Matrix completion with nonconvex regularization: spectral operators and scalable algorithms

As Published: https://doi.org/10.1007/s11222-020-09939-5

Publisher: Springer US

Persistent URL: https://hdl.handle.net/1721.1/131497

Version: Author's final manuscript: final author's manuscript post peer review, without publisher's formatting or copy editing

Matrix completion with nonconvex regularization: spectral operators and scalable algorithms

Cite this article as: Rahul Mazumder, Diego Saldana and Haolei Weng, Matrix completion with nonconvex regularization: spectral operators and scalable algorithms, Statistics and Computing https://doi.org/10.1007/s11222-020-09939-5

Statistics and Computing manuscript No.
(will be inserted by the editor)

Matrix Completion with Nonconvex Regularization: Spectral Operators and Scalable Algorithms

Rahul Mazumder · Diego Saldana · Haolei Weng

Received: date / Accepted: date

Abstract In this paper, we study the popularly dubbed matrix completion problem, where the task is to "fill in" the unobserved entries of a matrix from a small subset of observed entries, under the assumption that the underlying matrix is of low-rank. Our contributions herein, enhance our prior work on nuclear norm regularized problems for matrix completion (Mazumder et al., 2010) by incorporating a continuum of nonconvex penalty functions between the convex nuclear norm and nonconvex rank functions. Inspired by SOFT-IMPUTE (Mazumder et al., 2010; Hastie et al., 2016), we propose NC-IMPUTE — an EM-flavored algorithmic framework for computing a family of nonconvex penalized matrix completion problems with warm-starts. We present a systematic study of the associated spectral thresholding operators, which play an important role in the overall algorithm. We study convergence properties of the algorithm. Using structured low-rank SVD computations, we demonstrate the computational scalability of our proposal for problems up to the Netflix size (approximately, a $500,000 \times 20,000$ matrix with 10^8 observed entries). We demonstrate that on a wide range of synthetic and real data instances, our proposed nonconvex regularization framework leads to low-rank solutions with better predictive performance when compared to those obtained from nuclear norm problems. Implementations of algorithms proposed herein, written in the `R` language, are made available on `github`.

Keywords Matrix completion · Low rank · Spectral nonconvex penalties · MC+ penalty · Optimization · Degrees-of-freedom

Rahul Mazumder
MIT Sloan School of Management, Massachusetts Institute of Technology, Cambridge, MA 02142, USA
E-mail: rahulmaz@mit.edu

Diego Saldana
Data Scientist, Tapad, New York, NY 10010, USA
E-mail: diegofrasal@gmail.com

Haolei Weng
Department of Statistics and Probability, Michigan State University, East Lansing, MI 48824, USA
E-mail: wenghaol@msu.edu

1 Introduction

In several problems of contemporary interest, arising for instance, in recommender system applications, for example, the Netflix Prize competition (SIGKDD and Netflix, 2007), observed data is in the form of a large sparse matrix, Y_{ij}, $(i,j) \in \Omega$, where $\Omega \subset \{1,\ldots,m\} \times \{1,\ldots,n\}$, with $|\Omega| \ll mn$. Popularly dubbed as the matrix completion problem (Candès and Recht, 2009; Mazumder et al., 2010), the task is to predict the unobserved entries, under the assumption that the underlying matrix is of low-rank. This leads to the natural rank regularized optimization problem:

$$\min_X \quad \frac{1}{2}\|\mathcal{P}_\Omega(X-Y)\|_F^2 + \lambda \,\mathrm{rank}(X), \tag{1}$$

where, $\mathcal{P}_\Omega(X)$ denotes the projection of $X_{m\times n}$ onto the observed indices Ω and is zero otherwise; and $\|\cdot\|_F$ denotes the usual Frobenius norm of a matrix. Problem (1), however, is computationally difficult due to the presence of the combinatorial rank constraint (Chistov and Grigor'ev, 1984). A natural convexification (Fazel, 2002; Recht et al., 2010) of $\mathrm{rank}(X)$ is $\|X\|_*$, the nuclear norm of X, which leads to the following surrogate of Problem (1):

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

Candès and Recht (2009); Candès and Plan (2010) show that under some assumptions on the underlying "population" matrix, a solution to Problem (2) approximates a solution to

Problem (1) reasonably well. The estimator obtained from Problem (2) works quite well: the nuclear norm shrinks the singular values and simultaneously sets many of the singular values to zero, thereby encouraging low-rank solutions. It is thus not surprising that Problem (2) has enjoyed a significant amount of attention in the wider statistical community over the last decade. There have been impressive advances in understanding its statistical properties (Candès and Plan, 2010; Candès and Tao, 2010; Recht et al., 2010; Recht, 2011; Gross, 2011; Rohde and Tsybakov, 2011; Koltchinskii et al., 2011; Negahban and Wainwright, 2011; Chen, 2015; Lecué and Mendelson, 2018; Chen et al., 2019b). Motivated by the work of Candès and Recht (2009); Cai et al. (2010), the authors in Mazumder et al. (2010) proposed SOFT-IMPUTE, an EM-flavored (Dempster et al., 1977) algorithm for optimizing Problem (2). For some other computational work in developing scalable algorithms for Problem (2), see the papers Jaggi and Sulovský (2010); Freund et al. (2015); Hastie et al. (2016), and references therein. Typical assumptions under which the nuclear norm works as a good proxy for the low-rank problem require the entries of the singular vectors of the "true" low-rank matrix to be sufficiently spread, and the missing pattern to be roughly uniform. The proportion of observed entries needs to be sufficiently larger than the number of parameters of the matrix $O\left((m+n)r\right)$, where, r denotes the rank of the true underlying matrix. Some extensions under general sampling distribution has been made in Klopp (2014); Alquier (2015). Negahban and Wainwright (2012) proposes improvements with a (convex) weighted nuclear norm penalty in addition to spikiness constraints for the noisy matrix completion problem.

The nuclear norm penalization framework, however, has limitations. If some conditions mentioned above fail, Problem (2) may fall short of delivering reliable low-rank estimators with good prediction performance (on the missing entries). Since the nuclear norm shrinks the singular values, in order to obtain an estimator with good explanatory power, it often results in a matrix estimator with high numerical rank — thereby leading to models that have higher rank than what might be desirable. The limitations mentioned above, however, should not come as a surprise to an expert — especially, if one draws a parallel connection to the LASSO (Tibshirani, 1996), a popular sparsity inducing shrinkage mechanism effectively used in the context of sparse linear modeling and regression. In the linear regression context, the LASSO often leads to dense models and suffers when the features are highly correlated — the limitations of the LASSO are quite well known in the statistics literature, and there have been major strides in moving beyond the convex ℓ_1-penalty to more aggressive forms of nonconvex penalties (Fan and Li, 2001; Zou and Li, 2008; Mazumder et al., 2011; Zhang, 2010; Zhang and Zhang, 2012; Loh and Wainwright, 2015; Bertsimas et al., 2016; Zheng et al., 2017; Feng and Zhang, 2017). The key principle in these methods is the use of nonconvex regularizers that better approximate the ℓ_0-penalty, leading to possibly nonconvex estimation problems. Thusly motivated, we study herein, the following family of nonconvex regularized estimators for the task of (noisy) matrix completion:

$$\min_{X} \underbrace{\frac{1}{2}\|\mathcal{P}_{\Omega}(X-Y)\|_F^2 + \sum_{i=1}^{\min\{m,n\}} P(\sigma_i(X);\lambda,\gamma)}_{:=f(X)}, \tag{3}$$

where, $\sigma_i(X), i \geq 1$ are the singular values of X and $\sigma \mapsto P(\sigma;\lambda,\gamma)$ is a concave penalty function on $[0,\infty)$ that takes the value ∞ whenever $\sigma < 0$. We will denote an estimator obtained from Problem (3) by $\hat{X}_{\lambda,\gamma}$. The family of penalty functions $P(\sigma;\lambda,\gamma)$ is indexed by the parameters (λ,γ) — these parameters together control the amount of nonconvexity and shrinkage — see for example Mazumder et al. (2011); Zhang and Zhang (2012) and also Section 2, herein, for examples of such nonconvex families.

A caveat in considering problems of the form (3) is that they lead to nonconvex optimization problems and thus obtaining a certifiably optimal global minimizer is generally difficult. Fairly recently, Bertsimas et al. (2016); Mazumder and Radchenko (2015) have shown that subset selection problems in sparse linear regression can be computed using advances in mixed integer quadratic optimization. Such global optimization methods, however, do not apply to matrix variate problems involving spectral[1] penalties, as in Problems (1) or (3). The main focus in our work herein is to develop a computationally scalable algorithmic framework that allows us to obtain high quality stationary points or upper bounds[2] for Problem (3) — we obtain a path of solutions $\hat{X}_{\lambda,\gamma}$ across a grid of values of (λ,γ) for Problem (3) by employing warm-starts, following the path-following scheme proposed in Mazumder et al. (2011). Leveraging problem structure, modern advances in computationally scalable low-rank SVDs and appropriately advancing the tricks successfully employed in Mazumder et al. (2010); Hastie et al. (2016), we empirically demonstrate the computational scalability of our method for problems of the size of the Netflix dataset, a matrix of size (approx.) $480,000 \times 18,000$ with $\sim 10^8$ observed entries. Perhaps most importantly, we demonstrate

[1] We say that a function is a spectral function of a matrix X, if it depends only upon the singular values of X. The state of the art algorithmics in mixed integer Semidefinite optimization problems is in its nascent stage; and not even comparable to the technology for mixed integer quadratic optimization.

[2] Since the problems under consideration are nonconvex, our methods are not guaranteed to reach the global minimum – we thus refer to the solutions obtained as *upper bounds*. In many synthetic examples, however, the solutions are indeed seen to be globally optimal. We do show rigorously, however, that these solutions are first order stationary points for the optimization problems under consideration.

empirically that the resultant estimators lead to better statistical properties (i.e., the estimators have lower rank and enjoy better prediction performance) over nuclear norm based estimates, on a variety of problem instances.

Some recent works (Jain et al., 2010, 2013; Hardt, 2014; Hardt and Wootters, 2014; Chen and Wainwright, 2015; Ma et al., 2017; Chen et al., 2019a) study the scope of alternating minimization or (projected) gradient stylized algorithmic strategies for the rank constrained optimization problem, similar to Problem (1) — see also Hastie et al. (2016) for related discussions. We should emphasize that our work herein, studies the *entire* family of nonconvex spectral penalized problems of the form of Problem (3), and is hence more general than the class of estimation problems considered in those works. We establish empirically that this flexible family of nonconvex penalized estimators leads to solutions with better statistical properties than those available from particular instantiations of the penalty function — nuclear norm regularization (2) and rank regularization (1). Along the lines of the aforementioned works, there exists an active stream of research on characterizing the global optimality of local algorithms for various matrix factorization based formulations (Bhojanapalli et al., 2016; Ge et al., 2016; Sun and Luo, 2016; Zheng and Lafferty, 2016; Ge et al., 2017; Shapiro et al., 2018). Our paper focuses on a more general family of nonconvex regularization, with admittedly less strong algorithmic guarantees. Finally, a series of iterative reweighted algorithms have been proposed and discussed (Mazumder et al., 2010; Mohan and Fazel, 2010; Fornasier et al., 2011; Mohan and Fazel, 2012; Gu et al., 2017), largely motivated by the reweighting ideas from sparse recovery problems (Zou, 2006; Candes et al., 2008; Daubechies et al., 2010). Different weight formulas have been suggested to improve the statistical and computational efficiency. These are, however, beyond the scope of the current paper.

1.1 Contributions and Outline

The main contributions of our paper can be summarized as follows:

- We propose a computational framework for nonconvex penalized matrix completion problems of the form (3). Our algorithm: NC-IMPUTE, may be thought of as a novel adaptation (with important enhancements and modifications) of the EM-stylized procedure SOFT-IMPUTE (Mazumder et al., 2010) to more general nonconvex penalized thresholding operators.
- We present an in-depth investigation of nonconvex spectral thresholding operators, which form the main building block of our algorithm. We also study their effective degrees of freedom (*df*), which provide a simple and intuitive way to calibrate the two-dimensional grid of tuning parameters, extending the scope of the method proposed in nonconvex penalized (least squares) regression by Mazumder et al. (2011) to spectral thresholding operators. We propose computationally efficient methods to approximate the *df* using tools from random matrix theory.
- We provide comprehensive computational guarantees of our algorithm — this includes the number of iterations needed to reach a first order stationary point and the asymptotic convergence of the sequence of estimates produced by NC-IMPUTE.
- Every iteration of NC-IMPUTE requires the computation of a low-rank SVD of a structured matrix, for which we propose new methods. Using efficient warm-start tricks to speed up the low-rank computations, we demonstrate the effectiveness of our proposal to large scale instances up to the Netflix size in reasonable computation times.
- Over a wide range of synthetic and real-data examples, we show that our proposed nonconvex penalized framework leads to high quality solutions with excellent statistical properties, which are often found to be significantly better than nuclear norm regularized solutions in terms of producing low-rank solutions with good predictive performances.
- Implementations of our algorithms in the `R` programming language have been made publicly available on `github` at: `https://github.com/diegofrasal/ncImpute`.

The remainder of the paper is organized as follows. Section 2 studies several properties of nonconvex spectral penalties and associated spectral thresholding operators, including their effective degrees of freedom. Section 3 describes our algorithmic framework NC-IMPUTE and studies the convergence properties of the algorithm. Section 4 presents numerical experiments demonstrating the usefulness of nonconvex penalized estimation procedures in terms of superior statistical properties on several synthetic datasets — we also show the usefulness of these estimators on several real data instances. Section 5 contains the conclusions and discusses several important future research directions. To improve readability, some technical materials and empirical results are relegated to Section 6.

Notation: For a matrix $A_{m\times n}$, we denote its (i,j)th entry by a_{ij}. $\mathcal{P}_\Omega(A)$ is a matrix with its (i,j)th entry given by a_{ij} for $(i,j)\in\Omega$ and zero otherwise, with $\Omega\subset\{1,\ldots,m\}\times\{1,\ldots,n\}$. We use the notation $\mathcal{P}_\Omega^\perp(A)=A-\mathcal{P}_\Omega(A)$ to denote the projection of A onto the complement of Ω. Let $\sigma_i(A), i=1,\ldots,\max\{m,n\}$ denote the singular values of A, with $\sigma_i(A)\geq\sigma_{i+1}(A)$ (for all i) – we will use the notation $\boldsymbol{\sigma}(A)$ to denote the vector of singular values. When clear from the context, we will simply write $\boldsymbol{\sigma}$ instead of $\boldsymbol{\sigma}(A)$. For a vector $\mathbf{a}=(a_1,\ldots,a_n)\in\mathbb{R}^n$, we will use the

notation $\mathrm{diag}(\mathbf{a})$ to denote an $n \times n$ diagonal matrix with ith diagonal entry being a_i.

