MATRIX COMPLETION VIA NON-CONVEX PROGRAMMING

By Diego Franco Saldana and Haolei Weng

We consider the matrix completion problem in the noisy setting. To achieve statistically efficient estimation of the unknown low-rank matrix, solving convex optimization problems with nuclear norm constraints has been both theoretically and empirically proved a successful strategy under certain regularity conditions. However, the bias induced by the nuclear norm penalty may compromise the estimation accuracy. To address this problem, following a parallel line of research in sparse regression models, we study the performance of a family of non-convex regularizers in the matrix completion problem. In particular, a fast first-order algorithm is proposed to solve the non-convex programming problem. We also describe a degree-of-freedom based reparametrization to "refine" the search along the solution path. Numerical experiments show that these non-convex methods outperform the traditional use of nuclear norm regularization in both simulated and real data sets.

1. Introduction. The matrix completion problem consists in the recovery of a data matrix based on a sampling of its entries [1]. A popular example is the Netflix competition [2]. Denote the data matrix by $M_{m\times n}$, the rows corresponding to users and columns to movies. The entry M_{ij} is the rating user i gives to movie j. Typically, users only rate a few movies so that there is only a small fraction of the matrix M observed. Netflix would like to complete this matrix in order to recommend potentially appealing movies to users. Generally, it is impossible to correctly predict the ratings of users for unseen movies without additional information about M. A useful fact is that M is approximately low-rank because ratings are mainly determined by a small number of factors such as movie genre and user taste.

This problem can be formulated in a more familiar way to the statistics literature. Let $\beta \in \mathcal{R}^{mn}$ be the long vector obtained after stacking the columns of data matrix $M_{m\times n}$. The observed entries form a shorter vector $Y = X\beta$, where the matrix $X_{k\times mn}$ (k is the number of observations) is a linear operator mapping the data matrix to its observed entries. Since it is fairly common to have $k \ll mn$ in many statistical applications, the matrix completion problem can be considered as a high-dimensional regression problem. Based on the response Y and design matrix $X_{k\times mn}$, the objective

Keywords and phrases: Low-Rank, Nuclear Norm, Soft-Thresholding, Singular Value Decomposition, Non-Convex Penalty, Degrees-of-freedom, Recalibration.

is to estimate the coefficient vector β. It is well known that high-dimensional regression problems require extra structure in β to make statistical inference feasible. An active line of research consists in estimating a sparse β with only a few non-zero components [3, 4, 5, 6]. In our case, the structure of interest in the data matrix $M_{m\times n}$ is low-rank instead of sparsity. Although the matrix completion problem can be framed into the high-dimensional regression framework, it distinguishes itself as a more difficult problem in two respects. Firstly, from a computational point of view, the dimensionality of β can easily scale to the tens of millions and above. More importantly, low-rank inducing optimization problems involve advanced algorithmic analysis such as characterizing the subgradient of nuclear norm. Secondly, as well known in sparse regression setting, to achieve optimal estimation of β, the design matrix needs to be well "conditioned". For instance, the Restricted Isometry Property (RIP) in [6] requires that every set of columns with cardinality less than a given number behaves like an orthonormal system. However, due to the sampling nature of the matrix completion problem, $X_{k\times mn}$ never satisfies RIP [7]. Therefore, a less restrictive condition on X, or a more delicate low rank structure on β is required to recover the data matrix $M_{m\times n}$.

The report is organized as follows. We give a brief literature survey on the theoretical developments in estimating low-rank matrices in Section 2. In Section 3, we review a few algorithms for solving convex optimization problems with nuclear norm constraints. We then propose and study a non-convex programming approach in Section 4. Numerical experiments are conducted in Section 5. Section 6 discusses potential research directions and open problems.

2. Matrix Completion Theory. To fix notations, denote the unknown low-rank matrix by $M_{m\times n}$ $(m \geq n)$ with rank r, and the set of locations corresponding to observed entries of M by Ω. Let k be the cardinality of Ω. P_Ω is the projection mapping such that $(P_\Omega(X))_{ij} = X_{ij}$ if $(i,j) \in \Omega$ and $(P_\Omega(X))_{ij} = 0$ if $(i,j) \notin \Omega$. For a vector a, $\|a\|_p^p = \sum_i |a_i|^p$ $(1 \leq p \leq +\infty)$. For any matrix A, $\|A\|_* = \sum \sigma_i(A)$, $\|A\|_F^2 = \sum \sigma_i^2(A)$, $\|A\|_2 = \sigma_1(A)$, where $\sigma_1(A) \geq \sigma_2(A) \geq \cdots \geq \sigma_r(A) \geq 0$ are singular values of A.

2.1. *Exact Matrix Completion.* The first theoretical guarantee for exact recovery of a low-rank matrix appeared in the compressed sensing community [1]. The problem is to exactly recover M based on $P_\Omega(M)$. The authors in [1] propose the following convex optimization approach,

$$\begin{aligned} &\text{minimize} && \|X\|_* \\ &\text{subject to} && P_\Omega(X) = P_\Omega(M). \end{aligned} \tag{2.1}$$

A surprising result shown in their paper is that this simple approach can achieve exact recovery for a broad family of "incoherent" matrices. Let $M = U\Sigma V^T$ be the SVD of M with $U = (u_1, \ldots, u_r)$ and $V = (v_1, \ldots, v_r)$. Incoherence as defined in [1] is characterized by two conditions:

A. $\max_{1\le i\le m} \frac{m}{r}\|U^T e_i\|_2^2 \le \mu_0, \quad \max_{1\le j\le n} \frac{n}{r}\|V^T e_j\|_2^2 \le \mu_0$
B. $\max_{1\le i\le m, 1\le j\le n} \sqrt{\frac{mn}{r}}(UV^T)_{ij} \le \mu_1,$

where (e_i) is the standard basis, and μ_0, μ_1 are two constants. Condition **A** requires the row and column spaces of M to be diffuse such that information about unseen entries can be inferred from the observed ones. The main result in [1] claims that

> If M satisfies conditions **A** and **B**, and the observed set Ω is sampled uniformly at random, with $r \le \mu_0^{-1} m^{1/5}, k \ge C\mu_0 m^{6/5} r(\beta \log m)$, then the solution to (2.1) is unique and equal to M with probability at least $1 - cm^{-\beta}$.

Essentially, a low-rank matrix M with incoherence constant $\mu_0 = O(1)$ can be exactly recovered through $\Theta(m^{6/5} r \log m)$ entries with high probability. A fundamental and interesting observation is that any $m \times n$ matrix of rank r depends on $(m + n - r)r$ degrees of freedom. Clearly, exact matrix recovery is impossible if the number of observations is less than this intrinsic dimensionality. But, is it possible to recover M with a minimum number of measurements approaching the above limit (at least up to logarithmic multiplicative factors)? This problem has been addressed in [8]. The authors introduce the "strong incoherence" condition:

C. $|< Ue_i, Ue_j > - \frac{r}{m} I_{(i=j)}| \le \mu_2 \frac{\sqrt{r}}{m}, \quad 1 \le i, j \le m,$
$|< Ve_i, Ve_j > - \frac{r}{n} I_{(i=j)}| \le \mu_2 \frac{\sqrt{r}}{n}, \quad 1 \le i, j \le n.$

M is said to obey the "strong incoherence" property with parameter μ if it satisfies conditions **B** and **C** with $\max(\mu_1, \mu_2) \le \mu$. Condition **C** is generally more restrictive than condition **A**. For instance, condition **C** implies that $\frac{m}{r}\|U^T e_i\| \le 1 + \frac{\mu_2}{\sqrt{r}}$ and $\frac{n}{r}\|V^T e_j\| \le 1 + \frac{\mu_2}{\sqrt{r}}$. The authors in [8] show that

> If M obeys the strong incoherence property with parameter μ, and the observed set Ω is sampled uniformly at random with $k \ge C\mu^2 mr \log^6 m$, then M is the unique solution to (2.1) with probability at least $1 - m^{-3}$.

For strong incoherent matrices with $\mu = O(1)$, the number of measurements sufficient for exact recovery can be reduced from $\Theta(m^{6/5} r \log m)$ to $\Theta(mr \log^6 m)$. The authors further prove that, under a Bernoulli model where each entry is independently observed with probability $\frac{k}{n^2}$ $(m = n)$, if $-\log(1 - \frac{k}{m^2}) \ge \frac{\mu_0 r}{m} \log(\frac{m}{2\delta})$ does not hold, it is impossible to recover an incoherent matrix M with parameter μ_0 by any algorithm with probability larger than δ. In other words, a necessary condition to guarantee success

with probability at least δ is that

$$k \geq m^2(1 - e^{-\frac{\mu_0 r}{m}\log\frac{m}{2\delta}}).$$

When $\frac{\mu_0 r}{m}\log\frac{m}{2\delta} = o(1)$, it follows that $k = \Theta(\mu_0 mr\log(\frac{m}{2\delta}))$. That is to say, to recover an incoherent matrix with constant μ_0, the simple convex approach (2.1) only requires minimum (up to logarithmic factors) observations.

The proofs provided in [1, 8] involve sophisticated algebra manipulations and moment inequality arguments which are highly technical. More recently, [9] presented a much easier proof and sharpened the previous results. This paper shows that

> If M satisfies conditions **A** and **B**, and the observed set Ω is sampled uniformly at random, with $k \geq 32\max\{\mu_1^2, \mu_0\}r(n+m)\beta\log^2(2m)$ for $\beta > 1$, then M is the unique solution to (2.1) with probability at least $1 - 6\log(m)(n+m)^{2-2\beta} - m^{2-2\beta^{1/2}}$.

For $\mu_0 = O(1)$, the number of measurements needed is $\Theta(\mu_1^2 rm\log^2 m)$. For a large proportion of incoherent matrices, [1] shows that $\mu_1 = O(\log m)$. Thus, k only needs to be of the order $\Theta(rm\log^4 m)$. The proof in [9] is relatively simple; it is adapted from the quantum information community and only uses basic matrix analysis and concentration inequalities.

