matrix $A \in \mathbf{R}^{p\times p}$ is called simultaneously low-rank and sparse with parameters (s, r) if there exist $u_i \in \mathbf{R}^p$ and $v_i \in \mathbf{R}^p$ with $\|u_i\|_0 \le s$, $\|v_i\|_0 \le s$, $\|u_i\|_2 = \|v_i\|_2 = 1$, and $\theta_i \in \mathbf{R}$ such that

$$A = \sum_{i=1}^{r} \theta_i u_i v_i^T.$$

In this project, our main focus is to evaluate the minimax rate of the estimation of sparse low-rank matrices with parameters $(r, s) = (1, s)$, i.e., all rank-one matrices that have only one s by s block. Note that, although our definition of simultaneously sparse and low-rank matrices is close to that appeared in [1], the resulting minimax rate for these two classes of matrices are in fact different, as [1] requires additional assumptions on top eigenvalues. In addition, the techniques that evaluate the minimax rate in our class of matrices are also fundamentally different from those appeared in [1]. To evaluate the minimax rate of estimation for our defined matrix class, we borrow ideas from literatures that focus on sparse PCA, e.g., [2], [3] and [4]. In these papers, the authors analyze the fundamental limits under some different yet similar models, where they need to estimate the top sparse eigenvector of some low rank and sparse covariance matrices. In their papers, the authors prove that, the ℓ_2 minimax rate for estimation of the top eigenvectors is $\sqrt{s \log p/N}$ up to a constant. In addition to that, [4] shows that roughly no convex procedures can actually achieve this minimax rate, and the price for convex procedures to pay is that a factor of $\sqrt{s}$ will appear in their ℓ_2 error rate. In this project, we show that parallel results hold in the estimation of the entire matrices, where both the top eigenvectors and corresponding eigenvalues are simultaneously estimated. The minimax ℓ_2 rate for matrix estimation for the classes of matrices with $(r, s) = (1, s)$, to our surprise, is still $\sqrt{s \log p/N}$. We also provide a fast convex regularization procedure that can actually achieve the minimax ℓ_2 rate within a multiplication factor of $\sqrt{s}$. This suggests that a deeper connection may exist between our estimation problem and sparse PCA problem.

Main Results

Proposition 1 [Minimax Rate for Gaussian Sequence Model]
Under Gaussian sequence model, the minimax rate of the estimation of simultaneously low rank and sparse matrix with parameters $(r, s) = (1, s)$ is $\asymp \sigma\sqrt{s \log p}$. More precisely, there exist some universal constants $c, C, c_1, c_2 > 0$ such that

$$\inf_{\hat{\Theta}} \sup_{\Theta\in\Theta(1,s)} P\left(\|\hat{\Theta} - \Theta\|_F \ge c\sigma\left(\sqrt{s \log \frac{p}{s}}\right)\right) \ge \frac{1}{2}$$

$$\inf_{\hat{\Theta}} \sup_{\Theta\in\Theta(1,s)} P\left(\|\hat{\Theta} - \Theta\|_F \ge C\sigma\left(\sqrt{s \log \frac{p}{s}}\right)\right) \le c_1 \exp(-c_2(s \log p/s)/\sigma^2),$$

where

$$\Theta(r, s) = \left\{\Theta \in \mathbf{R}^{p\times p} \mid \Theta = \sum_{i=1}^{r} \theta_i u_i v_i^T, \|u_i\|_2 = \|v_i\|_2 = 1, \|u_i\|_0 \le s, \|v_i\|_0 \le s\right\}$$

Remark 1 Proposition 1 parallels the result in sparse PCA literature [3], where estimation of sparse eigenvector achieves the same minimax error rate. (Note that $\sigma \asymp \frac{1}{\sqrt{N}}$).
Remark 2 The result can be generalized to the estimation problem of sparse and low rank rectangular matrix with different height and width.

Proposition 2 [Soft-Thresholding achieves Sub-optimal Rate]
Let $l_\lambda(x) := \mathbf{sign}(x)(|x| - \lambda)_+$. Consider the following thresholding estimator $\hat{\Theta} := l_\lambda(X)$, where we do entry-wise soft thresholding on the observed matrix X. Then, if we choose λ to be greater than $\sigma\sqrt{2 \log p}$, then there exists some universal constants $c, C > 0$ such that

$$\|\hat{\Theta} - \Theta^*\|_F \le Cs\lambda$$

holds with probability greater than $1 - p^{-c}$.