2 Spectral Thresholding Operators

We begin our analysis by considering the fully observed version of Problem (3), given by:

$$\min_X \underbrace{\frac{1}{2}\|X - Z\|_F^2 + \sum_{i=1}^{\min\{m,n\}} P(\sigma_i(X); \lambda, \gamma)}_{:=g(X)} \tag{4}$$

where, for a given matrix Z, a minimizer of the function $g(X)$, denoted by $S_{\lambda,\gamma}(Z)$, is the *spectral thresholding operator* induced by the spectral penalty $\sum_i P(\sigma_i(X); \lambda, \gamma)$. Suppose $U\mathrm{diag}(\boldsymbol{\sigma})V'$ denotes the SVD of Z. For the nuclear norm regularized problem with the penalty function $P(\sigma_i(X); \lambda, \gamma) = \lambda\sigma_i(X)$, the corresponding thresholding operator, denoted by $S_{\lambda,\ell_1}(Z)$ (say), is given by the familiar soft-thresholding operator (Cai et al., 2010; Mazumder et al., 2010):

$$S_{\lambda,\ell_1}(Z) := U\mathrm{diag}(s_{\lambda,\ell_1}(\boldsymbol{\sigma}))V' \tag{5}$$

where, $s_{\lambda,\ell_1}(\sigma_i) := (\sigma_i - \lambda)_+$, $(\cdot)_+ = \max\{\cdot, 0\}$ and $s_{\lambda,\ell_1}(\sigma_i)$ is the ith entry of $s_{\lambda,\ell_1}(\boldsymbol{\sigma})$ (due to separability of the thresholding operator). Here, $S_{\lambda,\ell_1}(Z)$ is the the soft-thresholding operator on the singular values of Z and plays a crucial role in the SOFT-IMPUTE algorithm (Mazumder et al., 2010). For the rank regularized problem, with

$$P(\sigma_i(X); \lambda, \gamma) = \lambda \mathbb{1}(\sigma_i(X) > 0),$$

the thresholding operator denoted by $S_{\lambda,\ell_0}(Z)$ is given by the hard-thresholding operator (Mazumder et al., 2010):

$$S_{\lambda,\ell_0}(Z) := U\mathrm{diag}(s_{\lambda,\ell_0}(\boldsymbol{\sigma}))V' \tag{6}$$

with $s_{\lambda,\ell_0}(\sigma_i) = \sigma_i \mathbb{1}(\sigma_i > \sqrt{2\lambda})$. A closely related thresholding operator that retains the top r singular values and sets the remaining to zero formed the basis of the HARD-IMPUTE algorithm in Mazumder et al. (2010); Troyanskaya et al. (2001). The results in (5) and (6) suggest a curious link — the spectral thresholding operators (for the two specific choices of the spectral penalty functions given above) are tied to the corresponding thresholding functions that operate only on the singular values of the matrix — in other words, the operators $S_{\lambda,\ell_1}(Z), S_{\lambda,\ell_0}(Z)$ do *not* change the singular vectors of the matrix Z. It turns out that a similar result holds true for more general spectral penalty functions $P(\cdot; \lambda, \gamma)$ as the following proposition illustrates.

Proposition 1 *Let $Z = U\mathrm{diag}(\boldsymbol{\sigma})V'$ denote the SVD of Z, and $s_{\lambda,\gamma}(\boldsymbol{\sigma})$ denote the following thresholding operator on the singular values of Z:*

$$s_{\lambda,\gamma}(\boldsymbol{\sigma}) \in \arg\min_{\boldsymbol{\alpha}\geq \mathbf{0}} \underbrace{\frac{1}{2}\|\boldsymbol{\alpha} - \boldsymbol{\sigma}\|_2^2 + \sum_{i=1}^{\min\{m,n\}} P(\alpha_i; \lambda, \gamma)}_{:=\bar{g}(\boldsymbol{\alpha})}. \tag{7}$$

Then $S_{\lambda,\gamma}(Z) = U\mathrm{diag}(s_{\lambda,\gamma}(\boldsymbol{\sigma}))V'$.

Proof Note that by the Wielandt-Hoffman inequality (Horn and Johnson, 2012) we have that: $\|X - Z\|_F^2 \geq \|\boldsymbol{\sigma}(X) - \boldsymbol{\sigma}(Z)\|_2^2$, where, for a vector $\mathbf{a}$, $\|\mathbf{a}\|_2$ denotes the standard Euclidean norm. Equality holds when X and Z share the same left and right singular vectors. This leads to:

$$\begin{aligned}&\frac{1}{2}\|X - Z\|_F^2 + \sum_{i=1}^{\min\{m,n\}} P(\sigma_i(X); \lambda, \gamma)\\ \geq &\frac{1}{2}\|\boldsymbol{\sigma}(X) - \boldsymbol{\sigma}(Z)\|_2^2 + \sum_{i=1}^{\min\{m,n\}} P(\sigma_i(X); \lambda, \gamma).\end{aligned}$$

In the above inequality, note that the left hand side is $g(X)$ (defined in (4)) and right hand side is $\bar{g}(\boldsymbol{\sigma}(X))$ (defined in (7)). It follows that

$$\min_X \; g(X) \geq \min_{\boldsymbol{\sigma}(X)} \; \bar{g}(\boldsymbol{\sigma}(X)) = \bar{g}\left(s_{\lambda,\gamma}(\boldsymbol{\sigma})\right), \tag{8}$$

where, we used the observation that $\boldsymbol{\sigma}(X) \geq \mathbf{0}$ and $s_{\lambda,\gamma}(\boldsymbol{\sigma})$, as defined in (7) minimizes $\bar{g}(\boldsymbol{\sigma}(X))$. In addition, this minimum is attained by the function $g(X)$, at the choice $X = U\mathrm{diag}(s_{\lambda,\gamma}(\boldsymbol{\sigma}))V'$. This completes the proof of the proposition.

Due to the separability of the optimization Problem (7) across the coordinates, i.e., $\bar{g}(\boldsymbol{\alpha}) = \sum_i \bar{g}_i(\alpha_i)$ (where, $\bar{g}_i(\cdot)$ is defined in (9)), it suffices to consider each of the subproblems separately. Let $s_{\lambda,\gamma}(\sigma_i)$ denote a minimizer of $\bar{g}_i(\alpha)$, i.e.,

$$s_{\lambda,\gamma}(\sigma_i) \in \arg\min_{\alpha\geq 0} \; \bar{g}_i(\alpha) := \frac{1}{2}(\alpha - \sigma_i)^2 + P(\alpha; \lambda, \gamma). \tag{9}$$

It is easy to see that the ith coordinate of $s_{\lambda,\gamma}(\boldsymbol{\sigma})$ is given by $s_{\lambda,\gamma}(\sigma_i)$. This discussion suggests that our understanding of the spectral thresholding operator $S_{\lambda,\gamma}(Z)$ is intimately tied to the univariate thresholding operator (9). Thusly motivated, in the following, we present a concise discussion about univariate penalty functions and the resultant thresholding operators. We begin with some examples of concave penalties that are popularly used in statistics in the context of sparse linear modeling.

Families of Nonconvex Penalty Functions: Several types of nonconvex penalties are popularly used in high-dimensional regression frameworks—see for example, Nikolova (2000); Lv and Fan (2009); Zhang and Zhang (2012). For our setup, since these penalty functions operate on the singular values of a matrix, it suffices to consider nonconvex functions that are defined only on the nonnegative real numbers. We present a few examples below:

- The ℓ_γ penalty (Frank and Friedman, 1993) given by

  $$P(\sigma;\lambda,\gamma) = \lambda\sigma^\gamma,$$

  where $\lambda > 0$ and $0 \le \gamma < 1$.
- The SCAD penalty (Fan and Li, 2001) is defined via:

  $$P'(\sigma;\lambda,\gamma) = \lambda\mathbb{1}(\sigma \le \lambda) + \frac{(\gamma\lambda-\sigma)_+}{\gamma-1}\mathbb{1}(\sigma > \lambda),$$

  where $\lambda > 0, \gamma > 2$, and $P'(\sigma;\lambda,\gamma)$ denotes the derivative of $\sigma \mapsto P(\sigma;\lambda,\gamma)$ on $\sigma \ge 0$ with $P(0;\lambda,\gamma) = 0$.
- The MC+ penalty (Zhang, 2010; Mazumder et al., 2011) defined as

  $$P(\sigma;\lambda,\gamma) = \lambda\left(\sigma - \frac{\sigma^2}{2\lambda\gamma}\right)\mathbb{1}(0 \le \sigma < \lambda\gamma) + \frac{\lambda^2\gamma}{2}\mathbb{1}(\sigma \ge \lambda\gamma),$$

  with $\lambda > 0, \gamma > 0$.
- The log-penalty, with

  $$P(\sigma;\lambda,\gamma) = \lambda\log(\gamma\sigma+1)/\log(\gamma+1)$$

  on $\lambda > 0$ and $\gamma > 0$.

Figure 1 shows some members of the above nonconvex penalty families. The ℓ_γ penalty function is non differentiable at $\sigma = 0$, due to the unboundedness of $P'(\sigma;\lambda,\gamma)$ as $\sigma \to 0+$. The nonzero derivative at $\sigma = 0+$ encourages sparsity. The ℓ_γ penalty functions show a clear transition from the ℓ_1 penalty to the ℓ_0 penalty — similarly, the resultant thresholding operators show a passage from the soft-thresholding to the hard-thresholding operator. Let us examine the analytic form of the thresholding function induced by the MC+ penalty (for any $\gamma > 1$):

$$s_{\lambda,\gamma}(\sigma) = \begin{cases} 0, & \text{if } \sigma \le \lambda \\ \left(\frac{\sigma-\lambda}{1-1/\gamma}\right), & \text{if } \lambda < \sigma \le \lambda\gamma \\ \sigma, & \text{if } \sigma > \lambda\gamma. \end{cases} \tag{10}$$

It is interesting to note that for the MC+ penalty, the derivatives are all bounded and the thresholding functions are continuous for all $\gamma > 1$. As $\gamma \to \infty$, the threshold operator (10) coincides with the soft-thresholding operator. However, as $\gamma \to 1+$ the threshold operator approaches the discontinuous hard-thresholding operator $\sigma\mathbb{1}(\sigma \ge \lambda)$ — this is illustrated in Figure 1 and can also be observed by inspecting (10). Note that the ℓ_1 penalty penalizes small and large singular values in a similar fashion, thereby incurring an increased bias in estimating the larger coefficients. For the MC+ and SCAD penalties, we observe that they penalize the larger coefficients less severely than the ℓ_1 penalty — simultaneously, they penalize the smaller coefficients in a manner similar to that of the ℓ_1 penalty. On the other hand, the ℓ_γ penalty (for small values of γ) imposes a more severe penalty for values of $\sigma \approx 0$, quite different from the behavior of other penalty functions. In general, for a given family of nonconvex penalties $P(\sigma;\lambda,\gamma)$, the effect of (λ,γ) on the nonconvexity can be characterized through the general concavity quantity ϕ_P that is to be introduced in (11).

2.1 Properties of Spectral Thresholding Operators

The nonconvex penalty functions described in the previous section are concave functions on the nonnegative real line. We will now discuss measures that may be thought (loosely speaking) to measure the amount of concavity in the functions. For a univariate penalty function $\alpha \mapsto P(\alpha;\lambda,\gamma)$ on $\alpha \ge 0$, assumed to be differentiable on $(0,\infty)$, we introduce the following quantity (ϕ_P) that measures the amount of concavity (see also, Zhang (2010)) of $P(\alpha;\lambda,\gamma)$:

$$\phi_P := \inf_{\alpha,\alpha'>0} \frac{P'(\alpha;\lambda,\gamma) - P'(\alpha';\lambda,\gamma)}{\alpha - \alpha'}, \tag{11}$$

where $P'(\alpha;\lambda,\gamma)$ denotes the derivative of $P(\alpha;\lambda,\gamma)$ wrt α on $\alpha > 0$.

We say that the function $g(X)$ (as defined in (4)) is τ-*strongly convex* if the following condition holds:

$$g(X) \ge g(\widetilde{X}) + \langle \nabla g(\widetilde{X}), X - \widetilde{X}\rangle + \frac{\tau}{2}\|X - \widetilde{X}\|_F^2, \tag{12}$$

for some $\tau \ge 0$ and all $X, \widetilde{X}$. In inequality (12), $\nabla g(\widetilde{X})$ denotes *any* subgradient (assuming it exists) of $g(X)$ at $\widetilde{X}$. If $\tau = 0$ then the function is simply convex[3]. Using standard properties of spectral functions (Borwein and Lewis, 2006; Lewis, 1995), it follows that $g(X)$ is τ-strongly convex iff the vector function:

$$\bar{g}(\boldsymbol{\alpha}) = \frac{1}{2}\|\boldsymbol{\alpha} - \boldsymbol{\sigma}(Z)\|_2^2 + \sum_{i=1}^{\min\{m,n\}} P(\alpha_i;\lambda,\gamma) \tag{13}$$

is τ-strongly convex on $\{\boldsymbol{\alpha} : \boldsymbol{\alpha} \ge \mathbf{0}\}$, where $\boldsymbol{\sigma}(Z)$ denotes the singular values of Z. Let us recall the separable decomposition of $\bar{g}(\boldsymbol{\alpha}) = \sum_i \bar{g}_i(\alpha_i)$, with $\bar{g}_i(\alpha)$ as defined in (9). Clearly, the function $\boldsymbol{\alpha} \mapsto \bar{g}(\boldsymbol{\alpha})$ is τ-strongly convex (on the nonnegative reals) iff each summand $\bar{g}_i(\alpha)$ is τ-strongly

[3] Note that we consider $\tau \ge 0$ in the definition so that it includes the case of (non strong) convexity.