Note that the uniform sampling assumption appears in all results provided in [1, 8, 9]. Rather than being an intrinsic statistical modeling concern, these authors introduce the uniform sampling assumption in order to avoid worst-case sampling behavior. In reality, however, uniform sampling is likely to be an improper assumption. For example, in the Netflix data set, popular movies have more ratings and active users rate more movies than average users. The ratings available are expected to be highly non-uniform. More importantly, the key incoherence condition required in [1, 8, 9] is mainly for uniform sampling to capture the essential information in M with high probability. If the sampling can be adapted to the structure of M (e.g., more important entries are more likely to be sampled), then the incoherence conditions can be weakened. This is the spirit of [10], which defines the local coherences as follows:

$$\tilde{\mu}_i = \frac{m}{r}\|U^T e_i\|^2, \quad i = 1, \ldots, m.$$
$$\tilde{\nu}_j = \frac{n}{r}\|V^T e_j\|^2, \quad j = 1, \ldots, n.$$

The authors prove that

> If M has local coherence parameters $\{\tilde{\mu}_i, \tilde{\nu}_j\}$, and each element (i,j) is independently observed with probability p_{ij} such that $p_{ij} \geq \min\left\{c_0\frac{(\tilde{\mu}_i + \tilde{\nu}_j)r\log^2(n+m)}{n}, 1\right\}$,

$p_{ij} \geq \frac{1}{n^{10}}$, then M is the unique solution to (2.1) with probability at least $1 - c_1(n+m)^{-c_2}$, for some universal constants $c_0, c_1, c_2 > 0$.

Note that the expected number of observed entries satisfies

$$\begin{aligned} \sum_{i,j} p_{ij} &\geq \max\Big\{c_0 \frac{r\log^2(n+m)}{n}\sum_{i,j}(\tilde{\mu}_i + \tilde{\nu}_j), \sum_{i,j}\frac{1}{n^{10}}\Big\} \\ &= 2c_0 mr\log^2(n+m) \end{aligned}$$

which does not depend on the local coherence parameters. The Azuma-Hoeffding inequality then implies

Under the local coherence sampling given by the $\{p_{ij}\}$ above, the matrix M is unique solution to (2.1) and $k \leq 3c_0 mr\log^2(m+n)$ with probability at least $1 - c_1'(n+m)^{-c_2'}$.

For M with incoherence parameter μ_0, it is easy to see that $\tilde{\mu}_i \leq \mu_0$ and $\tilde{\nu}_j \leq \mu_0$. Hence, considering uniform sampling as a very special case of local coherence sampling with probability $p_{ij} \geq c\frac{\mu_0 r\log^2 m}{n}$, criterion (2.1) exactly recovers M using $k = \Theta(\mu_0 mr\log^2 m)$ entries with high probability. For $\mu_0 = O(1)$, $k = \Theta(mr\log^2 m)$ achieves the best existing scaling. More interestingly, condition **B** is dropped!

2.2. *Stable Matrix Completion.* In the real world, observations are likely corrupted by noise. Thus, a more realistic model is

$$Y_{ij} = M_{ij} + Z_{ij}, \quad (i,j) \in \Omega,$$

where (Z_{ij}) are zero-mean noise terms and (Y_{ij}) are observations. Under this model, criterion (2.1) tends to overfit. To achieve a better bias and variance trade-off, the authors in [7] propose to solve

$$\begin{aligned} &\text{minimize} \quad \|X\|_* \\ &\text{subject to} \quad \|P_\Omega(X - Y)\|_F \leq \delta, \end{aligned} \tag{2.2}$$

under the assumption that $\|P_\Omega(Z)\|_F \leq \delta$. Let the minimizer of (2.2) be $\hat{M}$. In [7], it is shown that

If M satisfies the strong incoherence condition, then with high probability, $\hat{M}$ obeys

$$\|\hat{M} - M\|_F \leq 4\sqrt{\frac{(2+p)n}{p}}\delta + 2\delta, \tag{2.3}$$

where $p = \frac{k}{mn}$.

The error upper bound is approximately of the order $\sqrt{\frac{n}{p}}\delta$. To have a better understanding of the bound, [7] compares it with two additional error bounds. The first one assumes the linear sampling operator $\mathcal{A}$ obeys the RIP,

$$(1-\Delta)\|X\|_F^2 \leq \frac{1}{p}\|\mathcal{A}(X)\|_F^2 \leq (1+\Delta)\|X\|_F^2$$

for all small rank matrices X and $\Delta < 1$. Note that the RIP does not hold for the linear operator P_Ω. Then, with high probability, $\hat{M}$ obeys

$$\|\hat{M} - M\|_F \leq C_0 p^{-1/2}\delta$$

The second one considers an oracle estimator. Let the tangent space of M be

$$T(M) = \{UX^T + YV^T \mid X \in \mathcal{R}^{n\times r}, Y \in \mathcal{R}^{m\times r}\}.$$

The linear space $T(M)$ is the analogous to the vector space with the same support as the least squares estimator in the linear regression setting. Suppose we have prior information about T, then a natural approach is solving

$$\begin{aligned} &\text{minimize} && \|P_\Omega(X) - P_\Omega(Y)\|_F \\ &\text{subject to} && X \in T(M). \end{aligned} \tag{2.4}$$

It it shown in [7] that the minimizer $\hat{M}$ of (2.4) obeys

$$\|\hat{M} - M\|_F \approx p^{-1/2}\delta.$$

In summary, the convex programming criterion (2.2) achieves stable matrix completion (error proportional to noise level), but loses efficiency by a factor of $\sqrt{n}$, compared to the oracle estimator and the "compressed sensing" estimator assuming the RIP.

A remarkably different approach using non-convex methods is proposed in [11]. The basic idea is as follows:

1. Trim $P_\Omega(Y)$ by setting the rows and columns with many revealed entries to zero.
2. Calculate the best rank r approximation to the trimmed matrix.
3. Minimize the cost function $F(P,Q) = \min_{S\in\mathcal{R}^{r\times r}} \mathcal{F}(P,Q,S)$, where $\mathcal{F}(P,Q,S) = \frac{1}{2}\sum_{(i,j)\in\Omega}(Y_{ij} - (PSQ^T)_{ij})^2$.

Step 3 is solved by using a gradient descent algorithm on matrix manifolds, where the projected trimmed matrix serves as an initial point. In practice, however, the rank r has to be estimated. For the method to work, [11] introduces a slightly different incoherent condition:

D. $\max_{1\le i\le m,1\le j\le n}\sqrt{\frac{mn}{r}}|\sum_{l=1}^{r}U_{il}(\sigma_l/\sigma_1)V_{jl}|\le\mu_3$,

where $\sigma_1\ge\cdots\ge\sigma_r>0$ are the nonzero singular values of the matrix M. Let $\kappa=\frac{\sigma_1}{\sigma_r}$ and the non-convex method estimator be $\hat{M}^*$. The authors in [11] show that

> If M obeys conditions **A** and **D**, and Ω is sampled uniformly at random with $k\ge C\sqrt{mn}\,\kappa^2\max\{\mu_0 r\sqrt{\frac{m}{n}}\log n,\mu_0^2 r^2\frac{m}{n}\kappa^4,\mu_3^2 r^2\frac{m}{n}\kappa^4\}$, then with probability at least $1-1/n^3$,
>
> $$\frac{1}{\sqrt{mn}}\|\hat{M}^*-M\|_F\le C'\kappa^2\frac{\sqrt{rmn}}{k}\|P_\Omega(Z)\|_2. \tag{2.5}$$

The bound given in [7] can be nearly reformulated as

$$\frac{1}{\sqrt{mn}}\|\hat{M}-M\|_F\le 7\sqrt{\frac{n}{k}}\|P_\Omega(Z)\|_F+\frac{2}{\sqrt{mn}}\|P_\Omega(Z)\|_F. \tag{2.6}$$

Generally, $\hat{M}^*$ and $\hat{M}$ are not directly comparable since $\hat{M}$ requires the strong incoherence condition while $\hat{M}^*$ needs the condition number κ to be bounded. In terms of the error bound, expression (2.5) does not include the second term in (2.6). Suppose $m=n$, then error bound in (2.5) is $\Theta(\frac{n}{k}\sqrt{r}\|P_\Omega(Z)\|_2)$ while the first term in (2.6) is $\Theta(\sqrt{\frac{n}{k}}\|P_\Omega(Z)\|_F)$. In the case where $P_\Omega(Z)$ is i.i.d Gaussian with variance τ^2 and rank $r=o(n)$, it is shown in [11] that the bound in (2.5) is $\Theta(\tau\sqrt{\frac{nr}{k}})$, having the same order as the oracle estimator provided in [7].

An alternative way to estimate the low-rank matrix is optimizing a penalized loss function as widely seen in the regression setting:

$$\min_X\ \frac{1}{2}\|P_\Omega(Y)-P_\Omega(X)\|_F^2+\lambda\|X\|_*\,.$$

The authors in [12] study this type of estimator in a more general setting with approximately low-rank structures and weighted sampling. For simplicity, we only present their results for exact low-rank matrices under uniform sampling of the observations. The key idea in establishing an error bound for the corresponding M-estimator is deriving a *restricted strong convexity* (RSC) condition as proposed in the unified framework introduced in [13]:

$$\frac{\|P_\Omega(X)\|_F}{\sqrt{n}}\ge c\|X\|_F\,. \tag{2.7}$$

As is easily seen, inequality (2.7) does not hold (with high probability) for matrices with only one nonzero entry. To avoid these overly "spiky" matrices, the authors in [12] define a tractable measure of "spikiness" as

$$\alpha_{sp}(X)=\sqrt{mn}\frac{\|X\|_\infty}{\|X\|_F}\,, \tag{2.8}$$

where $\|X\|_\infty$ is the elementwise ℓ_∞-norm. Note that $1 \le \alpha_{sp} \le \sqrt{mn}$, with $\alpha_{sp} = 1$ when X has entries that are all equal, and $\alpha_{sp} = \sqrt{mn}$ if X has a single nonzero entry. The authors consider the M-estimator

$$\hat{M} \in \arg \min_{\|X\|_\infty \le \frac{\alpha^*}{\sqrt{mn}}} \frac{1}{2k} \|P_\Omega(X) - P_\Omega(Y)\|_F^2 + \lambda_n \|X\|_* \,,$$

where $\alpha^* \ge 1$ is a measure of "spikiness". Assuming $n = m$, it has been shown in [12] that

> If $P_\Omega(Z)$ is i.i.d. sub-exponential with variance τ^2, M has rank at most r, Frobenius norm at most 1, and $\alpha_{sp}(M) \le \alpha^*$, then $\hat{M}$ with $\lambda_n = 4\tau\sqrt{\frac{n \log n}{k}}$ obeys
>
> $$\|\hat{M} - M\|_F^2 \le c_1' \max\{\tau^2, 1\} (\alpha^*)^2 \frac{rn \log n}{k} + \frac{c_1 (\alpha^*)^2}{k} \tag{2.9}$$
>
> with probability at least $1 - c_2 e^{-c_3 \log n}$.