Remark 3 Soft-thresholding is easy to compute in practice, yet the price to pay for the speed is the scaling of the statistical convergence rate by a factor of $\sqrt{s}$.

Proposition 3 [Minimax Rate for Regression Model]
Assume the following regularity condition on the design matrix X: $\exists \kappa, \eta > 0$ such that $\forall \Delta \in \{\Delta \mid \|\Delta\|_F = 1, \mathbf{rank}\,\Delta = 2\}$:

$$\kappa \le \frac{1}{N}\sum_{i=1}^{N}\left(x_i^T \Delta x_i\right)^2 \le \eta. \tag{1}$$

Then, the minimax rate of estimation of $\Theta^* \in \Theta(1, s)$ is $\asymp \sigma\sqrt{s \log p/s}$. More precisely, there exist some constants $c, C > 0$ and $c_1, c_2 > 0$ that are only dependent on κ and η, such that

$$\inf_{\hat{\Theta}} \sup_{\Theta\in\Theta(1,s)} P\left(\|\hat{\Theta} - \Theta\|_F \ge c\left(\sigma\sqrt{s \log \frac{p}{s}}\right)\right) \ge \frac{1}{2}$$

$$\inf_{\hat{\Theta}} \sup_{\Theta\in\Theta(1,s)} P\left(\|\hat{\Theta} - \Theta\|_F \ge C\left(\sigma\sqrt{s \log \frac{p}{s}}\right)\right) \le c_1 \exp(-c_2 N(s \log p/s))$$

Proposition 4 [ℓ_1 regularization achieves Sub-optimal Rate]
Consider the following penalized ℓ_1 regression:

$$\hat{\Theta} := \operatorname*{argmin}_{\Theta} \frac{1}{N}\sum_{i=1}^{N}(y_i - x_i^T \Theta x_i)^2 + \lambda\|\Theta\|_1$$

Suppose assumption (Eq.1) on X still holds. Then for any λ that is greater than $2\sigma\sqrt{\log p/N}$, there exists some constants $c, C > 0$ that are only dependent on κ and η such that

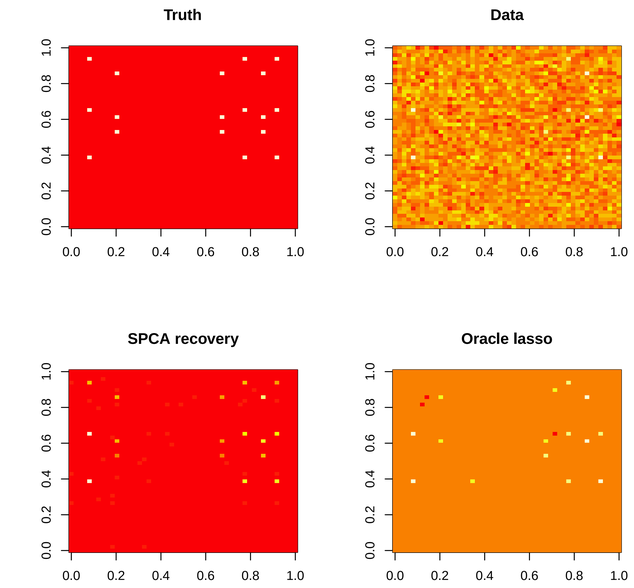
$$\|\hat{\Theta} - \Theta^*\|_F \le Cs\lambda$$

holds with probably at least $1 - p^{-c}$

Remark 4 The assumption Eq.1 can hold w.h.p in general random design as long as the row of X are i.i.d sampled from some distribution with compact support and the sample size $N = \Omega\left(\max\{s^3, s \log p\}\right)$

Simulation Studies

In our simulation setting, $p = 50$, $N = 100$, $r = 2$, $s = 3$. $\theta_1 = \theta_2 = 1$. The SPCA achieves better performance in both estimation and support recovery than that of the oracle lasso procedure.



References

[1] D. Yang, Z. Ma and A. Buja. *Journal of Machine Learning Research*, to appear.
[2] A. Amina and M. Wainwright. *The Annals of Statistics*, 2009, 37(5B): 2877-2921.
[3] I. Johnstone and A. Lu. *Journal of the American Statistical Association*, 2012.
[4] Q. Berthet and P. Rigollet. *The Annals of Statistics*, 2013, 41(4):1780-1815.