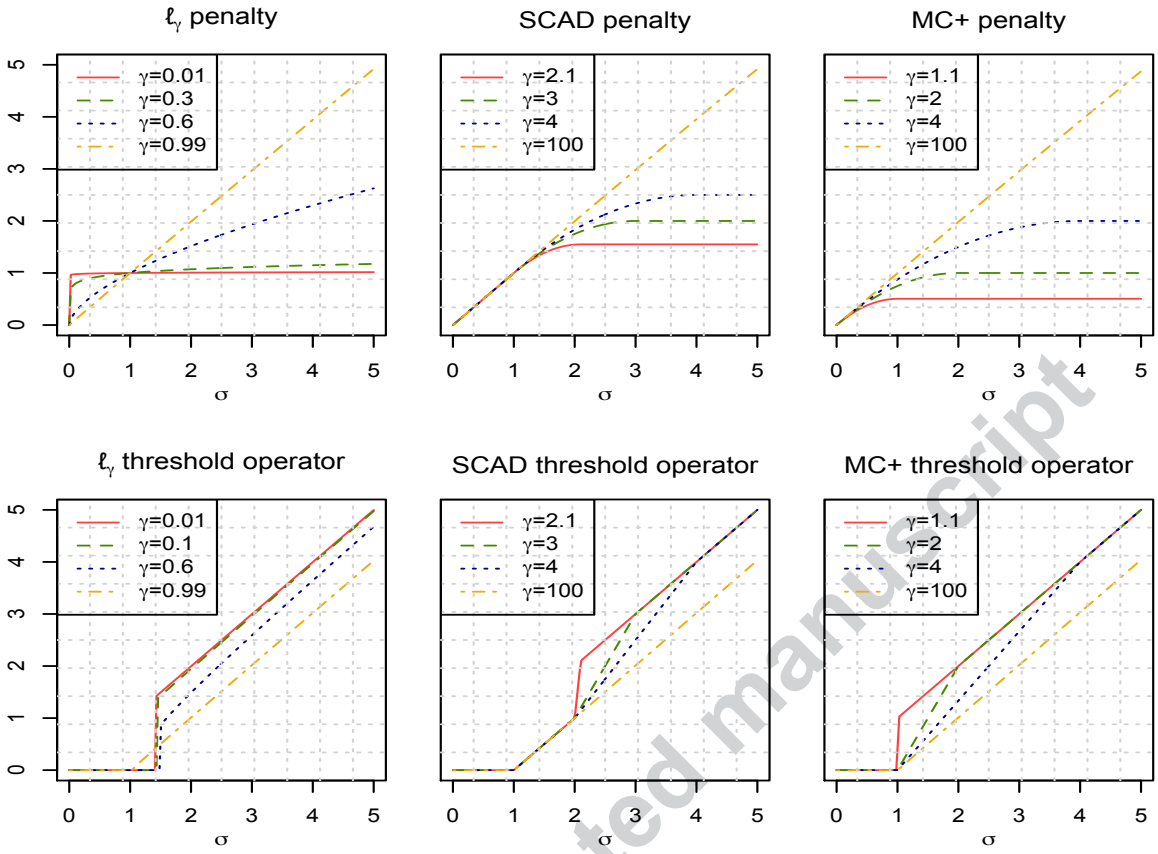


Fig. 1 [Top panel] Examples of nonconvex penalties $\sigma \mapsto P(\sigma;\lambda,\gamma)$ with $\lambda = 1$ for different values of γ. [Bottom Panel] The corresponding scalar thresholding operators: $\sigma \mapsto s_{\lambda,\gamma}(\sigma)$. At $\sigma = 1$, some of the thresholding operators corresponding to the ℓ_γ penalty function are discontinuous, and some of the other thresholding functions are "close" to being so.

convex on $\alpha \geq 0$. Towards this end, notice that $\bar{g}_i(\alpha)$ is convex on $\alpha \geq 0$ iff $1 + \phi_P \geq 0$ — in particular, $\bar{g}_i(\alpha)$ is τ-strongly convex with parameter $\tau = 1+\phi_P$, provided this number is nonnegative. In this vein, we have the following proposition:

Proposition 2 *Suppose $\phi_P > -1$, then the function $X \mapsto g(X)$ is τ-strongly convex with $\tau = 1 + \phi_P$.*

For the MC+ penalty, the condition $\tau = 1 + \phi_P > 0$ is equivalent to $\gamma > 1$. For the ℓ_γ penalty function, with $\gamma < 1$, the parameter $\tau = -\infty$, and thus the function $g(X)$ is not strongly convex.

Proposition 3 *Suppose $1 + \phi_P > 0$, then $Z \mapsto S_{\lambda,\gamma}(Z)$ is Lipschitz continuous with constant $\frac{1}{1+\phi_P}$, i.e, for all Z_1, Z_2 we have:*

$$\|S_{\lambda,\gamma}(Z_1) - S_{\lambda,\gamma}(Z_2)\|_F \leq \frac{1}{1+\phi_P}\|Z_1 - Z_2\|_F. \tag{14}$$

Proof We rewrite $g(X)$ as:

$$g(X) = \left\{\frac{1}{2}\|X - Z\|_F^2 - \frac{\psi}{2}\|X\|_F^2\right\} + \left\{\sum_{i=1}^{\min\{m,n\}} P(\sigma_i(X);\lambda,\gamma) + \frac{\psi}{2}\|X\|_F^2\right\}. \tag{15}$$

We have that $\|X\|_F^2 = \sum_{i=1}^{\min\{m,n\}} \sigma_i^2(X)$. Using the shorthand notation $\widetilde{P}(\sigma_i(X)) = P(\sigma_i(X);\lambda,\gamma) + \frac{\psi}{2}\sigma_i^2(X)$, and rearranging the terms in (15), it follows that $S_{\lambda,\gamma}(Z)$, a minimizer of $g(X)$, is given by:

$$S_{\lambda,\gamma}(Z) \in \arg\min_X \left\{\frac{1-\psi}{2}\|X - \frac{1}{1-\psi}Z\|_F^2 + \sum_{i=1}^{\min\{m,n\}} \widetilde{P}(\sigma_i(X))\right\}. \tag{16}$$

If $\psi + \phi_P > 0$, the function $\sigma_i \mapsto \widetilde{P}(\sigma_i)$ is convex for every i. If $1 - \psi > 0$, then the first term appearing in the objective function in (16) is convex. Thus, assuming

$\psi + \phi_P > 0, 1 - \psi > 0$ both summands in the above objective function are convex. In particular, the optimization problem (16) is convex and $Z \mapsto S_{\lambda,\gamma}(Z)$ can be viewed as a convex proximal map (Rockafellar, 1970). Using standard contraction properties of proximal maps, we have that:

$$\|S_{\lambda,\gamma}(Z_1) - S_{\lambda,\gamma}(Z_2)\|_F \leq \left\| \frac{Z_1}{1-\psi} - \frac{Z_2}{1-\psi} \right\|_F \leq \frac{1}{1-\psi}\|Z_1 - Z_2\|_F.$$

Since the above holds true for any ψ as chosen above, optimizing over the value of ψ such that Problem (16) remains convex gives us $\hat{\psi} = -\phi_P$, i.e., $1/(1-\hat{\psi}) = 1/(1+\phi_P)$, thereby leading to (14).

2.2 Effective Degrees of Freedom for Spectral Thresholding Operators

In this section, to better understand the statistical properties of spectral thresholding operators, we study their degrees of freedom. The effective degrees of freedom or *df* is a popularly used statistical notion that measures the amount of "fitting" performed by an estimator (Efron et al., 2004; Hastie et al., 2009; Stein, 1981). In the case of classical linear regression, for example, *df* is simply given by the number of features used in the linear model. This notion applies more generally to additive fitting procedures. Following Efron et al. (2004); Stein (1981), let us consider an additive model of the form:

$$Z_{ij} = \mu_{ij} + \varepsilon_{ij} \quad \text{with} \quad \varepsilon_{ij} \overset{\text{iid}}{\sim} N(0, \mathrm{v}^2), \tag{17}$$

for $i = 1, \ldots, m, j = 1, \ldots, n$. The *df* of $\hat{\mu} := \hat{\mu}(Z)$, for the fully observed model above, i.e., (17) is given by:

$$df(\hat{\mu}) = \sum_{ij} \mathrm{Cov}(\hat{\mu}_{ij}, Z_{ij})/\mathrm{v}^2,$$

where μ_{ij} denotes the (i,j)th entry of the matrix μ. For the particular case of a spectral thresholding operator we have $\hat{\mu} = S_{\lambda,\gamma}(Z)$. When $Z \mapsto \hat{\mu}(Z)$ satisfies a weak differentiability condition, the *df* may be computed via a divergence formula (Stein, 1981; Efron et al., 2004):

$$df(\hat{\mu}) = \mathbb{E}\left((\nabla \cdot \hat{\mu}(Z)) \cdot (Z)\right), \tag{18}$$

where $(\nabla \cdot \hat{\mu}) \cdot (Z) = \sum_{ij} \partial \hat{\mu}(Z_{ij})/\partial Z_{ij}$. For the spectral thresholding operator $S_{\lambda,\gamma}(\cdot)$, expression (18) holds if the map $Z \mapsto S_{\lambda,\gamma}(Z)$ is Lipschitz and hence weakly differentiable — see for example, Candès et al. (2013). In the light of Proposition 3, the map $Z \mapsto S_{\lambda,\gamma}(Z)$ is Lipschitz when $\phi_P + 1 > 0$. Under the model (17), the singular values of Z will have a multiplicity of one with probability one. We assume that the univariate thresholding operators are differentiable, i.e., $s'_{\lambda,\gamma}(\cdot)$ exists. With these assumptions in place, the divergence formula for $S_{\lambda,\gamma}(Z)$ can be obtained following Candès et al. (2013), as presented in the following proposition.

Proposition 4 *Assume that* $1 + \phi_P > 0$ *and the model* (17) *is in place. Then the degrees of freedom of the estimator* $S_{\lambda,\gamma}(Z)$ *is given by:*

$$df(S_{\lambda,\gamma}(Z)) = \mathbb{E}\sum_i \left(s'_{\lambda,\gamma}(\sigma_i) + |m-n|\frac{s_{\lambda,\gamma}(\sigma_i)}{\sigma_i} \right) + 2\mathbb{E}\sum_{i \neq j} \frac{\sigma_i s_{\lambda,\gamma}(\sigma_i)}{\sigma_i^2 - \sigma_j^2}, \tag{19}$$

where the σ_i*'s are the singular values of* Z.

We note that the above expression is true for any value of $1 + \phi_P > 0$. For the MC+ penalty function, expression (19) holds for $\gamma > 1$. As soon as $\gamma \leq 1$, the above method of deriving *df* does not apply due to the discontinuity in the map $Z \mapsto S_{\lambda,\gamma}(Z)$. Values of γ close to one (but larger), however, give an expression for the *df* near the hard-thresholding spectral operator, which corresponds to $\gamma = 1$.

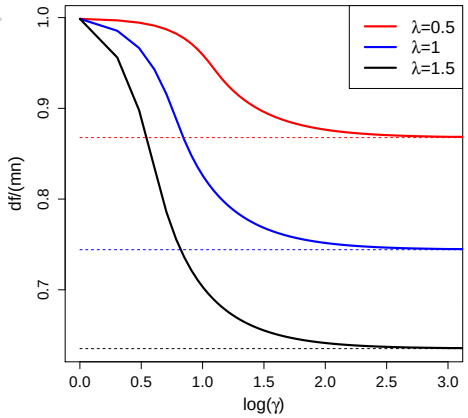


Fig. 2 Figure showing the *df* for the MC+ thresholding operator for a matrix with $m = n = 10$, $\mu = 0$ and $\mathrm{v} = 1$. The *df* profile as a function of γ (in the log scale) is shown for three values of λ. The dashed lines correspond to the *df* of the spectral soft-thresholding operator, corresponding to $\gamma = \infty$. We propose calibrating the (λ, γ) grid to a $(\widetilde{\lambda}, \widetilde{\gamma})$ grid such that the *df* corresponding to every value of $\widetilde{\gamma}$ matches the *df* of the soft-thresholding operator — as shown in Figure 3.

To understand the behavior of the *df* as a function of (λ, γ), let us consider the null model with $\mu = 0$ and the MC+ penalty function. In this case, for a fixed λ (see Figure 2 with a fixed $\lambda > 0$), the *df* is seen to increase with smaller γ values: the soft-thresholding function shrinks the large coefficients and sets all coefficients smaller than λ to be zero; the more aggressive (closer to the hard thresholding operator) shrinkage operators ($s_{\lambda,\gamma}(\sigma)$) shrink less for larger values of σ and set all coefficients smaller than λ to

zero. Thus, intuitively, the more aggressive thresholding operators should have larger *df* since they do more "fitting" — this is indeed observed in Figure 2. Mazumder et al. (2011) studied the *df* of the univariate thresholding operators in the linear regression problem, and observed a similar pattern in the behavior of the *df* across (λ, γ) values. For the linear regression problem, Mazumder et al. (2011) argued that it is desirable to choose a parametrization for (λ, γ) such that for a fixed λ, as one moves across γ, the *df* should be the same. We follow the same strategy for the spectral regularization problem considered herein — we reparametrize a two-dimensional grid of (λ, γ) values to a two-dimensional grid of $(\tilde{\lambda}, \tilde{\gamma})$ values, such that the *df* remain calibrated in the sense described above — this is illustrated in Figure 2 (see the horizontal dashed lines corresponding to the constant *df* values, after calibration). Figure 3 shows the lattice of $(\tilde{\lambda}, \tilde{\gamma})$ after calibration. The values of $(\tilde{\lambda}, \tilde{\gamma})$ on each curve induce the same *df*. As $\tilde{\gamma}$ moves down (the penalty becomes more "nonconvex"), the corresponding $\tilde{\lambda}$ (the shrinkage) has to increase to maintain the same *df*.

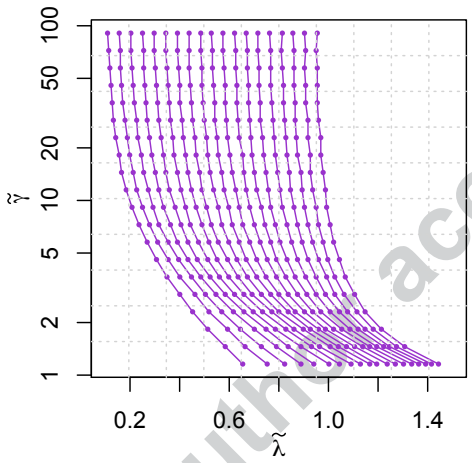


Fig. 3 Figure showing the calibrated $(\tilde{\lambda}, \tilde{\gamma})$ lattice — for every fixed value of $\tilde{\lambda}$, the *df* of the MC+ spectral threshold operators are the same across different $\tilde{\gamma}$ values. The *df* computations have been performed on a null model using Proposition 5.

The study of *df* presented herein provides a simple and intuitive explanation about the roles played by the parameters (λ, γ) for the fully observed problem. The notion of calibration provides a new parametrization of the family of penalties. From a computational viewpoint, since the general algorithmic framework presented in this paper (see Section 3) computes a regularization surface using warm-starts across adjacent (λ, γ) values on a two-dimensional grid; it is desirable for the adjacent values to be close — the *df* calibration also ensures this in a simple and intuitive manner.