To examine how good the proposed M-estimator is, the authors further establish an information-theoretic lower bound. Define the minimax risk in Frobenius norm

$$B(r) = \{X \in \mathcal{R}^{n \times n} \mid \operatorname{rank}(X) \le r, \alpha_{sp}(X) \le \sqrt{32 \log n}\}$$
$$\mathcal{R}(B(r)) = \inf_{\hat{M}} \sup_{M \in B(r)} E[\|\hat{M} - M\|_F^2] \,.$$

Then, there is a universal constant $c > 0$ such that

$$\mathcal{R}(B(r)) \ge c\tau^2 \frac{rn}{k} \,.$$

The dominant term in (2.9) matches the minimax lower bound up to logarithmic factors. The spikiness condition is essentially different from the incoherence or strong incoherence conditions introduced in [7, 11]. The incoherence conditions only involve the left and right singular vectors of M, while the spikiness condition also depends on its singular values. For exact recovery in the noiseless setting, dependency only on the singular vectors is reasonable because they determine the tangent space of M. In the noisy case, however, the authors argue that additional conditions on the singular values also play an important role. They point out an example where their spikiness condition holds and incoherence conditions are violated. The detailed comparison of error bounds with those in [7, 11] can also be found in [12].

2.3. *General Low Rank Matrix Estimation.* The sampling nature of P_Ω makes the theoretical development of the matrix completion problem of particular interest. On the other hand, there are several applications apart from the matrix completion setting that also explore low-rank structures of data matrices. Examples include compressed sensing, face recognition and multivariate regression [14, 15]. Thus, a general set up is

$$Y = \mathcal{A}(M) + Z\,,$$

where $\mathcal{A}$ is a general linear operator and Z is the noise matrix. A natural approach in estimating M is solving

$$\min_X \quad \|Y - \mathcal{A}(X)\|_F^2 + \lambda_n \|X\|_G\,, \tag{2.10}$$

where $\|X\|_G$ is a low-rank inducing penalty such as nuclear norm or Schatten p-norm. We refer to the papers [14, 15, 16, 17] for a detailed theoretical study of this general approach.

3. Matrix Completion Algorithms. We consider solving two optimization problems. Namely, the exact recovery approach

$$\begin{aligned} &\text{minimize} \quad \|X\|_* \\ &\text{subject to} \quad P_\Omega(X) = P_\Omega(M)\,, \end{aligned} \tag{3.1}$$

and the stable recovery method

$$\min_X \quad \frac{1}{2}\|P_\Omega(X) - P_\Omega(Y)\|_F^2 + \lambda \|X\|_*\,. \tag{3.2}$$

In [17], it is shown that

$$\begin{aligned} \|X\|_* = \min_{W_1, W_2} \quad & \frac{1}{2}(\mathrm{tr}(W_1) + \mathrm{tr}(W_2)) \\ \text{s.t.} \quad & \begin{bmatrix} W_1 & X \\ X^T & W_2 \end{bmatrix} \succeq 0. \end{aligned} \tag{3.3}$$

Hence, (3.1) can be reformulated as the semidefinite programming problem:

$$\begin{aligned} \min_{X, W_1, W_2} \quad & \frac{1}{2}(\mathrm{tr}(W_1) + \mathrm{tr}(W_2)) \\ \text{s.t.} \quad & \begin{bmatrix} W_1 & X \\ X^T & W_2 \end{bmatrix} \succeq 0 \\ & P_\Omega(X) = P_\Omega(M)\,. \end{aligned} \tag{3.4}$$

Interior-point methods are readily available to solve (3.4). However, this type of algorithms are not scalable to large matrices with over a million entries. The authors in [18] propose a fast, first-order algorithm which explores low-rank and sparse matrix structures to save computational and storage cost. The idea is to approximate the solution of (3.1) by solving a sequence of closely related problems as follows

$$\begin{aligned} &\text{minimize} \quad \tau\|X\|_* + \frac{1}{2}\|X\|_F^2 \\ &\text{subject to} \quad P_\Omega(X) = P_\Omega(M)\,. \end{aligned} \tag{3.5}$$

The unique solution $\hat{X}_\tau$ of (3.5) can be efficiently calculated by a Lagrange multiplier method. As $\tau \to \infty$, $\hat{X}_\tau$ converges to the solution of (3.1). We skip the implementation details here and focus instead on (3.2).

Problem (3.2) belongs to a generic class of optimization problems of the form

$$\min_x \Big\{ F(x) = f(x) + g(x) \Big\}, \tag{3.6}$$

where $f(x)$ is a continuously differentiable convex function with Lipschitz continuous gradient $L(f)$, and $g(x)$ is a continuous convex function which is possibly non-smooth. The authors in [19] present a fast iterative algorithm which solves an *approximation problem* at each iteration:

$$x_{t+1} = \arg\min_x \Big\{ Q_L(x, x_t) = f(x_t) + < x - x_t, \nabla f(x_t) > + \frac{L}{2}\|x - x_t\|^2 + g(x) \Big\}.$$

Equivalently,

$$x_{t+1} = \arg\min_x \frac{L}{2}\|x - (x_t - \frac{1}{L}\nabla f(x_t))\|^2 + g(x)\,. \tag{3.7}$$

It has been shown in [19] that

> If $\{x_t\}$ is the sequence generated by (3.7), then for any $t \geq 1$, $F(x_t) - F(x^*) \leq \frac{L(f)\|x_0 - x^*\|^2}{2t}$, where x^* is any solution to (3.6).

The key issue in this algorithm is choosing L, which is analogous to selecting a step size in gradient search methods. Without knowing $L(f)$, L can be chosen adaptively at each iteration with backtracking while still enjoying the same convergence properties. The authors further introduce an accelerated version of (3.7) to improve the complexity from $O(1/t)$ to $O(1/t^2)$. We refer interested readers to [19] for additional details. Back to (3.2), the convex function $\frac{1}{2}\|P_\Omega(X) - P_\Omega(Z)\|_F^2$ is continuously differentiable and $\lambda\|X\|_*$ is

continuous and convex. In order to apply this algorithm, the key step is solving (3.7)

$$X_{t+1} = \arg\min_X \frac{L}{2}\|X - (X_t + \frac{1}{L}(P_\Omega(Y) - P_\Omega(X_t)))\|_F^2 + \lambda\|X\|_* .$$

Define the soft-thresholding operator S_λ,

$$S_\lambda(X) = U S_\lambda(\Sigma) V^T, \quad S_\lambda(\Sigma) = \mathrm{diag}(\{(\sigma_i - \lambda)_+\}),$$

where $X = U\Sigma V^T$ is the SVD of X and $\Sigma = \mathrm{diag}(\sigma_1, \ldots, \sigma_r)$. A crucial point is that X_{t+1} can be obtained by soft-thresholding (see, for example, [18])

$$X_{t+1} = S_{\lambda/L}(X_t + \frac{1}{L}(P_\Omega(Y) - P_\Omega(X_t))) . \tag{3.8}$$

A similar iterative formula appears in [20], but derived from a different perspective. The authors use a fixed point iterative method inspired by [22]. The argument is as follows: X^* is a solution to (3.2) if and only if

$$0 \in \lambda\partial\|X^*\| + P_\Omega(X^*) - P_\Omega(Y) .$$

Equivalently,

$$0 \in \tau\lambda\partial\|X^*\| + X^* - [X^* - \tau(P_\Omega(X^*) - P_\Omega(Y))] \ (\tau > 0) .$$

Let $Y^* = [X^* - \tau(P_\Omega(Y) - P_\Omega(X^*))]$, then X^* is the optimal solution to

$$\min_X \frac{1}{2}\|X - Y^*\|_F^2 + \tau\lambda\|X\|_* .$$

Similar to (3.8), we then have

$$X^* = S_{\tau\lambda}(X^* - \tau(P_\Omega(Y) - P_\Omega(X^*))) .$$

Thus, X^* is a fixed point of the function $S_{\tau\lambda}(X - \tau(P_\Omega(Y) - P_\Omega(X)))$, and a natural iterative approach is

$$X_{t+1} = S_{\tau\lambda}(X_t - \tau(P_\Omega(Y) - P_\Omega(X_t))) . \tag{3.9}$$

Iterations (3.8) and (3.9) take the same form when $\tau = \frac{1}{L}$. Let A be the matrix mapping the vectorized $\mathrm{Vec}(X)$ to its observed entries, then [20] proves that

> The sequence $\{X_t\}$ generated by (3.9) with $\tau \in (0, 2/\lambda_{\max}(A^T A))$ converges to an optimal solution of (3.2).