Computation of df*:* The *df* estimate as implied by Proposition 4 depends only upon singular values (and not the singular vectors) of a matrix and can hence be computed with cost $O(\min\{m,n\}^2)$. The expectation can be approximated via Monte-Carlo simulation — these computations are easy to parallelize and can be done offline. Since we compute the *df* for the null model, for larger values of m, n we recommend using the Marchenko-Pastur law for iid Gaussian matrix ensembles to approximate the *df* expression (19). We illustrate the method using the MC+ penalty for $\gamma > 1$. Towards this end, let us define a function on $\beta \geq 0$

$$g_{\zeta,\gamma}(\beta) = \begin{cases} 0, & \text{if } \sqrt{\beta} \leq \zeta \\ \frac{\gamma}{\gamma-1}(1 - \frac{\zeta}{\sqrt{\beta}}), & \text{if } \zeta < \sqrt{\beta} \leq \zeta\gamma \\ 1, & \text{if } \sqrt{\beta} > \zeta\gamma\,. \end{cases}$$

For the following proposition, we will assume (for simplicity) that $m \geq n$.

Proposition 5 *Let $m, n \to \infty$ with $\frac{n}{m} \to \alpha \in (0, 1]$, then under the model $Z_{ij} \overset{iid}{\sim} N(0,1)$, we have*

$$\lim_{m,n\to\infty} \frac{df(S_{\lambda,\gamma}(Z))}{mn} = \begin{cases} 0 & \text{if } \frac{\lambda}{\sqrt{m}} \to \infty \\ (1-\alpha)\mathbb{E}g_{\zeta,\gamma}(T_1) + & \\ \alpha\mathbb{E}\left(\frac{T_1 g_{\zeta,\gamma}(T_1) - T_2 g_{\zeta,\gamma}(T_2)}{T_1 - T_2}\right) & \text{if } \frac{\lambda}{\sqrt{m}} \to \zeta \\ 1 & \text{if } \frac{\lambda}{\sqrt{m}} \to 0 \end{cases}$$

where $S_{\lambda,\gamma}(Z)$ is the thresholding operator corresponding to the MC+ penalty with $\lambda \geq 0, \gamma > 1$ and the expectation is taken with respect to T_1 and T_2 independently generated from the Marchenko-Pastur distribution (see Lemma 1, Section 6.1).

Proof For a proof, see Section 6.1.1.

Note that the variance v^2 in model (17) can always be assumed to be one (by adjusting the value of the tuning parameter accordingly[4]).

3 The NC-IMPUTE Algorithm

In this section, we present algorithm NC-IMPUTE. The algorithm is inspired by an EM-stylized procedure, similar to SOFT-IMPUTE (Mazumder et al., 2010), but has important innovations, as we will discuss shortly. It is helpful to recall that, for observed data: $\mathcal{P}_\Omega(Y)$, the algorithm SOFT-IMPUTE relies on the following update sequence

$$X_{k+1} = S_{\lambda,\ell_1}\left(\mathcal{P}_\Omega(Y) + \mathcal{P}_\Omega^\perp(X_k)\right), \tag{20}$$

[4] This follows from the simple observation that $s_{a\lambda,\gamma}(ax) = as_{\lambda,\gamma}(x)$ and $s'_{a\lambda,\gamma}(ax) = s'_{\lambda,\gamma}(x)$.

which can be interpreted as computing the nuclear norm regularized spectral thresholding operator for the following "fully observed" problem:

$$X_{k+1} \in \arg\min_X \left\{ \frac{1}{2} \left\| X - \left(\mathcal{P}_\Omega(Y) + \mathcal{P}_\Omega^\perp(X_k) \right) \right\|_F^2 + \lambda \|X\|_* \right\},$$

where, the missing entries are filled in by the current estimate, i.e., $\mathcal{P}_\Omega^\perp(X_k)$. We refer the reader to Mazumder et al. (2010) for a detailed study of the algorithm. Mazumder et al. (2010) suggest, in passing, the notion of extending SOFT-IMPUTE to more general thresholding operators; however, such generalizations were not pursued by the authors. In this paper, we present a thorough investigation about nonconvex generalized thresholding operators — we study their convergence properties, scalability aspects and demonstrate their superior statistical performance across a wide range of numerical experiments.

Update (20) suggests a natural generalization to more general nonconvex penalty functions, by simply replacing the spectral soft thresholding operator $S_{\lambda,\ell_1}(\cdot)$ with more general spectral operators $S_{\lambda,\gamma}(\cdot)$:

$$X_{k+1} = S_{\lambda,\gamma} \left(\mathcal{P}_\Omega(Y) + \mathcal{P}_\Omega^\perp(X_k) \right). \tag{21}$$

While the above update rule works quite well in our numerical experiments, it enjoys limited computational guarantees, as suggested by our convergence analysis in Section 3.1. We thus propose and study a seemingly minor generalization of the rule (21) — this modified rule enjoys superior finite time convergence rates to a first order stationary point. We develop our algorithmic framework below.

Let us define the following function:

$$F_\ell(X; X_k) := \frac{1}{2} \|\mathcal{P}_\Omega(X - Y)\|_F^2 + \frac{1}{2} \|\mathcal{P}_\Omega^\perp(X - X_k)\|_F^2 + \frac{\ell}{2} \|X - X_k\|_F^2 + \sum_{i=1}^{\min\{m,n\}} P(\sigma_i(X); \lambda, \gamma), \tag{22}$$

for $\ell \geq 0$. Note that $F_\ell(X; X_k)$ majorizes the objective function $f(X)$ defined in (3), i.e., $F_\ell(X; X_k) \geq f(X)$ for any X and X_k, with equality holding at $X = X_k$. In an attempt to obtain a minimum of Problem (3), we propose to iteratively minimize $F_\ell(X; X_k)$, an upper bound to $f(X)$, to obtain X_{k+1} — more formally, this leads to the following update sequence:

$$X_{k+1} \in \arg\min_X \; F_\ell(X; X_k). \tag{23}$$

Note that X_{k+1} is easy to compute; by some rearrangement of (22) we see:

$$X_{k+1} \in \arg\min_X \underbrace{\frac{\ell+1}{2} \|X - \widetilde{X}_k\|_F^2 + \sum_{i=1}^{\min\{m,n\}} P(\sigma_i(X); \lambda, \gamma)}_{:= S^\ell_{\lambda,\gamma}(\widetilde{X}_k)}, \tag{24}$$

where $\widetilde{X}_k = \left(\mathcal{P}_\Omega(Y) + \mathcal{P}_\Omega^\perp(X_k) + \ell X_k \right) / (\ell + 1)$. Note that (24) is a minor modification of (21) — in particular, if $\ell = 0$, then these two update rules coincide.

The sequence X_k defined via (24) has desirable convergence properties, as we discuss in Section 3.1. In particular, as $k \to \infty$, the sequence reaches (in a sense that will be made more precise later) a first order stationary point for Problem (3). We also provide a finite time convergence analysis of the update sequence (24).

We intend to compute an entire regularization surface of solutions to Problem (3) over a two-dimensional grid of (λ, γ)-values, using warm-starts. We take the MC+ family of functions as a running example, with $(\lambda, \gamma) \in \{\lambda_1 > \lambda_2 > \cdots > \lambda_N\} \times \{\infty := \gamma_1 > \gamma_2 > \ldots > \gamma_M\}$. At the beginning, we compute a path of solutions for the nuclear norm penalized problem, i.e., Problem (3) with $\gamma = \infty$ on a grid of λ values. For a fixed value of λ, we compute solutions to Problem (3) for smaller values of γ, gradually moving away from the convex problems. In this continuation scheme, we found the following strategies useful:

- For every value of (λ_i, γ_j), we apply two copies of the iterative scheme (23) initialized with solutions obtained from its two neighboring points $(\lambda_{i-1}, \gamma_j)$ and $(\lambda_i, \gamma_{j-1})$. From these two candidates, we select the one that leads to a smaller value of the objective function $f(\cdot)$ at (λ_i, γ_j).
- Instead of using a two-dimensional rectangular lattice, one can also use the recalibrated lattice, suggested in Section 2.2, as the two-dimensional grid of tuning parameters.

The algorithm outlined above, called NC-IMPUTE is summarized as Algorithm 1.

We now present an elementary convergence analysis of the update sequence (24). Since the problems under investigation herein are nonconvex, our analysis requires new ideas and techniques beyond those used in Mazumder et al. (2010) for the convex nuclear norm regularized problem.

3.1 Convergence Analysis

By the definition of X_{k+1} we have that:

$$F_\ell(X_{k+1}; X_k) = \min_X \; F_\ell(X; X_k) \leq F_\ell(X_k; X_k) = f(X_k).$$

Algorithm 1 NC-IMPUTE

1. Input: A search grid $\lambda_1 > \lambda_2 > \cdots > \lambda_N$; $+\infty := \gamma_1 > \gamma_2 > \cdots > \gamma_M$. Tolerance ε.
2. Compute solutions $\hat{X}_{\lambda_i,\gamma_1}$ for $i = 1,\ldots,N$, for the nuclear norm regularized problem.
3. For every $(\gamma,\lambda) \in \{\gamma_2,\ldots,\gamma_M\} \times \{\lambda_1,\ldots,\lambda_N\}$:
   (a) Initialize
   $$X^{\text{old}} = \arg\min_X \left\{ f(X), X \in \left\{ \hat{X}_{\lambda_{i-1},\gamma_j},\ \hat{X}_{\lambda_i,\gamma_{j-1}} \right\} \right\}.$$
   (b) Repeat until convergence, i.e., $\|X^{\text{new}} - X^{\text{old}}\|_F^2 < \varepsilon \|X^{\text{old}}\|_F^2$:
      (i) Compute $X^{\text{new}} \in \arg\min_X F_\ell(X; X^{\text{old}})$.
      (ii) Assign $X^{\text{old}} \leftarrow X^{\text{new}}$.
   (c) Assign $\hat{X}_{\lambda_i,\gamma_j} \leftarrow X^{\text{new}}$.
4. Output: $\hat{X}_{\lambda_i,\gamma_j}$ for $i = 1,\ldots,N$, $j = 1,\ldots,M$.

Let us define the quantities:

$$\nu(\ell) := 1 + \phi_P + \ell \quad \text{and} \quad \nu^\dagger(\ell) := \max\{\nu(\ell), 0\},$$

where, if $\nu(\ell) \geq 0$, then the function $X \mapsto F_\ell(X; X_k)$ is $\nu(\ell)$-strongly convex. In particular, from (23), it follows that $\nabla F_\ell(X_{k+1}; X_k)$, a subgradient of the map $X \mapsto F_\ell(X; X_k)$ (evaluated at X_{k+1}) equals zero. We thus have:

$$F_\ell(X_k; X_k) - F_\ell(X_{k+1}; X_k) \geq \frac{\nu(\ell)}{2} \|X_{k+1} - X_k\|_F^2. \tag{25}$$

Now note that, by the definition of X_{k+1}, we always have: $F_\ell(X_k; X_k) \geq F_\ell(X_{k+1}; X_k)$, which combined with (25) leads to (replacing $\nu(\ell)$ by $\nu^\dagger(\ell)$):

$$F_\ell(X_k; X_k) - F_\ell(X_{k+1}; X_k) \geq \frac{\nu^\dagger(\ell)}{2} \|X_{k+1} - X_k\|_F^2. \tag{26}$$

In addition, we have:

$$\begin{aligned} & F_\ell(X_{k+1}; X_k) \\ =& \frac{1}{2} \|\mathcal{P}_\Omega(X_{k+1} - Y)\|_F^2 + \sum_{i=1}^{\min\{m,n\}} P(\sigma_i(X_{k+1}); \lambda, \gamma) \\ & + \frac{1}{2} \|\mathcal{P}_\Omega^\perp(X_{k+1} - X_k)\|_F^2 + \frac{\ell}{2} \|X_{k+1} - X_k\|_F^2 \\ =& f(X_{k+1}) + \frac{1}{2} \|\mathcal{P}_\Omega^\perp(X_{k+1} - X_k)\|_F^2 + \frac{\ell}{2} \|X_{k+1} - X_k\|_F^2. \end{aligned} \tag{27}$$

Combining (26) and (27), and observing that $F_\ell(X_k; X_k) = f(X_k)$, we have:

$$\begin{aligned} & f(X_k) - f(X_{k+1}) \\ \geq & \frac{\nu^\dagger(\ell)}{2} \|X_{k+1} - X_k\|_F^2 + \frac{\ell}{2} \|X_{k+1} - X_k\|_F^2 \\ & + \frac{1}{2} \|\mathcal{P}_\Omega^\perp(X_{k+1} - X_k)\|_F^2 \\ = & \underbrace{\frac{\nu^\dagger(\ell) + \ell}{2} \|X_{k+1} - X_k\|_F^2 + \frac{1}{2} \|\mathcal{P}_\Omega^\perp(X_{k+1} - X_k)\|_F^2}_{:= \Delta_\ell(X_k; X_{k+1})}. \end{aligned} \tag{28}$$

Since $\Delta_\ell(X_k; X_{k+1}) \geq 0$, the above inequality immediately implies that $f(X_k) \geq f(X_{k+1})$ for all k; and the improvement in objective values is at least as large as the quantity $\Delta_\ell(X_k; X_{k+1})$. The term $\Delta_\ell(X_k; X_{k+1})$ is a measure of progress of the algorithm, as formalized by the following proposition.

Proposition 6 **(a)**: *Let $\nu^\dagger(\ell) + \ell > 0$ and for any X_a, let us consider the update $X_{a+1} \in \arg\min_X F_\ell(X; X_a)$. Then the following are equivalent:*

(i) $f(X_{a+1}) = f(X_a)$
(ii) $\Delta_\ell(X_a; X_{a+1}) = 0$
(iii) X_a *is a fixed point, i.e.,* $X_{a+1} = X_a$.

(b): *If $\nu^\dagger(\ell), \ell = 0$ and $\Delta_\ell(X_a; X_{a+1}) = 0$ then X_{a+1} is a fixed point.*

Proof Proof of Part (a):
We will show that (i) $\implies$ (ii) $\implies$ (iii) $\implies$ (i); by analyzing (28). If $f(X_{a+1}) = f(X_a)$ then $\Delta_\ell(X_a; X_{a+1}) = 0$. Since $\nu^\dagger(\ell) + \ell > 0$, we have that $X_{a+1} = X_a$, which trivially implies (i).