The authors in [21] provide an alternative way to tackle (3.2). The idea is to iteratively impute the unobserved entries, and solve the imputed version of (3.2) by soft-thresholding. It turns out this algorithm fits into the general class of minorize-maximize (MM) algorithms [23], an extension of the well-known EM algorithms. Let $F_Y(X) = \frac{1}{2}\|P_\Omega(X) - P_\Omega(Y)\|_F^2 + \lambda\|X\|_*$ and $Q_Y(X;Z) = \frac{1}{2}\|X - P_{\Omega^\perp}(Z) - P_\Omega(Y)\|_F^2 + \lambda\|X\|_*$. Then for any X, we have

$$\begin{aligned} Q_Y(X;Z) &= \frac{1}{2}\|P_\Omega(X) - P_\Omega(Y) + P_{\Omega^\perp}(X) - P_{\Omega^\perp}(Z)\|_F^2 + \lambda\|X\|_* \\ &= \frac{1}{2}\|P_\Omega(X) - P_\Omega(Y)\|_F^2 + \frac{1}{2}\|P_{\Omega^\perp}(X) - P_{\Omega^\perp}(Z)\|_F^2 + \lambda\|X\|_* \\ &\geq F_Y(X) \end{aligned}$$

and $Q_Y(Z;Z) = F_Y(Z)$. Thus, at each iteration, the updating rule is as follows:

$$\begin{aligned} X_{t+1} &= \arg\min_X \; Q_Y(X;X_t) \\ &= S_\lambda(P_\Omega(Y) + P_{\Omega^\perp}(X_t)) \\ &= S_\lambda(X_t + P_\Omega(Y) - P_\Omega(X_t))\,. \end{aligned} \tag{3.10}$$

Since $X_t + P_\Omega(Y) - P_\Omega(X_t) = P_\Omega(Y) + P_{\Omega^\perp}(X_t)$, at each step, we impute the unobserved entries with the corresponding entries from the estimate in the previous step. Iterations (3.8), (3.9) and (3.10) are all equal when $\tau = 1$ and $L = 1$. The nice thing about procedure (3.10) is that $\{F_Y(X_t)\}$ is a non-increasing sequence since

$$F_Y(X_{t+1}) \leq Q_Y(X_t;X_t) = F_Y(X_t)\,.$$

It is further shown in [21] that

> The sequence $\{X_t\}$ generated by (3.10) converges to X^*, an optimal solution of (3.2). Besides, $F_Y(X_t) - F_Y(X^*) \leq \frac{2\|X_0 - X^*\|_F^2}{t+1}$, where X_0 is an initial value.

To make the algorithm scalable to large matrices, the authors present an efficient way to speed up the SVD calculation in (3.10). Note that $P_\Omega(Y) - P_\Omega(X_t)$ is a sparse matrix, and X_t is a low-rank matrix because of the soft-thresholding operation. This "sparse+low rank" matrix structure can remarkably reduce the computational cost of matrix multiplication. The SVD calculation is then performed by power methods, whose building blocks are matrix-vector multiplications.

The authors in [24] solve (3.2) by considering the equivalent problem

$$\min_{\|X\|_* \leq \frac{\mu}{2}} \quad \frac{1}{2}\|P_\Omega(X) - P_\Omega(Y)\|_F^2\,. \tag{3.11}$$

Equivalence means that for any λ and any optimal solution X^* of (3.2), there exists a $\mu \geq 0$ such that X^* is also a solution to (3.11), and vice versa. By (3.3), it follows that (3.11) is also equivalent to solving

$$\min_{\substack{Z \in \mathcal{R}^{(n+m)\times(n+m)} \\ Z \succeq 0, \mathrm{tr}(Z)=\mu}} \frac{1}{2}\|P_\Omega(X) - P_\Omega(Y)\|_F^2 \,, \tag{3.12}$$

where $Z = \begin{bmatrix} A & X \\ X^T & B \end{bmatrix}$. Under this formulation, [24] follows a simple gradient descent type algorithm proposed in [25]. The algorithm guarantees ε-small primal-dual error after at most $O(\frac{1}{\varepsilon})$ iterations. At each iteration, the main cost is in calculating an approximate largest eigenvector of a given matrix.

4. Matrix Completion via Non-convex Programming.

4.1. *A Family of Non-convex Penalties.* We consider the matrix completion problem in the noisy setting

$$P_\Omega(Y) = P_\Omega(M) + P_\Omega(Z) \,. \tag{4.1}$$

In section 1 we discussed that (4.1) can be treated as a high-dimensional linear regression model; completing the matrix is equivalent to estimating the coefficient parameter $\beta \in \mathcal{R}^{mn}$. As the ℓ_1-penalty induces bias in estimation in the sparse setting, the nuclear norm also biases the estimation of singular values, and thus of low-rank matrices. Motivated by [26], we propose solving the non-convex optimization problem

$$\min_X \quad \{G_{\lambda,\gamma}(X;Y) = \frac{1}{2}\|P_\Omega(X) - P_\Omega(Y)\|_F^2 + \lambda \sum_i P(\sigma_i;\lambda,\gamma)\} \,, \tag{4.2}$$

where $\lambda P(\sigma_i;\lambda,\gamma)$ is the MC+ family of penalties [5] defined by

$$\lambda P(t;\lambda,\gamma) = \lambda(|t| - \frac{t^2}{2\lambda\gamma})I(|t| < \lambda\gamma) + \frac{\lambda^2\gamma}{2} I(|t| \geq \lambda\gamma) \,. \tag{4.3}$$

The reason why we choose this family of non-convex penalties is based on the following one-dimensional problem

$$S_\gamma(\bar{\beta},\lambda) = \arg\min_\beta \ \frac{1}{2}(\beta - \bar{\beta})^2 + \lambda P(|\beta|;\lambda,\gamma) \,. \tag{4.4}$$

In this simplest scenario, the MC+ penalized estimator has an explicit form when $\gamma > 1$:

$$S_\gamma(\bar{\beta},\lambda) = \begin{cases} 0 & \text{if } |\bar{\beta}| \leq \lambda \\ \mathrm{sign}(\bar{\beta})(\frac{|\bar{\beta}|-\lambda}{1-1/\gamma}) & \text{if } \lambda < |\bar{\beta}| \leq \lambda\gamma \\ \bar{\beta} & \text{if } |\bar{\beta}| > \lambda\gamma \,. \end{cases} \tag{4.5}$$

We list a few desirable properties of the univariate minimizer $S_\gamma(\bar{\beta}, \lambda)$ which are well explained in [26]:

1. Criterion (4.4) is strictly convex and $S_\gamma(\bar{\beta}, \lambda)$ is unique.
2. $S_\gamma(\bar{\beta}, \lambda)$ is a continuous function of $\bar{\beta}$.
3. $S_\gamma(\bar{\beta}, \lambda)$ is a continuous function of γ on $(1, \infty)$.
4. $S_\gamma(\bar{\beta}, \lambda)$ bridges the gap between soft-thresholding (ℓ_1) and hard-thresholding (ℓ_0). In particular, as $\gamma \to +\infty, S_\gamma(\bar{\beta}, \lambda) \to \text{sign}(\bar{\beta})(|\bar{\beta}| - \lambda)_+$; and as $\gamma \to 1+, S_\gamma(\bar{\beta}, \lambda) \to \bar{\beta} I(|\bar{\beta}| \geq \lambda)$.

4.2. *Fast First Order Algorithm.* We follow the general procedure in [21] to solve (4.2). At each iteration, we compute

$$X_{t+1} = \arg\min_X \ \frac{1}{2}\|X - P_\Omega(Y) - P_{\Omega^\perp}(X_t)\|_F^2 + \lambda \sum_i P(\sigma_i; \lambda, \gamma)\,, \tag{4.6}$$

where the $\{\sigma_i\}$ are the singular values of X. The following proposition provides a closed form expression for X_{t+1}.

PROPOSITION 4.1. *In (4.6), for any $\gamma > 1, X_{t+1}$ is unique. Define the spectral mapping $\mathcal{S}_{\lambda,\gamma} : \mathcal{R}^{m\times n} \to \mathcal{R}^{m\times n}$ by $\mathcal{S}_{\lambda,\gamma}(X) = U\mathcal{S}_{\lambda,\gamma}(\Sigma)V^T$, with $\mathcal{S}_{\lambda,\gamma}(\Sigma) = diag(\{S_\gamma(\sigma_i, \lambda)\})$, where $X = U\Sigma V^T$ is the SVD of X. Then X_{t+1} has the closed form expression*

$$X_{t+1} = \mathcal{S}_{\lambda,\gamma}(P_\Omega(Y) + P_{\Omega^\perp}(X_t)). \tag{4.7}$$

PROOF. The proof is a direct application of von Neumann's trace inequality (A.1) given in Appendix A. Let $W = P_\Omega(Y) + P_{\Omega^\perp}(X_t)$, the inequality implies

$$\begin{aligned}
\frac{1}{2}\|X - W\|_F^2 + \lambda \sum_i P(\sigma_i; \lambda, \gamma) &= \frac{1}{2}\Big\{\text{tr}(X^T X) + \text{tr}(W^T W) - 2\text{tr}(X^T W)\Big\} \\
&\quad + \lambda \sum_i P(\sigma_i; \lambda, \gamma) \\
&\overset{(a)}{\geq} \frac{1}{2}\Big\{\sum_i \sigma_i^2 + \sum_i \sigma_i(W)^2 - 2\sum_i \sigma_i \sigma_i(W)\Big\} \\
&\quad + \lambda \sum_i P(\sigma_i; \lambda, \gamma) \\
&= \sum_i \Big\{\frac{1}{2}(\sigma_i - \sigma_i(W))^2 + \lambda P(\sigma_i; \lambda, \gamma)\Big\}\,.
\end{aligned}$$

The sum on the right-hand side is separable. By (4.5), there exists a unique $\sigma_i = S_\gamma(\sigma_i(W), \lambda)$ minimizing each term. To achieve the equality in (a), X and W should share the same singular vectors. □

PROPOSITION 4.2. *The sequence $\{G_{\lambda,\gamma}(X_t); Y\}$ is non-increasing, where $\{X_t\}$ is generated by iteration (4.7). Further, any local optimum of (4.2) is a fixed point of $\mathcal{S}_{\lambda,\gamma}(P_\Omega(Y) + P_{\Omega^\perp}(X))$, as a function of X.*

PROOF.