Proof of Part (b):
If $\nu^\dagger(\ell) + \ell = 0$, Part (a) needs to be slightly modified. Note that $\Delta_\ell(X_a; X_{a+1}) = 0$ iff $\mathcal{P}_\Omega^\perp(X_{a+1}) = \mathcal{P}_\Omega^\perp(X_a)$. Since $\ell = 0$, we have that $X_{a+2} = S_{\lambda,\gamma}\left(\mathcal{P}_\Omega(Y) + \mathcal{P}_\Omega^\perp(X_{a+1})\right)$. The condition $\mathcal{P}_\Omega^\perp(X_{a+1}) = \mathcal{P}_\Omega^\perp(X_a)$, implies that

$$S_{\lambda,\gamma}(\mathcal{P}_\Omega(Y) + \mathcal{P}_\Omega^\perp(X_{a+1})) = S_{\lambda,\gamma}\left(\mathcal{P}_\Omega(Y) + \mathcal{P}_\Omega^\perp(X_a)\right),$$

where the term on the right equals X_{a+1}. Thus, $X_{a+1} = X_{a+2} = \cdots$, i.e., X_{a+1} is a fixed point.

Since the $f(X_k)$'s form a decreasing sequence which is bounded from below, they converge to $\hat{f}$, say — this implies that $\Delta_\ell(X_k; X_{k+1}) \to 0$ as $k \to \infty$. Let us now consider two cases, depending upon the value of $\nu^\dagger(\ell) + \ell$. If $\nu^\dagger(\ell) + \ell > 0$, then we have $X_{k+1} - X_k \to 0$ as $k \to \infty$. On the other hand, if the quantities $\nu^\dagger(\ell) = 0, \ell = 0$, the conclusion needs to be modified: $\Delta_\ell(X_k; X_{k+1}) \to 0$ implies that $\mathcal{P}_\Omega^\perp(X_{k+1} - X_k) \to 0$ as $k \to \infty$.

Motivated by the above discussion, we make the following definition of a first order stationary point for Problem (3).

Definition 1 *X_a is said to be a first order stationary point for Problem (3) if $\Delta_\ell(X_a; X_{a+1}) = 0$. X_a is said to be an ϵ-accurate first order stationary point for Problem (3) if $\Delta_\ell(X_a; X_{a+1}) \leq \epsilon$.*

Proposition 7 *The sequence $f(X_k)$ is decreasing and suppose it converges to $\hat{f}$. Then the rate of convergence of X_k to this first order stationary point is given by:*

$$\min_{1\leq k\leq \mathcal{K}} \Delta_\ell(X_k; X_{k+1}) \leq \frac{1}{\mathcal{K}}\left(f(X_1) - \hat{f}\right). \tag{29}$$

Proof The arguments presented preceding Proposition 7 establish that the sequence $f(X_k)$ is decreasing and converges to $\hat{f}$, say.

Consider (28) for any $1 \leq k \leq \mathcal{K}$. We have that $\Delta_\ell(X_k; X_{k+1}) \leq f(X_k) - f(X_{k+1})$ — summing this inequality for $k = 1, \ldots, \mathcal{K}$ we obtain:

$$\begin{aligned}\mathcal{K} \min_{1\leq k\leq \mathcal{K}} \Delta_\ell(X_k; X_{k+1}) &\leq \sum_{1\leq k\leq \mathcal{K}} \Delta_\ell(X_k; X_{k+1}) \\ &\leq f(X_1) - f(X_{\mathcal{K}+1}) \leq f(X_1) - \hat{f},\end{aligned}$$

where in the last inequality we used the simple fact that $f(X_k) \downarrow \hat{f}$. Gathering the left and right parts of the above chain of inequalities leads to (29).

Proposition 7 shows that the sequence X_k reaches an ϵ-accurate first order stationary point within $K_\epsilon = (f(X_1) - \hat{f})/\epsilon$ many iterations. The number of iterations K_ϵ, depends upon how close the initial estimate $f(X_1)$ is to the eventual solution $\hat{f}$. Since NC-IMPUTE employs warm-starts, the constant appearing in the rhs of (29) suggests that the number of iterations required to a reach an approximate first order stationary point is quite low — this is indeed observed in our experiments, and this feature of using warm-starts makes our algorithm particularly attractive from a practical viewpoint.

3.1.1 Rank Stabilization

Let us consider the thresholding function $S^\ell_{\lambda,\gamma}(\widetilde{X}_k)$ defined in (24), which expresses X_{k+1} as a function of X_k. Using the development in Section 2, it is easy to see that the spectral operator $S^\ell_{\lambda,\gamma}(\widetilde{X}_k)$ is closely tied to the following vector thresholding operator (30), acting on the singular values of $\widetilde{X}_k$. Formally, for a given nonnegative vector $\widetilde{\mathbf{x}}$, if we denote:

$$\begin{aligned}s^\ell_{\lambda,\gamma}(\widetilde{\mathbf{x}}) \in \arg\min_{\boldsymbol{\alpha}\geq \mathbf{0}} \Bigg\{ &\frac{\ell+1}{2}\|\boldsymbol{\alpha} - \widetilde{\mathbf{x}}\|_2^2 \\ &+ \sum_{i=1}^{\min\{m,n\}} P(\alpha_i; \lambda, \gamma)\Bigg\},\end{aligned} \tag{30}$$

then

$$S^\ell_{\lambda,\gamma}(\widetilde{X}) = \widetilde{U}\mathrm{diag}(s^\ell_{\lambda,\gamma}(\widetilde{\mathbf{x}}))\widetilde{V}',$$

where $\widetilde{X} = \widetilde{U}\mathrm{diag}(\widetilde{\mathbf{x}})\widetilde{V}'$ is the SVD of $\widetilde{X}$. Thus, properties of the thresholding function $S^\ell_{\lambda,\gamma}(\widetilde{X})$ are closely related to those of the vector thresholding operator $s^\ell_{\lambda,\gamma}(\widetilde{\mathbf{x}})$. Due to the separability of the vector thresholding operator $s^\ell_{\lambda,\gamma}(\widetilde{\mathbf{x}})$, across each coordinate of $\widetilde{\mathbf{x}}$, we denote by $s^\ell_{\lambda,\gamma}(\widetilde{x}_i)$, the ith coordinate of $s^\ell_{\lambda,\gamma}(\widetilde{\mathbf{x}})$.

We now investigate what happens to the rank of the sequence X_k as defined via (23). In particular, does this rank converge? We show that the rank stabilizes after finitely many iterations, under an additional assumption — namely the spectral thresholding operator is discontinuous — see Figure 1 for examples of discontinuous thresholding functions.

Proposition 8 *Consider the update sequence*

$$X_{k+1} = S^\ell_{\lambda,\gamma}(\widetilde{X}_k)$$

as defined in (24); and let $\nu^\dagger(\ell) + \ell > 0$. Suppose that there is a $\lambda_S > 0$ such that, for any scalar $\widetilde{x} \geq 0$, the following holds: $s^\ell_{\lambda,\gamma}(\widetilde{x}) \neq 0 \Longrightarrow |s^\ell_{\lambda,\gamma}(\widetilde{x})| > \lambda_S$ — i.e., the scalar thresholding operator $\widetilde{x} \mapsto s^\ell_{\lambda,\gamma}(\widetilde{x})$ is discontinuous. Then there exists an integer $\mathcal{K}^*$ such that for all $k \geq \mathcal{K}^*$, we have $\mathrm{rank}(X_k) = r$, i.e., the rank stabilizes after finitely many iterations.

Proof Using (28) it follows that

$$\begin{aligned}f(X_k) - f(X_{k+1}) &\geq \frac{\nu^\dagger(\ell)+\ell}{2}\|X_{k+1} - X_k\|_F^2 \\ &\geq \frac{\nu^\dagger(\ell)+\ell}{2}\|\boldsymbol{\sigma}_{k+1} - \boldsymbol{\sigma}_k\|_2^2,\end{aligned}$$

where the last inequality follows from Wielandt-Hoffman inequality (Horn and Johnson, 2012) and $\boldsymbol{\sigma}_k := \boldsymbol{\sigma}(X_k)$ denotes the vector of singular values of X_k. Let $\mathbb{1}(\boldsymbol{\sigma})$ be an indicator vector with ith coordinate being equal to $1(\sigma_i \neq 0)$. We will prove the result of rank stabilization via the method of contradiction. Suppose the rank does *not* stabilize, then $\mathbb{1}(\boldsymbol{\sigma}_{k+1}) \neq \mathbb{1}(\boldsymbol{\sigma}_k)$ for infinitely many k values. Thus there are infinitely many k' values such that:

$$\|\boldsymbol{\sigma}_{k'+1} - \boldsymbol{\sigma}_{k'}\|_2^2 \geq \sigma^2_{k'+1,i}\,,$$

where i is taken such that $\sigma_{k'+1,i} \neq 0$ but $\sigma_{k',i} = 0$. Note that by the property of the thresholding function $s^\ell_{\lambda,\gamma}(\cdot)$ we have that $s^\ell_{\lambda,\gamma}(\widetilde{x}) \neq 0 \Longrightarrow |s^\ell_{\lambda,\gamma}(\widetilde{x})| > \lambda_S$. This implies that $\|\boldsymbol{\sigma}_{k'+1} - \boldsymbol{\sigma}_{k'}\|_2^2 \geq \lambda_S^2$ for infinitely many k' values, which is a contradiction to the convergence: $f(X_{k+1}) - f(X_k) \to 0$. Thus the support of $\boldsymbol{\sigma}(X_k)$ converges, and necessarily after finitely many iterations — leading to the existence of an iteration number $\mathcal{K}^*$, after which the rank of X_k remains fixed. This completes the proof of the proposition.

Remark 1 If $\ell = 0$, the discontinuity of the thresholding operator $s_{\lambda,\gamma}(\cdot)$ (as demanded by Proposition 8) occurs for the MC+ penalty function as soon as $\gamma \leq 1$. For a general $\ell > 0$, discontinuity in $s^\ell_{\lambda,\gamma}(\cdot)$ occurs as soon as $\gamma \leq \frac{1}{\ell+1}$.

3.1.2 Subspace Stabilization

We study herein, the properties of the left and right singular subspaces associated with the sequence X_k. The *stabilization* of subspaces has important implications in the main bottleneck of the NC-IMPUTE algorithm, i.e., the SVD computations — we discuss this in further detail in Section 3.2. The study of singular subspace stabilization requires subtle analysis based on matrix perturbation theory (Stewart and Sun, 1990), since the left (and right) singular subspace, corresponding to the top r singular values of a matrix is not a continuous function of the matrix argument.

Towards this end, we first recall a standard notion of distance between two subspaces (with same dimension) in terms of canonical angles.

Definition 2 *Let $S_1 \in \mathbb{R}^{m\times\ell}$ and $S_2 \in \mathbb{R}^{m\times\ell}$ be two orthonormal matrices and let us define $S_1^{\perp}$ such that $[S_1, S_1^{\perp}]$ forms an orthonormal basis for $\mathbb{R}^m$. The canonical angles between these two subspaces denoted by the vector $\Theta(S_1, S_2)$ are defined as:*

$$\Theta(S_1, S_2) := \sin^{-1}\left(\sigma_1(X), \ldots, \sigma_\ell(X)\right),$$

where, $\sigma_i(X), i \leq \ell$ are the singular values of the matrix $X := (S_1^{\perp})' S_2$.

We now present a result regarding perturbation of singular subspaces, taken from Stewart and Sun (1990). Before stating the proposition, we introduce some notation. Let $U_1 \in \mathbb{R}^{m\times r_1}$ ($V_1 \in \mathbb{R}^{n\times r_1}$) denote a matrix of the r_1 left singular vectors (respectively, right) of a matrix A — with Σ_1 being a diagonal matrix of the corresponding top r_1 singular values. Similarly, we use the notation $\widetilde{U}_1, \widetilde{V}_1, \widetilde{\Sigma}_1$ to denote the triplet of left and right singular vectors and singular values (corresponding to the top r_1 singular values) for a matrix $\widetilde{A}$. We use the following matrices

$$R = A\widetilde{V}_1 - \widetilde{U}_1\widetilde{\Sigma}_1, \qquad Q = A'\widetilde{U}_1 - \widetilde{V}_1\widetilde{\Sigma}_1,$$

to measure a notion of proximity between A and $\widetilde{A}$. The distance between the left (and also right) singular subspaces (corresponding to the top r_1 singular values) of A and $\widetilde{A}$ may be measured by the following quantity:

$$\rho_{r_1}(A, \widetilde{A}) := \max\left\{\left\|\sin\left(\Theta(U_1, \widetilde{U}_1)\right)\right\|_2, \left\|\sin\left(\Theta(V_1, \widetilde{V}_1)\right)\right\|_2\right\}, \quad (31)$$

where, the notation $\|A\|_2$ denotes the spectral norm of A. With the above notations in place, we present the following proposition (Stewart and Sun, 1990) regarding the perturbation of singular subspaces of matrices.

Proposition 9 *Suppose there exists $\alpha, \delta > 0$ such that*

$$\min(\widetilde{\Sigma}_1) \geq \alpha + \delta, \quad \textit{and} \quad \max(\Sigma_2) \leq \alpha,$$

where, Σ_2 is a diagonal matrix with the remaining singular values of A. Then,

$$\rho_{r_1}(A, \widetilde{A}) \leq \max\left\{\|R\|_2, \|Q\|_2\right\}/\delta.$$

The above proposition informs us about the proximity of the left (and also right) singular subspaces across successive iterates X_k, as presented in the following proposition:

Proposition 10 *Suppose $\nu^{\dagger}(\ell) + \ell > 0$ and let*

$$\delta_{k,p} = \sigma_{p+1}(X_k) - \sigma_p(X_{k+1}),$$

for $1 \leq p \leq \min\{m, n\}$. If $\liminf_{k\to\infty} \delta_{k,p} > 0$ then $\rho_p(X_k, X_{k+1}) \to 0$ as $k \to \infty$.

Proof The proof is presented in Section 6.1.2.

Remark 2 Let the assumptions of Proposition 8 be in place – this implies that there exists an integer $\mathcal{K}^*$ such that

$$\operatorname{rank}(X_k) = r, \qquad \text{for all } k \geq \mathcal{K}^*.$$

Hence, in particular, there is a separation between $\sigma_r(X_k)$ and $\sigma_{r+1}(X_{k+1})$ for all k sufficiently large. This implies that $\rho_r(X_k, X_{k+1}) \to 0$ as $k \to \infty$, i.e., in words: the distance between the left (and right) singular subspaces corresponding to the top r singular values of X_k and X_{k+1} converges to zero, as $k \to \infty$.

3.1.3 Asymptotic Convergence.

We now investigate the asymptotic convergence properties of the sequence $X_k, k \geq 1$. Proposition 8 shows that under suitable assumptions, the sequence $\operatorname{rank}(X_k), k \geq 1$ converges. The existence of a limit point of X_k is guaranteed if the singular values of $\boldsymbol{\sigma}(X_k)$ remain bounded. It is not immediately clear whether the sequence $\boldsymbol{\sigma}(X_k)$ will remain bounded since several spectral penalty functions (like the MC+ penalty) are bounded[5]. We address herein, the existence of a limit point of the sequence $\boldsymbol{\sigma}(X_k)$, and hence the sequence X_k.

For the following proposition, we will assume that the concave penalty function $\sigma \mapsto P(\sigma; \lambda, \gamma)$ on $\sigma \geq 0$ is differentiable and the gradient is bounded.

Proposition 11 *Let $U_k \operatorname{diag}(\boldsymbol{\sigma}_k) V_k'$ denote the rank-reduced SVD of X_k. Let $\bar{U}_{m\times r}, \bar{V}_{m\times r}$ denote a limit point of the sequence $\{U_k, V_k\}, k \geq 1$, such that $(U_{n_k}, V_{n_k}) \to (\bar{U}, \bar{V})$ along a subsequence $n_k \to \infty$. Let $\bar{u}_i$ denote the ith column of $\bar{U}$ (and similarly for $\bar{v}_i, \bar{V}$) and let us denote $\bar{\Theta} = [vec(\mathcal{P}_{\Omega}(\bar{u}_1\bar{v}_1')), \ldots, vec(\mathcal{P}_{\Omega}(\bar{u}_r\bar{v}_r'))]$. We have the following:*

[5] Due to the boundedness of the penalty function, the boundedness of the objective function does not necessarily imply that the sequence $\boldsymbol{\sigma}(X_k)$ will remain bounded.

(a) If $\operatorname{rank}(\bar{\Theta}) = r$*, then the sequence* X_{n_k} *has a limit point which is a first order stationary point.*

(b) If $\lambda_{\min}(\bar{\Theta}'\bar{\Theta}) + \phi_P > 0$*, then the sequence* X_{n_k} *converges to a first order stationary point:* $\bar{X} = \bar{U}\operatorname{diag}(\bar{\boldsymbol{\sigma}})\bar{V}'$*, where* $\boldsymbol{\sigma}_{n_k} \to \bar{\boldsymbol{\sigma}}$.

Proof See Section 6.1.3

Proposition 8 describes sufficient conditions under which the rank of the sequence X_k stabilizes after finitely many iterations — it does *not* describe the boundedness of the sequence X_k, which is addressed in Proposition 11. Note that Proposition 11 does not imply that the rank of the sequence X_k stabilizes after finitely many iterations (recall that Proposition 11 does not assume that the thresholding operators are discontinuous, an assumption required by Proposition 8).

3.2 Computing the Thresholding Operators

The operator (24) requires computing a thresholded SVD of the matrix $\widetilde{X}_k$, as demonstrated by Proposition 1. The thresholded singular values $s^{\ell}_{\lambda,\gamma}(\cdot)$ as in (30) will have many zero coordinates due to the "sparsity promoting" nature of the concave penalty. Thus, computing the thresholding operator (24) will typically require performing a low-rank SVD on the matrix $\widetilde{X}_k$. While direct factorization based SVD methods can be used for smaller problems where $\min\{m, n\}$ is of the order of a thousand or so; for larger matrices, such methods become computationally prohibitive — we thus resort to iterative methods for computing low-rank SVDs for large scale problems. Algorithms such as the block power method; also known as block QR iterations, or those based on the Lanczos method (Golub and Van Loan, 1983) are quite effective in computing the top few singular value and vectors of a matrix A, especially when the operations of multiplying Ab_1 and $A'b_2$ (for vectors b_1, b_2 of matching dimensions) can be done efficiently. Indeed, such matrix-vector multiplications turn out to be quite computationally attractive for our problem, since the computational cost of multiplying $\widetilde{X}_k$ and $\widetilde{X}'_k$ with vectors of matching dimensions is quite low. This is due to the structure of:

$$\begin{aligned}\widetilde{X}_k &= \left(\mathcal{P}_{\Omega}(Y) + \mathcal{P}^{\perp}_{\Omega}(X_k) + \ell X_k\right)/(\ell+1) \\ &= \frac{1}{\ell+1}\underbrace{\mathcal{P}_{\Omega}(Y - X_k)}_{\text{Sparse}} + \underbrace{X_k}_{\text{Low-rank}},\end{aligned} \tag{32}$$

which admits a decomposition as the sum of a sparse matrix and a low-rank matrix[6]. Note that the sparse matrix has the same sparsity pattern as the observed indices Ω. Decomposition (32) is inspired by a similar decomposition that was exploited effectively in the algorithm SOFT-IMPUTE (Mazumder et al., 2010), where the authors use PROPACK (Larsen, 2004) to compute the low-rank SVDs. In this paper, we use the Alternating Least Squares (ALS)-stylized procedure, which computes a low-rank SVD by solving the following nonlinear optimization problem:

$$\min_{U_{m\times\tilde{r}}, V_{n\times\tilde{r}}} \quad \frac{1}{2}\|\widetilde{X}_k - UV'\|_F^2, \tag{33}$$

using alternating least squares—this is in fact, equivalent to the block power method (Golub and Van Loan, 1983), in computing a rank $\tilde{r}$ SVD of the matrix $\widetilde{X}_k$. Across the iterations of NC-IMPUTE, we pass the warm-start information in the U, V's obtained from a low-rank SVD of $\widetilde{X}_k$ to compute the low-rank SVD for $\widetilde{X}_{k+1}$. Empirically, this warm-start strategy is found to be significantly more advantageous than a black-box low-rank SVD stylized approach, as used in the SOFT-IMPUTE algorithm (for example), where, at every iteration, a new low-rank SVD is computed from scratch via PROPACK. This strategy quite naturally leads to a loss of useful information about the left and right singular vectors, which become closer to each other along the course of the SOFT-IMPUTE iterations (as formalized by Section 3.1.2). Using warm-start information across successive iterations (i.e., k values) leads to notable gains in computational speed (often reduces the total time to compute a family of solutions by orders of magnitude), when compared to black-box SVD stylized methods that do not rely on such warm-start strategies. This improvement is also supported by theory — the computational guarantee of block power iterations (Golub and Van Loan, 1983) states that the subspace spanned by the U matrix (in the factorization UV' in (33)) converges to that of the top $\tilde{r}$ left singular vectors at the rate: $C\gamma^q$, where, q denotes the number of power iterations, γ depends upon the ratio between the $\tilde{r}+1$ and $\tilde{r}$ singular values of the matrix $\widetilde{X}_k$; and C depends upon the distance between: the initial estimate of (the subspace spanned by) U and the left top-$\tilde{r}$ set of singular vectors of $\widetilde{X}_k$. The constant C is smaller with a good warm-start, when compared to a random initialization. A similar argument applies for the right set of singular vectors.

4 Numerical Experiments

In this section, we present a systematic experimental study of the statistical properties of estimators obtained from (3) for different choices of penalty functions. We perform our

[6] We note that it is not guaranteed that the X_k's will be of low-rank across the iterations of the algorithm for $k \geq 1$, even if they are eventually, for k sufficiently large. However, in the presence of warm-starts across (λ, γ) they are indeed, empirically, found to have low-rank as long as the regularization parameters are large enough to result in a small rank solution. Typically, as we have observed in our experiments, in the presence of warm-starts, the rank of X_k is found to remain low across all iterations.

experiments on a wide array of synthetic and real data instances. Recall that the majority of the algorithmic guarantees proved in Section 3 rely on the condition $\nu^\dagger(\ell)+\ell>0$. For MC+ penalty functions, it is straightforward to verify that $\nu^\dagger(0)>0$ as long as $\gamma\in(1,\infty]$. Hence we will use $\ell=0$ in NC-IMPUTE throughout this section.

4.1 Synthetic Examples

We study three different examples, where, for the true low-rank matrix $M=L\Phi R'$, we vary both the structure of the left and right singular vectors in L and R, as well as the sampling scheme used to obtain the observed entries in Ω. Our basic model is $Y_{ij}=M_{ij}+\varepsilon_{ij}$, where we observe entries $(i,j)\in\Omega$. We consider different types of missing patterns for Ω, and various signal-to-noise (SNR) ratios for the Gaussian error term ε, defined here to be:

$$\text{SNR}=\frac{\text{var}(\text{vec}(M))}{\text{var}(\text{vec}(\varepsilon))}.$$

Accordingly, the (standardized) training and test error for the model are defined as:

$$\text{Training Error}=\frac{\|\mathcal{P}_\Omega(Y-\hat{M})\|_F^2}{\|\mathcal{P}_\Omega(Y)\|_F^2},$$

$$\text{Test Error}=\frac{\|\mathcal{P}_\Omega^\perp(L\Phi R'-\hat{M})\|_F^2}{\|\mathcal{P}_\Omega^\perp(L\Phi R')\|_F^2},$$

where a value greater than one for the test error indicates that the computed estimate $\hat{M}$ does a worse job at estimating M than the zero solution, and the training error corresponds to the fraction of the error explained on the observed entries by the estimate $\hat{M}$ relative to the zero solution.

Example-A: In our first simulation setting, we use the model

$$Y_{m\times n}=L_{m\times r}\Phi_{r\times r}R'_{r\times n}+\varepsilon_{m\times n},$$

where L and R are matrices generated from the *random orthogonal model* (Candès and Recht, 2009), and the singular values $\Phi=\text{diag}(\phi_1,\ldots,\phi_r)$ are randomly selected as $\phi_1,\ldots,\phi_r\overset{\text{iid}}{\sim}\text{Uniform}(0,100)$. The set Ω is sampled uniformly at random. Recall that for this model, exact matrix completion in the noiseless setting is guaranteed as long as $|\Omega|\geq Cmr\log^4 m$, for some universal constant C (Recht, 2011). Under the noisy setting, Mazumder et al. (2010) show superior performance of nuclear norm regularization *vis-à-vis* other matrix recovery algorithms (Cai et al., 2010; Keshavan et al., 2010) in terms of achieving smaller test error. For the purposes herein, we fix $(m,n)=(800,400)$ and set the fraction of missing entries to $|\Omega^c|/mn=0.9$ and $|\Omega^c|/mn=0.95$.

Example-B: In our second setting, we also consider the model

$$Y_{m\times n}=L_{m\times r}\Phi_{r\times r}R'_{r\times n}+\varepsilon_{m\times n},$$

but we now select matrices L and R which do not satisfy the incoherence conditions required for full matrix recovery. Specifically, for the choices of $(m,n,r)=(800,400,10)$ and $|\Omega^c|/mn=0.9$, we select L and R to be block-diagonal matrices of the form

$$L=\text{diag}(L_1,\ldots,L_5),\qquad R=\text{diag}(R_1,\ldots,R_5),$$

where $L_i\in\mathbb{R}^{160\times 2}$ and $R_i\in\mathbb{R}^{80\times 2}$, $i=1,\ldots,5$, are random matrices with scaled Gaussian entries. The singular values are again sampled as $\phi_1,\ldots,\phi_r\overset{\text{iid}}{\sim}\text{Uniform}(0,100)$ with Ω being uniformly random over the set of indices. For this model, successful matrix completion is not guaranteed even for the noiseless problem with the nuclear norm relaxation, as the left and right singular vectors are not sufficiently spread. We observe the usefulness of the nonconvex regularized estimators in this regime, in our experimental results.

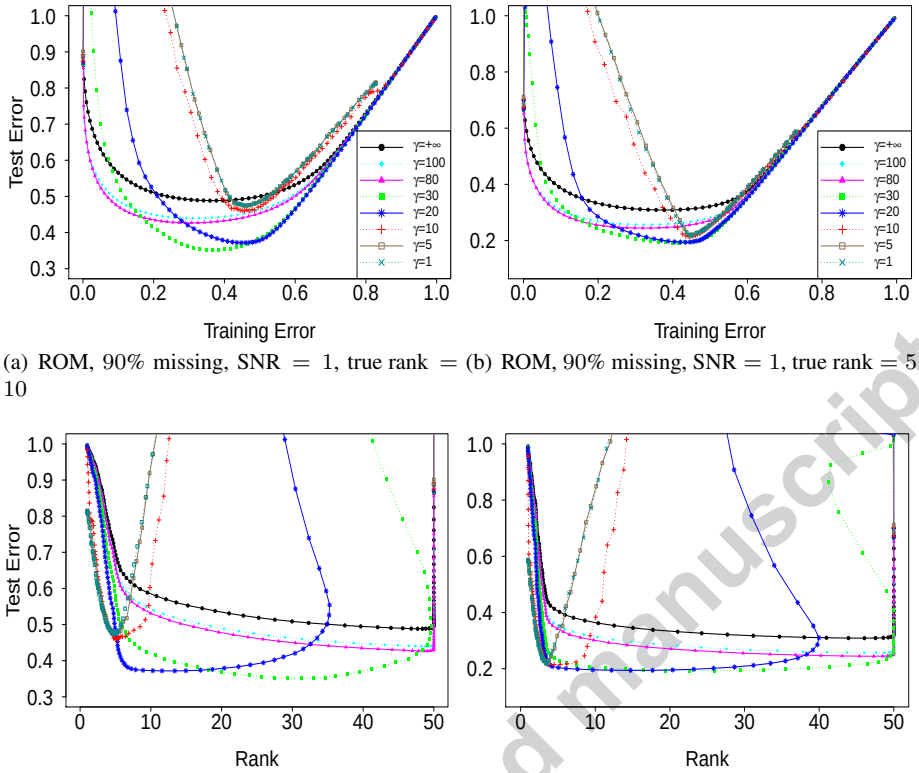
Example-C: In our third simulation setting, for the choice of $(m,n,r)=(100,100,10)$, we also generate $Y_{m\times n}=L_{m\times r}\Phi_{r\times r}R'_{r\times n}+\varepsilon_{m\times n}$ from the random orthogonal model as in our first setting, but we now allow the observed entries in Ω to follow a nonuniform sampling scheme. In particular, we fix $\Omega^c=\{1\leq i,j\leq 100:1\leq i\leq 50,51\leq j\leq 100\}$ so that

$$\mathcal{P}_\Omega(Y)=\begin{bmatrix}Y_{11} & 0\\ Y_{21} & Y_{22}\end{bmatrix}\quad\text{where,}\quad Y=\begin{bmatrix}Y_{11} & Y_{12}\\ Y_{21} & Y_{22}\end{bmatrix},$$

with the fraction of missing entries thus being $|\Omega^c|/mn=0.25$. This is again a challenging simulation setting in which both the uniform (Candès and Recht, 2009) and independent (Chen et al., 2014) sampling scheme assumptions in Ω are violated. Our aim again is to explore whether the nonconvex MC+ family is able to outperform nuclear norm regularization in this regime.