$$\begin{aligned} G_{\lambda,\gamma}(X_{t+1}; Y) &= \arg\min_X \frac{1}{2}\|X - P_\Omega(Y) - P_{\Omega^\perp}(X_t)\|_F^2 + \lambda \sum_i P(\sigma_i; \lambda, \gamma) \\ &\le \frac{1}{2}\|X_t - P_\Omega(Y) - P_{\Omega^\perp}(X_t)\|_F^2 + \lambda \sum_i P(\sigma_i(X_t); \lambda, \gamma) \\ &= \frac{1}{2}\|P_\Omega(X_t) - P_\Omega(Y)\|_F^2 + \lambda \sum_i P(\sigma_i(X_t); \lambda, \gamma) \\ &= G_{\lambda,\gamma}(X_t; Y)\,. \end{aligned}$$

Let $H(X) = \frac{1}{2}\|X - P_\Omega(Y) - P_{\Omega^\perp}(X_*)\|_F^2 + \lambda \sum_i P(\sigma_i; \lambda, \gamma)$, where X_* is a local optimum of $G_{\lambda,\gamma}(X; Y)$. Then X_* is a fixed point if and only if

$$X_* = \arg\min_X \; H(X)\,.$$

This is true since, for any $X \in \{X : \|X - X_*\|_F \le r\}$ such that $G_{\lambda,\gamma}(X_*; Y) \le G_{\lambda,\gamma}(X; Y)$, we have

$$\begin{aligned} H(X) &\ge \frac{1}{2}\|P_\Omega(X) - P_\Omega(Y)\|_F^2 + \lambda \sum_i P(\sigma_i(X); \lambda, \gamma) \\ &= G_{\lambda,\gamma}(X; Y) \ge G_{\lambda,\gamma}(X_*; Y) \\ &= \frac{1}{2}\|X_* - P_\Omega(Y) - P_{\Omega^\perp}(X_*)\|_F^2 + \lambda \sum_i P(\sigma_i(X_*); \lambda, \gamma) \\ &= H(X_*)\,. \end{aligned}$$

□

Thus, X_* is also a local optimum of $H(X)$. If the mapping $H(\cdot)$ is convex, then since H has only one global optimum, the local optimum X_* must be the unique solution. It remains to show $H(X)$ is a convex function of X. Without loss of generality, we prove $\tilde{H}(X) = \frac{1}{2}\|X - W\|_F^2 + \lambda \sum_i P(\sigma_i(X); \lambda, \gamma)$ is convex. Note that

$$\tilde{H}(X) = -\mathrm{tr}(W^T X) + \sum_i \sigma_i(W)\sigma_i + \sum_i \Big\{\frac{1}{2}(\sigma_i - \sigma_i(W))^2 + \lambda P(\sigma_i(X); \lambda, \gamma)\Big\},$$

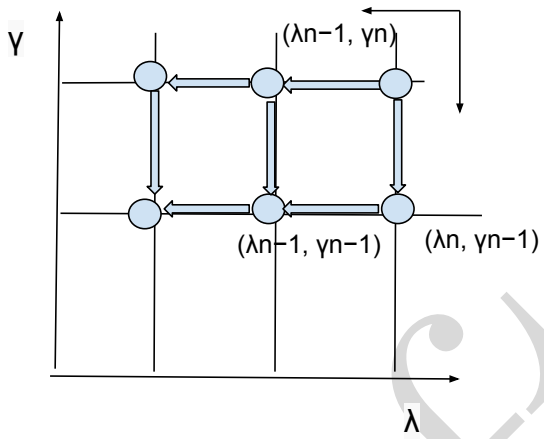

FIG 1. *Grid search pattern.*

where $-\mathrm{tr}(W^T X)$ is linear and $\sum_i \sigma_i(W)\sigma_i$ is a unitarily invariant matrix norm [28], hence both convex. It is not hard to verify the third term is a convex function of the singular values $\sigma_i(X)$. Since it is a unitarily invariant matrix function, it is convex [29].

4.3. *Computing the Solution Path.* Note that there are two tuning parameters (λ, γ) in (4.2), where $\lambda > 0$ and $\gamma \in (1, +\infty)$. We need to compute a series of solutions for different pairs of (λ, γ). We search over a grid (λ_i, γ_j) $(1 \le i \le n, 1 \le j \le n)$, where $0 < \lambda_1 < \cdots < \lambda_n$ and $0 < \gamma_1 < \cdots < \gamma_{n-1} < \gamma_n = +\infty$. Starting by calculating the solution for $(\lambda_{n-1}, +\infty)$, using the solution from the previous pair $(\lambda_n, +\infty)$ as warm-start, we then compute the solution for $(\lambda_{n-1}, \gamma_{n-1})$ based on the solutions from neighbors $(\lambda_n, \gamma_{n-1})$ and $(\lambda_{n-1}, \gamma_n)$ in the grid. The solution which gives a lower value of the objective function is chosen as the initial warm-started value. Figure 1 illustrates the search pattern along the grid. We are now in position to present Algorithm 1 for computing the entire solution path $\{(\lambda_i, \gamma_j)\}$.

4.4. *Computing the Solution Path with Recalibration.* The key issue in non-convex penalized optimization problems is that there may exist local optima. Any algorithm "stuck" on a local solution may compromise the final performance. To avoid bad local optima, Algorithm 1 uses a warm-start strategy to seek good initial values. The rationale is that the solution changes continuously as the tuning parameters vary. A good solution from a previous pair is likely to help the algorithm find a good solution for the

Algorithm 1 Non-convex Impute

1: Input: a search grid: $\lambda_1 < \cdots < \lambda_n, \gamma_1 < \cdots < \gamma_n = +\infty$. Tolerance ε.
2: Use Soft-Impute algorithm in [21] to compute solutions $\hat{X}_{\lambda_i,+\infty}$ for parameter values $(\lambda_i, +\infty)$ $(1 \le i \le n)$.
3: Do for $\gamma_1 < \gamma_2 < \cdots < \gamma_{n-1}$:
  Do for $\lambda_1 < \lambda_2 < \cdots < \lambda_n$:
    (a) Initialize $X^{old} = \arg\min_{X=\hat{X}_{\lambda_{n-1},\gamma_n} \text{ or } \hat{X}_{\lambda_n,\gamma_{n-1}}} G_{\lambda_{n-1},\gamma_{n-1}}(X;Y)$.
    (b) Repeat:
      i. Compute $X^{new} \leftarrow \mathcal{S}_{\lambda_{n-1},\gamma_{n-1}}(P_\Omega(Y) + P_{\Omega^\perp}(X^{old}))$.
      ii. If $\frac{\|X^{new}-X^{old}\|_F^2}{\|X^{old}\|_F^2} < \varepsilon$, exit.
      iii. Assign $X^{old} \leftarrow X^{new}$.
    (c). Assign $\hat{X}_{\lambda_{n-1},\gamma_{n-1}} \leftarrow X^{new}$.
4: Output $\hat{X}_{\lambda_i,\gamma_j}$ $(1 \le i \le n, 1 \le j \le n)$.

current one. From Figure 1, we see the search pattern is naturally horizontal and vertical. But nothing stops us from conducting a curved search. Would there exist a feasible curved search for obtaining better solutions of problem (4.2)? We borrow a nice recalibration approach from [26]. Before going into any details, we need a measure of model complexity. To this end, we review Stein's unbiased risk estimation (SURE) theory [27, 31].

PROPOSITION 4.3. *Suppose that* $Y_{ij} \overset{i.i.d.}{\sim} N(X_{ij}, \tau^2)$. *Consider an estimator* $\hat{X}$ *of the form* $\hat{X} = Y + g(Y)$, *where* $g_{ij} : \mathcal{R}^{m\times n} \to \mathcal{R}$ *is weakly differentiable with respect to* Y_{ij} *and*

$$E\Big\{|Y_{ij}g_{ij}(Y)| + |\frac{\partial}{\partial Y_{ij}}g_{ij}(Y)|\Big\} < \infty, \text{ for } 1 \le i \le m, 1 \le j \le n\,.$$

Then

$$E\|\hat{X} - X\|_F^2 = E\Big\{-\tau^2 mn + 2\tau^2 \sum_{i,j} \frac{\partial \hat{X}_{ij}}{\partial Y_{ij}} + \|g(Y)\|_F^2\Big\}\,.$$

Thus, $-\tau^2 mn + 2\tau^2 \sum_{i,j} \frac{\partial \hat{X}_{ij}}{\partial Y_{ij}} + \|g(Y)\|_F^2$ is an unbiased estimator for risk estimation. The authors in [32] define the "degrees-of-freedom" for $\hat{X}$ as $\sum_{i,j} \mathrm{cov}(\hat{X}_{ij}, Y_{ij})/\tau^2$. When the conditions of Proposition 4.3 hold, [30] shows that $E\{\sum_{i,j} \frac{\partial \hat{X}_{ij}}{\partial Y_{ij}}\} = 2\sum_{i,j} \mathrm{cov}(\hat{X}_{ij}, Y_{ij})/\tau^2$. SURE theory provides an insight into how model complexity affects prediction error, with $\|g(Y)\|_F^2$

being the training error and $2\tau^2 \sum_{i,j} \frac{\partial \hat{X}_{ij}}{\partial Y_{ij}}$ measuring model complexity. When the model grows, $\|g(Y)\|_F^2$ typically decreases while $2\tau^2 \sum_{i,j} \frac{\partial \hat{X}_{ij}}{\partial Y_{ij}}$ tends to increase.

The basic idea of recalibration consists in calculating a reparametrization $(\lambda_S(\lambda,\gamma),\gamma)$ of (λ,γ) such that for any fixed λ, the degrees of freedom for the fitting procedure $\hat{X}_{\lambda_S(\lambda,\gamma),\gamma} = \arg\min_X G_{\lambda_S(\lambda,\gamma),\gamma}(X;Y)$ remain constant as γ varies. Each mapping $(\lambda_S(\lambda,\gamma),\gamma)$ characterizes a curve for a fixed λ. The solution path will then be computed along these curves. Under this model complexity driven coordinates, the solution path is expected to be more smooth and able to avoid bad local optima. Indeed, in the sparse regression setting, numerical experiments in [26] show the improvements over the search strategy discussed in Section 4.3.