For all three settings above, we choose a 100×25 grid of (λ,γ) values as follows. In each simulation instance we fix $\lambda_1=\|\mathcal{P}_\Omega(Y)\|_2$, the smallest value of λ for which the nuclear norm regularized solution is zero, and set $\lambda_{100}=0.001\cdot\lambda_1$. Keeping in mind that NC-IMPUTE benefits greatly from using warm starts, we construct an equally spaced sequence of 100 values of λ decreasing from λ_1 to λ_{100}. We choose 25 γ-values in a logarithmic grid from 5000 to 1.1. The results displayed in Figures 4 – 6 show averages of training and test errors, as well as recovered ranks of the solution matrix $\hat{M}_{\lambda,\gamma}$ for the values of (λ,γ), taken over 50 simulations under all three problem instances. The plots including rank reveal how effective the MC+ family is at recovering the true rank while minimizing prediction error. Throughout the simulations we keep an upper bound of the operating rank as 50.

Example-A (Low SNR, less missing entries)



(a) ROM, 90% missing, SNR $=$ 1, true rank $=$ 10 (b) ROM, 90% missing, SNR $=$ 1, true rank $=$ 5

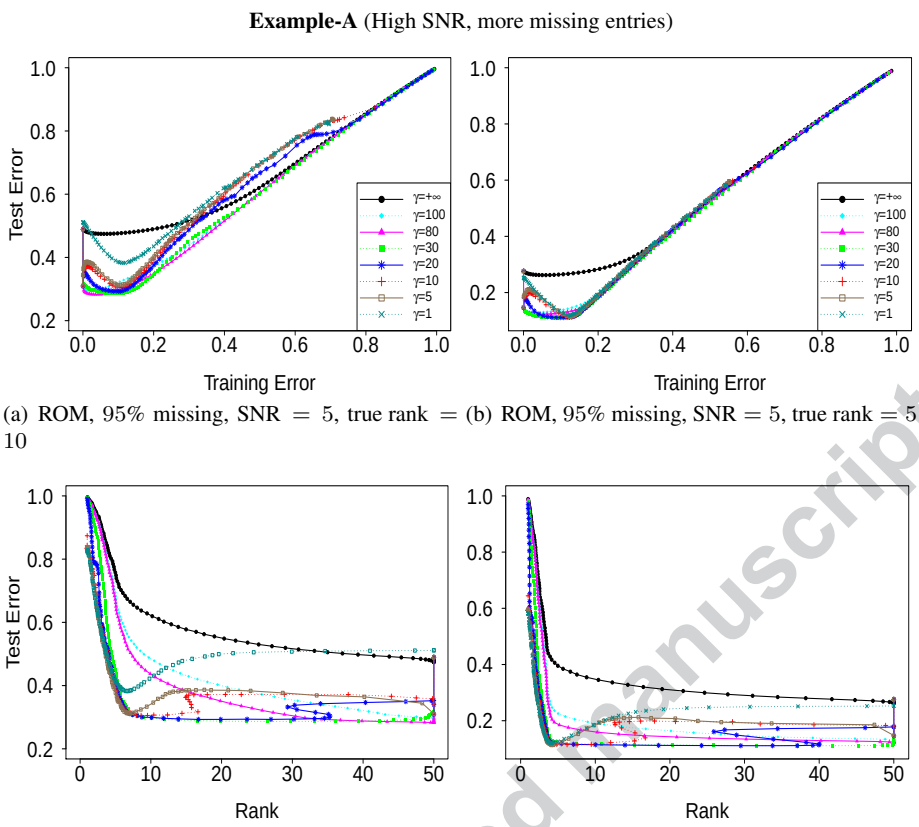
(c) ROM, 90% missing, SNR $=$ 1, true rank $=$ 10 (d) ROM, 90% missing, SNR $=$ 1, true rank $=$ 5

Fig. 4 (Color online) Random Orthogonal Model (ROM) simulations with SNR $= 1$. The choice $\gamma = +\infty$ refers to nuclear norm regularization as provided by the SOFT-IMPUTE algorithm. We also include the choice $\gamma = 1$ to represent the rank regularized approach. The least nonconvex alternatives at $\gamma = 100$ and $\gamma = 80$ behave similarly to nuclear norm, although with better prediction performance. The choices of $\gamma = 1, 5, 10$ result in excessively aggressive fitting behavior for the true rank $= 10$ case, but improve significantly in prediction error and recovering the true rank in the sparser true rank $= 5$ setting. In both scenarios, the intermediate models with $\gamma = 30$ and $\gamma = 20$ fare the best, with the former achieving the smallest prediction error, while the latter estimates the actual rank of the matrix. Values of test error larger than one are not displayed in the figure.

4.1.1 Discussion of Experimental Results

We devote Figures 4 and 5 to analyze the simpler random orthogonal model (Example-A), leaving the more challenging coherent and nonuniform sampling settings (Example-B and Example-C) for Figure 6. In each case, the captions detail the results which we summarize here. The noise is quite high in Figure 4 with SNR $= 1$ and 90% of the entries missing in both displayed settings, while the model complexity decreases from a true rank of 10 to 5. The underlying true ranks remain the same in Figure 5, but the noise level has decreased to SNR $= 5$ with the missing entries increasing to 95%. For each model setting considered, all nonconvex methods from the MC+ family outperform nuclear norm regularization in terms of prediction performance, while members of the MC+ family with smaller values of γ are better at estimating the correct rank. The choices of $\gamma = 30$ and $\gamma = 20$ have the best performance in Figure 4 (best prediction errors around the true ranks), while more nonconvex alternatives fare better in the high-sparsity, low-noise setting of Figure 5. In both figures, the performance of nuclear norm regularization is somewhat similar to the least nonconvex alternative displayed at $\gamma = 100$, however, the bias induced in the estimation of the singular values of the low-rank matrix M leads to the worst bias-variance trade-off among all training versus test error plots for the settings considered.

While the nuclear norm relaxation provides a good convex approximation for the rank of a matrix (Recht et al., 2010), these examples show that nonconvex regularization methods provide a superior mechanism for rank estimation. This is reminiscent of the performance of the MC+ penalty in the context of variable selection within high-dimensional sparse regression models. Although the ℓ_1 penalty function represents the best convex approximation to the ℓ_0 penalty, the gap bridged by the nonconvex MC+ penalty family pro-



(a) ROM, 95% missing, SNR = 5, true rank = 10 (b) ROM, 95% missing, SNR = 5, true rank = 5

(c) ROM, 95% missing, SNR = 5, true rank = 10 (d) ROM, 95% missing, SNR = 5, true rank = 5

Fig. 5 (Color online) Random Orthogonal Model (ROM) simulations with SNR = 5. The benefits of nonconvex regularization are more evident in this high-sparsity, high-missingness scenario. While the $\gamma = 100$ and $\gamma = 80$ models distance themselves more from nuclear norm, the remaining members of the MC+ family essentially minimize prediction error while correctly estimating the true rank. This is especially true in panel (d), where the best predictive performance of the model $\gamma = 5$ at the correct rank is achieved under a low-rank truth and high SNR setting.

vides a better basis for model selection, and hence rank estimation in the low-rank matrix completion setting.

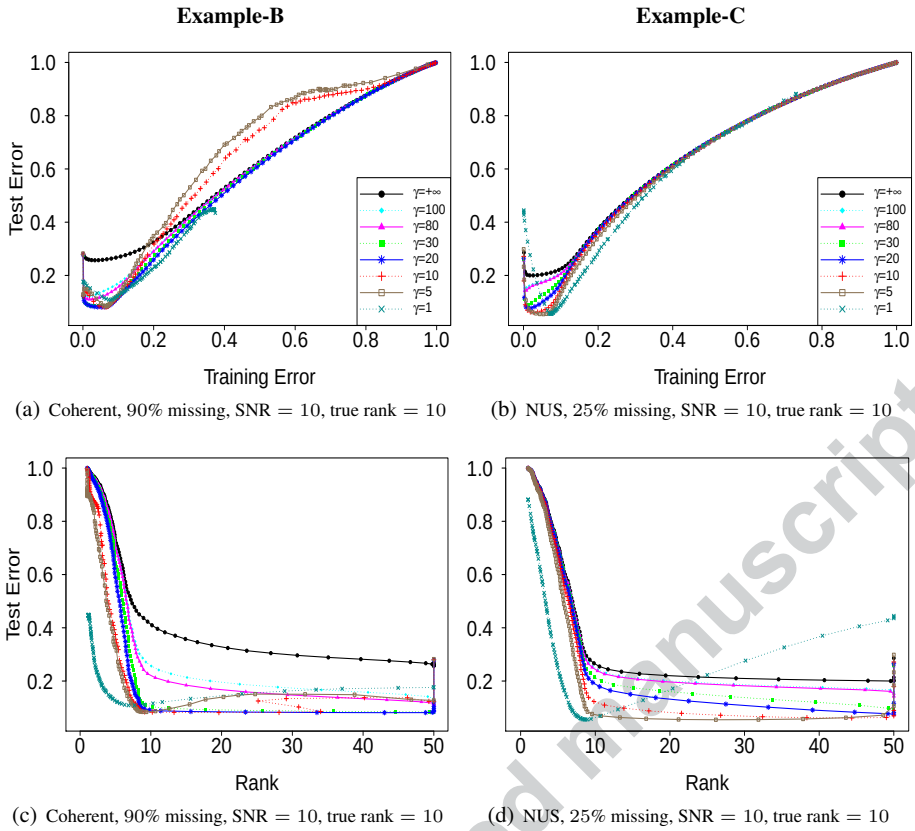
For the coherent and nonuniform sampling settings of Figure 6, we choose the small noise scenario SNR = 10 in order to favor all considered models. Despite the absence of any theoretical guarantees for successful matrix recovery, the nuclear norm regularization approach achieves a relatively small prediction error in all displayed instances. Nevertheless, the nonconvex MC+ family of penalties seems empirically more adept at overcoming the limitations of nuclear norm penalizated matrix completion in these challenging simulation settings. In particular, the most aggressive nonconvex fitting behavior at $\gamma = 5$ achieves excellent prediction performance in the nonuniform sampling setting while correctly estimating the true rank of the coherent model.

4.2 Real Data Examples: MovieLens and Netflix datasets

We now use the real world recommendation system datasets `ml100k` and `ml1m` provided by MovieLens[7], as well as the famous Netflix competition data to compare the usual nuclear norm approach with the MC+ regularizers. The dataset `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 datasets, for all regularization methods considered, a random subset of 80% of the ratings were used for training purposes; the remaining were used as the test set.

We also choose a similar 100×25 grid of (λ, γ) values, but for each value of λ in the decreasing sequence, we use an "operating rank" threshold somewhat larger than the rank of the previous solution, with the goal of always obtaining solution ranks smaller than the operating threshold. Following the approach of Hastie et al. (2016), we perform row and col-

[7] Available at `http://grouplens.org/datasets/movielens/`



(a) Coherent, 90% missing, SNR = 10, true rank = 10 (b) NUS, 25% missing, SNR = 10, true rank = 10

(c) Coherent, 90% missing, SNR = 10, true rank = 10 (d) NUS, 25% missing, SNR = 10, true rank = 10

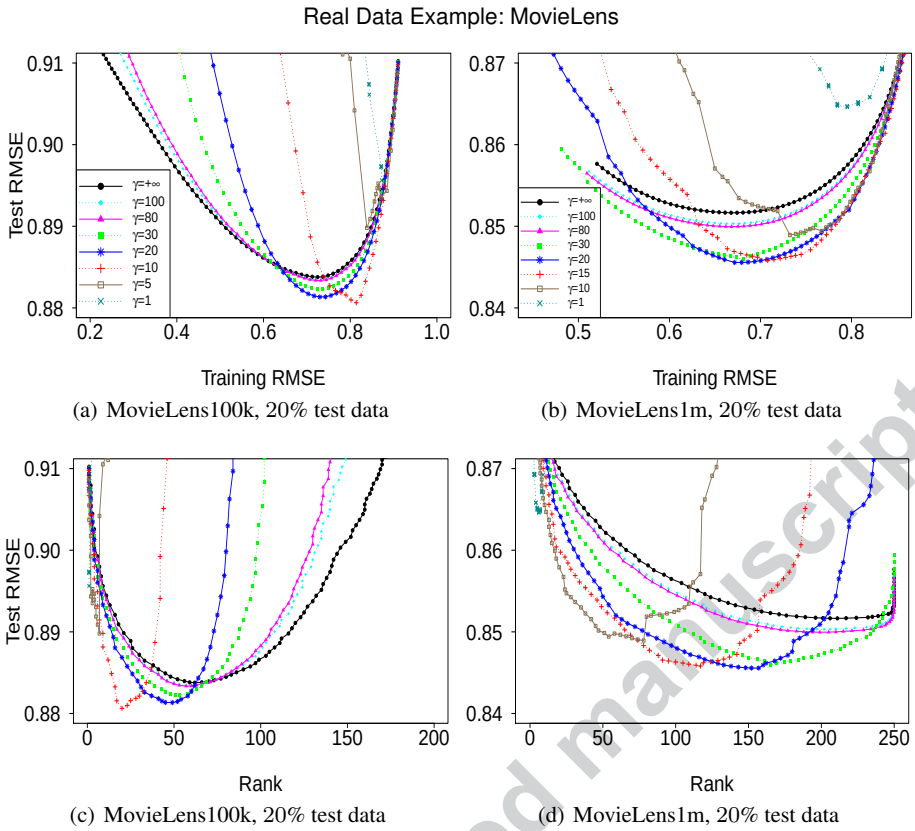
Fig. 6 (Color online) Coherent and Nonuniform Sampling (NUS) simulations with SNR = 10. nonconvex regularization also proves to be a successful strategy in these challenging scenarios, particularly in the nonuniform sampling setting where the MC+ family exhibits a monotone decrease in prediction error as γ approaches 1. Again, the model $\gamma = 5$ estimates the correct rank under high SNR settings. Although nuclear norm achieves a relatively small prediction error, compared with previous simulation settings, the MC+ family still provides a superior and more robust mechanism for regularization.

umn centering of the corresponding (incomplete) data matrices as a preprocessing step.