To apply recalibration in solving (4.2), the key step is to calculate the degrees of freedom. Generally, this is impossible to accomplish because not much information about the global optimum $\hat{X}_{\lambda_S(\lambda,\gamma),\gamma}$ is known. This is, after all, what we want to compute in the first place. As in [26], we consider a simplified fitting procedure

$$\tilde{X}_{\lambda,\gamma} = \arg\min_X \frac{1}{2}\|X - \tilde{Y}\|_F^2 + \lambda \sum_i P(\sigma_i;\lambda,\gamma) = \mathcal{S}_{\lambda,\gamma}(\tilde{Y})\,,$$

where $\tilde{Y} \overset{i.i.d.}{\sim} N(0,1)$. Note that this type of fitting is the building block of Algorithm 1. Denote the degrees of freedom for $\tilde{X}_{\lambda,\gamma}$ as $df(\tilde{X}_{\lambda,\gamma})$. To show that $\tilde{X}_{\lambda,\gamma}$ satisfies the conditions in the SURE theory, we first prove $\mathcal{S}_{\lambda,\gamma}(\tilde{Y})$ is Lipschitz continuous in the following.

PROPOSITION 4.4. *For any $Y_1, Y_2 \in \mathcal{R}^{m\times n}$, the thresholding operator $\mathcal{S}_{\lambda,\gamma}(\cdot)$ obeys*

$$\|\mathcal{S}_{\lambda,\gamma}(Y_1) - \mathcal{S}_{\lambda,\gamma}(Y_2)\|_F \le \frac{\gamma}{\gamma-1}\|Y_1 - Y_2\|_F\,. \tag{4.8}$$

The equality holds if and only if $\mathcal{S}_{\lambda,\gamma}(Y_1) - \mathcal{S}_{\lambda,\gamma}(Y_2) = \frac{\gamma}{\gamma-1}(Y_1 - Y_2)$.

PROOF. Let $Y_1 = U_1\Sigma V_1^T, Y_2 = U_2\Phi V_2^T$ be the SVD of Y_1 and Y_2, and $\{\sigma_i\},\{\phi_i\}$ be the singular values of Y_1, Y_2, respectively. We use $\mathcal{S}(\cdot)$ to rep-

resent $\mathcal{S}_{\lambda,\gamma}(\cdot)$ and $S(\cdot)$ to denote $S_\gamma(\cdot;\lambda)$. Choosing $L = [\frac{\gamma}{\gamma-1}]^2$ yields

$$\begin{aligned}
& L\|Y_1 - Y_2\|_F^2 - \|\mathcal{S}(Y_1) - \mathcal{S}(Y_2)\|_F^2 \\
= \;& \sum_i \{L(\sigma_i^2 + \phi_i^2) - S(\sigma_i)^2 - S(\phi_i)^2\} - 2L\mathrm{tr}(Y_1^T Y_2) + 2\mathrm{tr}(\mathcal{S}(Y_1)^T \mathcal{S}(Y_2)) \\
= \;& \sum_i \{L(\sigma_i^2 + \phi_i^2) - S(\sigma_i)^2 - S(\phi_i)^2\} \\
& -2\mathrm{tr}[(\sqrt{L}Y_1 - \mathcal{S}(Y_1))^T(\sqrt{L}Y_2 - \mathcal{S}(Y_2))] - 2\mathrm{tr}[(\sqrt{L}Y_1 - \mathcal{S}(Y_1))^T \mathcal{S}(Y_2)] \\
& -2\mathrm{tr}[\mathcal{S}(Y_1)^T(\sqrt{L}Y_2 - \mathcal{S}(Y_2))] \\
\overset{(b)}{\geq} \;& \{\sum_i L(\sigma_i^2 + \phi_i^2) - S(\sigma_i)^2 - S(\phi_i)^2\} \\
& -2\{\sum_i (\sqrt{L}\sigma_i - S(\sigma_i))(\sqrt{L}\phi_i - S(\phi_i))\} - 2\{\sum_i (\sqrt{L}\sigma_i - S(\sigma_i))S(\phi_i)\} \\
& -2\{\sum_i S(\sigma_i)(\sqrt{L}\phi_i - S(\phi_i))\} \\
= \;& \sum_i [L(\sigma_i - \phi_i)^2 - (S(\sigma_i) - S(\phi_i))^2] \\
\overset{(c)}{\geq} \;& 0\,,
\end{aligned}$$

where (b) holds by applying von Neumann's trace inequality (A.1) and (c) follows from (4.5). Equality holds if and only if equality holds throughout (b) and (c). Equality can be achieved in (b) if and only if $U_1 = U_2$ and $V_1 = V_2$. Equality holds in (c) if and only if $\lambda \leq \sigma_i, \phi_i \leq \lambda\gamma$, implying $\frac{S(\sigma_i) - S(\phi_i)}{\sigma_i - \phi_i} = \frac{\gamma}{\gamma-1}$. □

Consider now the set of matrices

$$\begin{aligned}
\mathcal{F} = \{Y \in \mathcal{R}^{m\times n} : & \; Y \text{ is simple, of full rank, and no singular values} \\
& \text{are equal to } \lambda \text{ and } \lambda\gamma\}.
\end{aligned}$$

Since the singular values are continuous matrix functions, $\mathcal{F}$ is an open set. It is also not hard to see that the complement set $\mathcal{F}^c$ has Lebesgue measure zero under the noisy setting (4.1). It is shown in [27] that $\mathcal{S}_{\lambda,\gamma}(Y)$ is differentiable over $\mathcal{F}$, and

$$\begin{aligned}
\sum_{i,j} \frac{\partial[\mathcal{S}_{\lambda,\gamma}(Y)]_{ij}}{\partial Y_{ij}} \;=\; & \sum_i \Big(S'_\gamma(\sigma_i;\lambda) + |m-n|\frac{S_\gamma(\sigma_i;\lambda)}{\sigma_i}\Big) \\
& +2\sum_{i\neq j} \frac{\sigma_i S_\gamma(\sigma_i;\lambda)}{\sigma_i^2 - \sigma_j^2}\,.
\end{aligned} \tag{4.9}$$

Thus $[\mathcal{S}_{\lambda,\gamma}(Y)]_{ij}$ is weakly differentiable. Furthermore, Proposition 4.4 implies

$$\begin{aligned} E\{\Big|\frac{\partial[\mathcal{S}_{\lambda,\gamma}(Y)]_{ij}}{\partial Y_{ij}}\Big|\} &\leq \frac{\gamma}{\gamma-1} < \infty \\ E\{Y_{ij}[\mathcal{S}_{\lambda,\gamma}(Y)]_{ij}\} &\leq E^{1/2}[Y_{ij}^2]E^{1/2}[\mathcal{S}^2_{\lambda,\gamma}(Y)]_{ij}] \\ &\leq C\, E^{1/2}\|\mathcal{S}_{\lambda,\gamma}(Y) - \mathcal{S}_{\lambda,\gamma}(0)\|_F^2 \\ &\leq C\,\frac{\gamma}{\gamma-1}E^{1/2}\|Y\|_F^2 < \infty\,. \end{aligned}$$

Thus, we have shown the following.

PROPOSITION 4.5. *The degrees of freedom of the fitting procedure $\tilde{X}_{\lambda,\gamma}$ obey*

$$df(\tilde{X}_{\lambda,\gamma}) = E\Big\{\sum_i \Big(S'_\gamma(\sigma_i;\lambda) + |m-n|\frac{S_\gamma(\sigma_i;\lambda)}{\sigma_i}\Big) + 2\sum_{i\neq j}\frac{\sigma_i S_\gamma(\sigma_i;\lambda)}{\sigma_i^2-\sigma_j^2}\Big\},$$

where the $\{\sigma_i\}$ are the singular values of $\tilde{Y}$, with $\tilde{Y}_{ij} \overset{i.i.d.}{\sim} N(0,1)$.

The degrees of freedom calculation in Proposition 4.5 involves a high-dimensional integral, which itself is a difficult problem. Instead, we use the Marchenko-Pastur law (A.2) to approximate it. Adopting the notation in Appendix A, it is not hard to verify that

$$\begin{aligned} &\sum_i \Big(S'_\gamma(\sigma_i;\lambda) + |m-n|\frac{S_\gamma(\sigma_i;\lambda)}{\sigma_i}\Big) + 2\sum_{i\neq j}\frac{\sigma_i S_\gamma(\sigma_i;\lambda)}{\sigma_i^2-\sigma_j^2} = \\ &nE_{\mu_n}\Big\{S'_\gamma(\sqrt{x};\lambda) + |m-n|\frac{S_\gamma(\sqrt{x};\lambda)}{\sqrt{x}}\Big\} + 2n^2E_{\mu_n}\Big\{\frac{\sqrt{x}S_\gamma(\sqrt{x};\lambda)}{x-y}I(x\neq y)\Big\}, \end{aligned}$$

where $x, y \overset{i.i.d.}{\sim} \mu_n$. Let $\frac{n}{m} = \alpha$, by the Marchenko-Pastur law, we have

$$\begin{aligned} df(\tilde{X}_{\lambda,\gamma}) \;\approx\; & nE_\mu\Big\{S'_\gamma(\sqrt{mx};\lambda) + |m-n|\frac{S_\gamma(\sqrt{mx};\lambda)}{\sqrt{mx}}\Big\} \\ & +2n^2E_\mu\Big\{\frac{\sqrt{mx}S_\gamma(\sqrt{mx};\lambda)}{mx-my}I(x\neq y)\Big\}, \end{aligned} \tag{4.10}$$

where $x, y \overset{i.i.d.}{\sim} \mu$ with parameter α. Note that (4.10) only requires two-dimensional integrals which can be approximated by Monte Carlo methods. We now present Algorithm 2 with recalibration.

Algorithm 2 Non-convex Impute with Recalibration

1: Input: a search grid: $\lambda_1 < \cdots < \lambda_n, \gamma_1 < \cdots < \gamma_n = +\infty$. Tolerance ε.
2: Use Soft-Impute algorithm in [21] to compute the solutions $\hat{X}_{\lambda_i,+\infty}$ for parameter values $(\lambda_i, +\infty)$ $(1 \le i \le n)$.
3: Generate i.i.d. samples from the Marchenko-Pastur law with parameter $\alpha = n/m$.
4: Do for $\gamma_1 < \gamma_2 < \cdots < \gamma_{n-1}$:
Do for $\lambda_1 < \lambda_2 < \cdots < \lambda_n$:
(a) Use Monte Carlo methods to calculate the degrees of freedom df_i for $(\lambda_i, \gamma_{j-1})$ based on (4.10).
(b) Compute $\lambda_S(\lambda_i, \gamma_j)$ such that the degrees of freedom for $(\lambda_S(\lambda_i, \gamma_j), \gamma_j)$ remains df_i.
(c) Initialize $X^{old} = \hat{X}_{\lambda_S(\lambda_i,\gamma_{j-1}),\gamma_{j-1}}$.
(d) Repeat:
i. Compute $X^{new} \leftarrow \mathcal{S}_{\lambda_S(\lambda_i,\gamma_j),\gamma_j}(P_\Omega(Y) + P_{\Omega^\perp}(X^{old}))$.
ii. If $\frac{\|X^{new} - X^{old}\|_F^2}{\|X^{old}\|_F^2} < \varepsilon$, exit.
iii. Assign $X^{old} \leftarrow X^{new}$.
(e). Assign $\hat{X}_{\lambda_S(\lambda_i,\gamma_j),\gamma_j} \leftarrow X^{new}$.
5: Output $\hat{X}_{\lambda_S(\lambda_i,\gamma_j),\gamma_j}$ $(1 \le i \le N, 1 \le j \le N)$.

5. Numerical Experiments. In this section we show the advantages of using non-convex regularizers in the noisy matrix completion problem (4.1) via simulations and two real data sets. For simplicity, we only analyze solutions to (4.2) as provided by Algorithm 1, leaving the recalibration approach of Algorithm 2 for future work. In addition to the MC+ penalty (4.3), we explore matrix completion solutions via the non-convex SCAD penalty introduced in [4].

5.1. *Simulation Settings.* We analyze two different simulation settings where, for the true underlying matrix $M = U\Sigma V^T$, we vary both the structure of the left and right singular vectors in U and V, as well as the "spikiness" in the singular values $\Sigma = \mathrm{diag}(\sigma_1, \ldots, \sigma_r)$. For the observed matrix $Y = M + Z$, we also consider different variance terms τ^2 for the Gaussian noise Z.

In our first simulation setting, we use the model $Y_{m\times n} = U_{m\times r}\Sigma_{r\times r}V^T_{r\times n} + Z_{m\times n}$, where U and V are matrices generated from the Random Orthogonal Model [1]. Recall that for this model, exact matrix completion is guaranteed as long as $|\Omega| \ge Crm\log^4 m$. The singular values of M are randomly selected as either $\sigma_1, \ldots, \sigma_r \overset{i.i.d.}{\sim} \mathrm{Uniform}(20, 50)$ or $\sigma_1, \ldots, \sigma_r \overset{i.i.d.}{\sim} \mathrm{Pareto}(1, 1.0001)$. With the noise term Z having variance $\tau^2 \in \{1/25, 1/100\}$, the resulting

matrix Y should have low "spikiness" (as defined in (2.8)) in the Uniform case and the Pareto case when there is a single dominant sampled singular value, and higher "spikiness" when the observations sampled from the Pareto are clustered around one.

We also consider the model $Y_{m\times n} = U_{m\times r}\Sigma_{r\times r}V^T_{r\times n} + Z_{m\times n}$ in our second simulation setting, but we now select matrices $U = (u_1, \cdots, u_r)$ and $V = (v_1, \cdots, v_r)$ which do not satisfy the incoherence conditions in [1, 9]. Specifically, we select the first five columns of U and V to be of the form

$$u_i = v_i = \sum_{j=1}^{4}(-1)^{j-1}\,\frac{1}{2}e_{j+4(i-1)} \quad \text{for } i = 1, \ldots, 5\,,$$

where $(e_i)_{i=1}^m$ is the standard canonical basis. The remaining columns in U, V are generated from the Random Orthogonal Model. The singular values are sampled in the same way as in the first setting. For this model, successful matrix completion is not guaranteed even for the exact case with the nuclear norm (LASSO) regularizer. We would like to investigate the performance of the non-convex regularizers based on the SCAD and MC+ penalties in this challenging scenario. A superior performance of these penalties over the usual nuclear norm might signal that non-convex penalties are able to weaken the strong incoherence conditions required for successful nuclear norm recovery. For the same choice of noise variance τ^2, given the structure of the factors U and V, the spikiness $\alpha_{sp}(Y)$ should be much larger in this setting.

The standardized test error for both models is defined as

$$\text{Test Error} = \frac{\|P_{\Omega^\perp}(U\Sigma V^T - \hat{Y})\|_F^2}{\|P_{\Omega^\perp}(U\Sigma V^T)\|_F^2}\,,$$

where a value greater than one indicates that the computed estimator $\hat{Y}$ does a worse job at estimating M than the zero solution. Tables 1–2 show median and robust standard deviation (RSD = IQR/1.34) over 20 repetitions for all three of the regularizers mentioned above in both problem instances. In all examples, we fix $(m, n, r) = (1000, 500, 15)$. The set Ω of observed entries is sampled uniformly at random over the indices of the matrix. We choose a 20×20 grid of (λ, γ) values as follows. We set $\lambda_{20} = \lambda_{\max}(P_\Omega(Y))$, $\lambda_1 = 0.001 \cdot \lambda_{20}$, and construct a sequence of 20 values of λ decreasing from λ_{20} to λ_1 on the log scale. The selected grid of concavity parameters for γ is given by $\{100, 70, 50, 30, 20, 15, 10, 9, 8, 7, 6, 5, 4.5, 4, 3.5, 3, 2.5, 2, 1.5, 1.1\}$, where the last three are trimmed for the SCAD penalty. For each simulated data set, we only record the solution along the grid achieving smallest test

TABLE 1
Comparison of LASSO, SCAD, and MC+ regularized estimations over 20 repetitions from the Random Orthogonal Model. Results are in the form median (robust standard deviation).

Dist. Sing. Val.	$\|\Omega\|/(m \cdot n)$	τ^2	**LASSO**	**SCAD**	**MC+**
Uniform(20, 50)	0.5	1/100	0.053(0.005)	0.025(0.003)	0.025(0.003)
		1/25	0.174(0.014)	0.110(0.014)	0.105(0.011)
	0.2	1/100	0.158(0.015)	0.077(0.010)	0.077(0.011)
		1/25	0.402(0.026)	0.286(0.035)	0.290(0.026)
Pareto(1, 1.0001)	0.5	1/100	0.216(0.447)	0.135(0.406)	0.136(0.375)
		1/25	0.473(0.592)	0.390(0.646)	0.341(0.617)
	0.2	1/100	0.373(0.613)	0.266(0.606)	0.250(0.647)
		1/25	0.788(0.843)	0.694(0.929)	0.605(0.960)

error. Throughout these simulations, if the warm-started solution X^{old} has rank higher than 50 along the (λ, γ) grid, we truncate the smallest singular values so as to deliver solutions with rank at most 50.

5.2. *Observations.* For both simulation settings, models are displayed as increasingly challenging in Tables 1–2, ranging from only 50% of entries missing and small noise, to 80% of entries missing and relatively high noise in a "spiky" matrix setup. From the simplest to the more complex structures, the non-convex regularizers outperform the usual nuclear norm estimator in terms of testing error. The SCAD and MC+ non-convex regularizers have almost identical performance, with MC+ doing slightly better in the more complex Random Orthogonal Models from Table 1.

The improvements that the non-convex programming approach can bring in the Coherent Model of Table 2 are quite marginal, even in the simplest setting of higher percentage of observed entries and smaller noise. This suggests, as pointed out in [10], that the incoherence assumptions are required as a result of the uniform sampling assumption on Ω, and there is not much improvement that non-convex regularizers can achieve under this form of sampling on Ω. Nonetheless, for all settings in Table 2, SCAD and MC+ improve on the results obtained by the usual nuclear norm.

When the singular values of M are sampled from the heavy-tailed Pareto distribution, the standard deviation of the test error increases with respect to the Uniform case. As explained above, when there is a single dominant singular value sampled from the Pareto $(1, 1.0001)$ distribution, the spikiness ratio $\alpha_{sp}(Y)$ will be closer to one, making the matrix completion problem easier [12]. When this was the case throughout our random simulations, the resulting non-convex estimators were able to achieve test errors as small

TABLE 2
Comparison of LASSO, SCAD, and MC+ regularized estimations over 20 repetitions from the Coherent (vis-à-vis the standard basis (e_i)) Model. Results are in the form median (robust standard deviation).

Dist. Sing. Val.	$\|\Omega\|/(m \cdot n)$	τ^2	**LASSO**	**SCAD**	**MC+**
Uniform$(20, 50)$	0.5	1/100	0.266(0.065)	0.225(0.074)	0.225(0.073)
		1/25	0.388(0.066)	0.337(0.066)	0.329(0.060)
	0.2	1/100	0.557(0.062)	0.555(0.061)	0.556(0.063)
		1/25	0.719(0.055)	0.711(0.057)	0.713(0.059)
Pareto$(1, 1.0001)$	0.5	1/100	0.710(0.327)	0.648(0.341)	0.628(0.325)
		1/25	0.863(0.337)	0.860(0.397)	0.847(0.394)
	0.2	1/100	0.967(0.057)	0.933(0.095)	0.931(0.092)
		1/25	0.988(0.023)	0.980(0.087)	0.976(0.094)

as 0.003. Tightly clustered singular values around one made the problem more complicated for the non-convex regularizers in the Pareto setting, thus agreeing with the intuition in sparse regression that non-convex penalized methods perform better in the presence of a few large, sparse signals.

5.3. *Applications to the MovieLens Data Sets.* We now use the real world recommendation system data sets `ml100k` and `ml1m` provided by MovieLens (http://grouplens.org/datasets/movielens/) to compare the usual nuclear norm approach with the SCAD and MC+ regularizers. The data set `ml100k` consists of $100,000$ movie ratings (1–5) from 943 users on $1,682$ movies, whereas `ml1m` includes $1,000,209$ anonymous ratings from $6,040$ users on $3,952$ movies. In both cases, and for all three regularization methods, 90% of the ratings were used for training, the others were used as the test set.

The entries from the estimators in (4.2) were rounded in the natural way to produce integer-valued matrices. Table 3 shows the prediction error (RMSE) obtained from the left out portion of the data. The performance of the SCAD and MC+ regularizers is almost identical in these examples, however, the improvement over the nuclear norm approach is considerable, especially in the smaller data set `ml100k`. These results are encouraging enough, and should stimulate future work on more sophisticated non-convex variants achieving smaller RMSE rates.

6. Summary and Discussions. We gave a brief overview of algorithms and theoretical results in the matrix completion problem. We then proposed a non-convex penalized approach for estimating low-rank matrices. Following the work in [21, 26], we presented a fast algorithm and a recalibartion approach to warm-starts. Numerical results show potential ad-

TABLE 3
Root Mean Squared Error of the LASSO, SCAD, and MC+ regularized estimators on the MovieLens testing data sets.

Data Set	$\|\Omega\|/(m \cdot n)$	**LASSO**	**SCAD**	**MC+**
`ml100k`	0.0567	1.0203	0.9764	0.9771
`ml1m`	0.0377	0.9368	0.9106	0.9127

vantages for non-convex programming methods if the local optima issues can be properly solved.

For this work, there are several open problems left. The convergence analysis of Algorithm 1 has not been studied. Numerical experiments regarding the recalibration approach need to be conducted. For large matrices, scalability is also an important problem to be explored. Furthermore, it is desirable to develop theoretical results to analytically compare with the nuclear norm based approach. Another interesting and difficult direction is to quantify the performance of local optima.

Considering matrix completion as a high-dimensional regression problem, we think a number of ideas from the sparse regression literature can be borrowed to ask and solve interesting questions in the field. For instance, what is the analogy of matrix completion to variable selection in regression? Is it selecting the tangent space? Can we eliminate unimportant subspaces easily as we screen out noises by marginal information in regression? In sparse regression, sparsity patterns and inducing penalties have been extended to some interesting structures like group and hierarchy. Would there exist motivating examples in matrix completion having more interesting structures other than low-rank? If yes, what would be the corresponding inducing penalty? We leave these questions for future research.

APPENDIX A

Theorem 1. (von Neumann's trace inequality). *For any matrices $A, B \in \mathcal{R}^{m\times n}$, let $\sigma_1(A) \geq \sigma_2(A) \geq \ldots \geq 0$ and $\sigma_1(B) \geq \sigma_2(B) \geq \ldots \geq 0$ be the descending singular values of A and B, respectively. Let $A = U_A\Sigma_A V_A^T$ and $B = U_B\Sigma_B V_B^T$ be their singular value decompositions. Then,*

$$|tr(A^T B)| \leq \sum_i \sigma_i(A)\sigma_i(B)\,. \tag{A.1}$$

The equality holds if and only if $U_A = U_B$ and $V_A = V_B$.

Theorem 2. (Marchenko-Pastur law). *Let $X \in \mathcal{R}^{m\times n}$, where X_{ij} are i.i.d. with $E(X_{ij}) = 0, E(X_{ij}^2) = 1$, and $m > n$. Let $\lambda_1 \leq \lambda_2 \leq \cdots \leq \lambda_n$ be*

the eigenvalues of $S_m = \frac{1}{m} X^T X$. *Define the random spectral measure*

$$\mu_n = \frac{1}{n} \sum_{i=1}^{n} \delta_{\lambda_i} .$$

Then, assuming $n/m \to \alpha \in (0, 1]$, *we have*

$$\mu_n(\cdot, \omega) \to \mu \quad a.s.,$$

where μ *is a deterministic measure with density*

$$\frac{d\mu}{dx} = \frac{\sqrt{(\alpha_+ - x)(x - \alpha_-)}}{2\pi\alpha x} I(\alpha_- \leq x \leq \alpha_+). \tag{A.2}$$

Here, $\alpha_+ = (1 + \sqrt{\alpha})^2, \alpha_- = (1 - \sqrt{\alpha})^2$.

REFERENCES

[1] E. Candes and B. Recht. (2009). Exact matrix completion via convex optimization. *Foundations of Computational Mathematics,* **9** 717-772.

[2] ACM SIGKDD and Netflix. *Proceedings of KDD Cup and Workshop, 2007.*

[3] R. Tibshirani. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society, Ser. B,* **58** 267-288.

[4] J. Fan and R. Li. (2001). Variable selection via non concave penalized likelihood and its oracle properties. *Journal of the American Statistical Association,* **96** 1348-1360.

[5] C. Zhang. (2010). Nearly unbiased variable selection under minimax concave penalty. *The Annals of Statistics,* **38** 894-942.

[6] E. Candes and T. Tao. (2007). The Dantzig selector: statistical estimation when p is much larger than n. *The Annals of Statistics,* **35** 23132351.

[7] E. Candes and Y. Plan. (2010). Matrix completion with noise. *Proceedings of the IEEE,* **98**, 925-936.

[8] E. Candes and T. Tao. (2010). The power of convex relaxation: Near-optimal matrix completion. *IEEE Transactions on Information Theory,* **56**, 2053-2080.

[9] B. Recht. (2011). A simpler approach to matrix completion. *Journal of Machine Learning Research,* **12**, 3413-3430.

[10] Y. Chen, S. Bhojanapalli, S. Sanghavi and R. Ward. (2014). Coherent Matrix Completion. *Proceedings of the* 31^{st} *International Conference on Machine Learning, 2014.*

[11] R. Keshavan, A. Montanari and S. Oh. (2010). Matrix completion from noisy entries. *Journal of Machine Learning Research,* **11**, 2057-2078.

[12] S. Negahban and M. Wainwright. (2012). Restricted strong convexity and weighted matrix completion: optimal bounds with noise. *Journal of Machine Learning Research,* **13**, 1665-1697.

[13] S. Negahban, P. Ravikumar, M. Wainwright and B. Yu. (2012). A unified framework for high-dimensional analysis of M-estimators with decomposable regularizers. *Statistical Science,* **27**, 538-557.

[14] E. Candes and Y. Plan. (2010). Tight oracle inequalities for low-rank matrix recovery from a minimal number of noisy random measurements. *IEEE Transactions on Information Theory,* **57**(4), 2342-2359.

[15] S. Negahban and M. Wainwright. (2011). Estimation of (near) low-rank matrices with noise and high-dimensional scaling. *The Annals of Statistics,* **39**, 1069-1097.

[16] A. Rohde and A. Tsybakov. (2011). Estimation of high-dimensional low-rank matrices. *The Annals of Statistics,* **39**, 887-930.

[17] B. Recht, M. Fazel and P. Parrilo. (2010). Guaranteed minimum-rank solutions of linear matrix equations via nuclear norm minimization. *SIAM Review,* **52**, 471-501.

[18] J. Cai, E. Candes and Z. Shen. (2010). A singular value thresholding algorithm for matrix completion. *SIAM Journal on Optimization,* **20**, 1956-1982.

[19] A. Beck and M. Teboulle. (2009). A fast iterative shrinkage-thresholding algorithm for linear inverse problems. *SIAM Journal on Imaging Sciences,* **2**, 183-202.

[20] S. Ma, D. Goldfarb and L. Chen. (2011). Fixed point and Bregman iterative methods for matrix rank minimization. *Math.Program., Ser. A* **128**, 321-353.

[21] R. Mazumder, T. Hastie and R. Tibshirani. (2010). Spectral regularization algorithms for learning large incomplete matrices. *Journal of Machine Learning Research,* **11**, 2287-2322.

[22] E. Hale, W. Yin and Y. Zhang. (2007). A fixed-point continuation method for ell_1-regularized minimization with applications to compressed sensing. Technical Report, CAAM TR07-07.

[23] D. Hunter and R. Li. (2005). Variable selection using MM algorithms. *The Annals of Statistics,* **33**, 1617-1642.

[24] M. Jaggi and M. Sulovsky. (2010). A simple algorithm for nuclear norm regularized problems. *Proceedings of the* 27^{th} *International Conference on Machine Learning, 2010.*

[25] E. Hazan. (2008). Sparse approximation solutions to semidefinite programs. *LATIN*, pp. 306-316.

[26] R. Mazumder, J. Friedman and T. Hastie. (2011). SparseNet: Coordinate descent with nonconvex penalties. *Journal of the American Statistical Association*, **106** ,1125-1138.

[27] E. Candes, C. Sing-Long and J. Trzasko. (2013). Unbiased risk estimates for singular value thresholding and spectral estimators. *IEEE Transactions on Signal Processing,* **61**, 4643-4657.

[28] R. Horn and C. Johnson. (1985). *Matrix Analysis.* Cambridge University Press, Cambridge.

[29] A. Lewis. (1995). The convex analysis of unitarily invariant matrix functions. *Journal of Convex Analysis,* **2**, 173-183.

[30] B. Efron. (2004). The estimation of prediction error: Covariance penalties and cross-validation (with discussion). *J. Amer. Statist. Assoc.* **99**, 619-632.

[31] C. Stein. (1981). Estimation of the mean of a multivariate normal distribution. *The Annals of Statistics,* **9**, 1135-1151.

[32] B. Efron, T. Hastie, I. Jonstone and R. Tibshirani. (2004). Least angle regression. *The Annals of Statistics,* **32**, 407-499.