Figure 7 compares the performance of nuclear norm regularization with the MC+ family of penalties on these datasets, in terms of the prediction error (RMSE) obtained from the left out portion of the data. While the fitting behavior at $\gamma = 5$ seems to be overly aggressive in these instances, the choice $\gamma = 10$ achieves the best test set RMSE with a minimum solution rank of 20 for the `ml100k` data. With a higher test RMSE, nuclear norm regularization achieves its minimum with a less parsimonious model of rank 62. Similar results hold for the `ml1m` data, where the model $\gamma = 15$ achieves near optimal test RMSE at a solution rank of 115, while the best estimation accuracy of SOFT-IMPUTE occurs for ranks well over 200.

The Netflix competition data consists of $100,480,507$ ratings from $480,189$ users on $17,770$ movies. A designated probe set, a subset of $1,408,395$ of these ratings, was distributed to participants for calibration purposes, leaving $99,072,112$ for training. We did not consider the probe set as part of this numerical experiment, instead choosing $1,500,000$ randomly selected entries as test data with the remaining $97,572,112$ used for training purposes. Similar to the MovieLens data, we select a 20×25 grid of (λ, γ) values which adaptively chooses an operating rank threshold, and also remove row and columns means for prediction purposes.

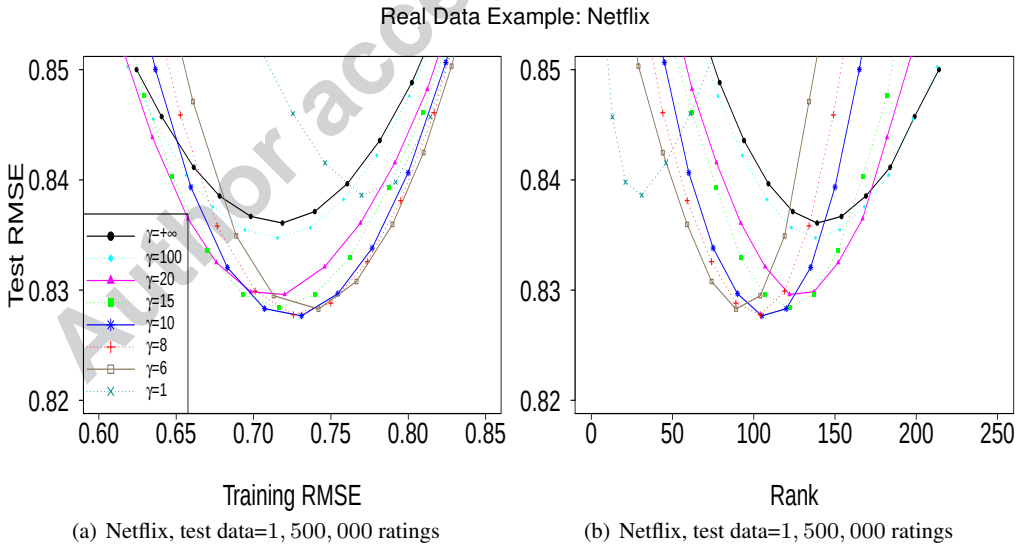
As shown in Figure 8, the MC+ family again yields better prediction performance under more parsimonious models. On average, and for a convergence tolerance of 0.001 in Algorithm 1, the sequence of twenty models took under 10.5 hours of computing on an Intel E5-2650L cluster with 2.6 GHz processor. We note that our main goal here is to show the feasibility of applying NC-IMPUTE to the MC+ family on a Netflix sized dataset, and further reductions in computation time may be possible with specialized implementations. It seems that using a family of nonconvex penalties leads to models with better statistical properties, when compared to the nuclear norm regularized problem and the rank constrained problem (obtained via HARD-IMPUTE, for example).



(a) MovieLens100k, 20% test data (b) MovieLens1m, 20% test data

(c) MovieLens100k, 20% test data (d) MovieLens1m, 20% test data

Fig. 7 (Color online) MovieLens 100k and 1m data. For each value of λ in the solution path, an operating rank threshold (capped at 250) larger than the rank of the previous solution was employed.



(a) Netflix, test data=$1,500,000$ ratings (b) Netflix, test data=$1,500,000$ ratings

Fig. 8 (Color online) Netflix competition data. The model $\gamma = 10$ achieves optimal test set RMSE of 0.8276 for a solution rank of 105.

5 Conclusions and Discussions

In this paper we present a computational study for the noisy matrix completion problem with nonconvex spectral penalties — we consider a family of spectral penalties that bridge the convex nuclear norm penalty and the rank penalty, leading to a family of estimators with varying degrees of shrinkage and nonconvexity. We propose NC-IMPUTE— an al-

gorithm that appropriately modifies and enhances the EM-stylized procedure SOFT-IMPUTE (Mazumder et al., 2010), to compute a two dimensional family of solutions with specialized warm-start strategies. The main computational bottleneck of our algorithm is a low-rank SVD of a structured matrix, which is performed using a block QR stylized strategy that makes effective use of singular subspace warm-start information across iterations. We discuss computational guarantees of our algorithm, including a finite time complexity analysis to a first order stationary point. We present a systematic study of various statistical and structural properties of spectral thresholding functions, which form a main building block in our algorithm. We demonstrate the impressive gains in statistical properties of our framework on a wide array of synthetic and real datasets. The current work leaves open several important directions for future research.

- *Statistical guarantees for the nonconvex method*. In addition to the comprehensive algorithmic analysis presented in the paper, it is of great importance to establish statistical theories such as estimation error bounds to shed new lights on the empirical success of the proposed nonconvex method. In particular, given that our algorithm returns stationary points, it would be interesting to obtain statistical errors for these local optima. In fact, such type of results have been derived for regularized M-estimators in some general multivariate analysis settings when the penalty function is separable (Loh and Wainwright, 2015). A notable difficulty in the matrix completion problem is that the penalty is imposed on the singular values and is hence a non-separable function of the matrix, which will require more delicate analyses. Moreover, we should point out that statistical analysis of nonconvex optimization methods for matrix completion has been actively investigated in recent years. See the works surveyed in the last paragraph of Introduction and Chi et al. (2019) for a thorough review. However, the nonconvex methods studied in this line of research are exclusively based on low-rank matrix factorization formulation, and the regularization from these methods is less general than the ones in our method. The nonconvexity of the former methods is largely due to the matrix factorization which enables the reduction of memory and computation costs, while the nonconvexity of our method arises from the nonconvex penalties that aim to attenuate the bias.
- *Sharp comparison between nonconvex and convex methods*. Referring to both simulation study and real data analysis in Section 4, we observe that the value of γ leading to the optimal matrix completion performance lies between 5 and 30. Recall that the penalty parameter γ in the MC+ penalties controls the amount of nonconvexity in the regularization. As γ decreases from ∞ down to 1, the penalty behaves closer to ℓ_0 and farther away from ℓ_1. Hence, the empirical results in Section 4 demonstrate that neither the convex approach ($\gamma = +\infty$) nor the most aggressive nonconvex one ($\gamma = 1$) is the optimal choice. This phenomenon is in fact a manifestation of bias-variance tradeoff. A smaller value of γ brings more "nonconvexity" to the regularization and hence induces less bias as expected. On the other hand, more "nonconvexity" means more aggressiveness in the selection of low rank matrices and thus results in larger variance. Consequently, for a given level of signal-to-noise ratio in the observations, the optimal γ is the one that strikes the best balance between bias and variance. This point of view lends further support to our proposed nonconvex method which incorporates the entire family of nonconvex penalties instead of some particular instantiations. Recent works by a subset of the authors (Zheng et al., 2017; Wang et al., 2019) have given sharp theoretical characterizations of such a phenomenon in the high-dimensional sparse regression and variable selection problems. The results there reveal that among the ℓ_q penalties for $q \in [0, 2]$, as the signal-to-noise ratio (SNR) decreases, the optimal value of q will move from 0 towards 2. See also the work of Hazimeh and Mazumder (2019); Mazumder et al. (2017) for similar observations regarding the overfitting of ℓ_0-based estimators for low SNR regimes. For the MC+ penalties in the matrix completion problem, γ plays a similar role as q does in the regression problem. It would be of great interest to derive analogue theories for the matrix completion problem and establish a sharp characterization of the proposed nonconvex method.

6 Appendix

6.1 Additional Technical Material

Lemma 1 *(Marchenko-Pastur law (Bai and Silverstein, 2010)). Let $X \in \mathbb{R}^{m\times n}$, where X_{ij} are iid with $\mathbb{E}(X_{ij}) = 0, \mathbb{E}(X_{ij}^2) = 1$, and $m > n$. Let $\lambda_1 \le \lambda_2 \le \cdots \le \lambda_n$ be the eigenvalues of $Q_m = \frac{1}{m}X'X$. 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

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

where μ is a deterministic measure with density

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

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

6.1.1 Proof of Proposition 5.

Proof In the following proof, we make use of the notation: $\Theta_1(\cdot)$ and $\Theta_2(\cdot)$, defined as follows. For two positive sequences a_k and b_k, we say $a_k = \Theta_2(b_k)$ if there exists a constant $c > 0$ such that $a_k \geq cb_k$ and we say $a_k = \Theta_1(b_k)$, whenever, $a_k = \Theta_2(b_k)$ and $b_k = \Theta_2(a_k)$.

We first consider the case $\lambda_n = \Theta_1(\sqrt{m})$ For simplicity, we assume $\lambda_n = \zeta\sqrt{m}$ for some constant $\zeta > 0$. Denote $df(S_{\lambda_n,\gamma}(Z)) = D_{\lambda_n,\gamma}$, and use $\mathcal{T}_{t_1,t_2}$ to represent

$$\frac{\sqrt{mt_1}s_{\lambda_n,\gamma}(\sqrt{mt_1}) - \sqrt{mt_2}s_{\lambda_n,\gamma}(\sqrt{mt_2})}{mt_1 - mt_2}\mathbb{1}(t_1 \neq t_2).$$

Adopting the notation from Lemma 1, it is not hard to verify that

$$\begin{aligned} D_{\lambda_n,\gamma} =& n\mathbb{E}_{\mu_n}\left\{s'_{\lambda_n,\gamma}(\sqrt{mt_1}) + |m-n|\frac{s_{\lambda_n,\gamma}(\sqrt{mt_1})}{\sqrt{mt_1}}\right\} \\ &+ n^2\mathbb{E}_{\mu_n}(\mathcal{T}_{t_1,t_2}), \end{aligned}$$

where $t_1, t_2 \overset{\text{iid}}{\sim} \mu_n$. A quick check of the relation between $s_{\lambda_n,\gamma}$ and $g_{\zeta,\gamma}$ yields

$$\begin{aligned} \frac{D_{\lambda_n,\gamma}}{mn} =& \frac{1}{m}\mathbb{E}_{\mu_n}s'_{\lambda_n,\gamma}(\sqrt{mt_1}) + \left(1-\frac{n}{m}\right)\mathbb{E}_{\mu_n}g_{\zeta,\gamma}(t_1) + \\ & \frac{n}{m}\mathbb{E}_{\mu_n}\left\{\frac{t_1g_{\zeta,\gamma}(t_1) - t_2g_{\zeta,\gamma}(t_2)}{t_1-t_2}\mathbb{1}(t_1 \neq t_2)\right\}. \end{aligned}$$

Due to the Lipschitz continuity of the functions $s_{\lambda_n,\gamma}(x)$ and $xg_{\zeta,\gamma}(x)$, we obtain

$$\left|\frac{D_{\lambda_n,\gamma}}{mn}\right| \leq \frac{\gamma}{m(\gamma-1)} + \left(1-\frac{n}{m}\right) + \frac{n}{m}\left(\frac{2\gamma-1}{2\gamma-2}\right).$$

Hence, there exists a positive constant C_α, such that for sufficiently large n,

$$\left|\frac{D_{\lambda_n,\gamma}}{mn}\right| \leq C_\alpha, \quad a.s.$$

Let T_1, T_2 be two independent random variables generated from the Marchenko-Pastur distribution μ. If we can show

$$\begin{aligned} &\frac{D_{\lambda_n,\gamma}}{mn} \overset{a.s.}{\to} \\ &(1-\alpha)\mathbb{E}g_{\zeta,\gamma}(T_1) + \alpha\mathbb{E}\left(\frac{T_1g_{\zeta,\gamma}(T_1) - T_2g_{\zeta,\gamma}(T_2)}{T_1-T_2}\right), \end{aligned}$$

then by the Dominated Convergence Theorem (DCT), we conclude the proof in the $\lambda_n = \Theta_1(\sqrt{m})$ regime. Note immediately that

$$\frac{1}{m}\mathbb{E}_{\mu_n}s'_{\lambda_n,\gamma}(\sqrt{mt_1}) \to 0 \quad a.s. \tag{34}$$

Moreover, given that $g_{\zeta,\gamma}(\cdot)$ is bounded and continuous, the Marchenko-Pastur theorem in Lemma 1 implies

$$\left(1-\frac{n}{m}\right)\mathbb{E}_{\mu_n}g_{\zeta,\gamma}(t_1) \to (1-\alpha)\mathbb{E}_{\mu}g_{\zeta,\gamma}(T_1) \quad a.s. \tag{35}$$

Since $(t_1, t_2) \overset{d}{\to} (T_1, T_2)$, and the discontinuity set of the function $\frac{t_1g_{\zeta,\gamma}(t_1)-t_2g_{\zeta,\gamma}(t_2)}{t_1-t_2}\mathbb{1}(t_1 \neq t_2)$ has zero probability under the measure induced by (T_1, T_2), by the continuous mapping theorem,

$$\begin{aligned} &\frac{t_1g_{\zeta,\gamma}(t_1) - t_2g_{\zeta,\gamma}(t_2)}{t_1-t_2}\mathbb{1}(t_1 \neq t_2) \overset{d}{\to} \\ &\frac{T_1g_{\zeta,\gamma}(T_1) - T_2g_{\zeta,\gamma}(T_2)}{T_1-T_2}\mathbb{1}(T_1 \neq T_2) \quad \text{as } n \to \infty. \end{aligned}$$

Also, due to the boundedness of $\frac{t_1g_{\zeta,\gamma}(t_1)-t_2g_{\zeta,\gamma}(t_2)}{t_1-t_2}\mathbb{1}(t_1 \neq t_2)$, it holds that

$$\begin{aligned} &\mathbb{E}_{\mu_n}\left\{\frac{t_1g_{\zeta,\gamma}(t_1) - t_2g_{\zeta,\gamma}(t_2)}{t_1-t_2}\mathbb{1}(t_1 \neq t_2)\right\} \overset{a.s.}{\to} \\ &\mathbb{E}_{\mu}\left\{\frac{T_1g_{\zeta,\gamma}(T_1) - T_2g_{\zeta,\gamma}(T_2)}{T_1-T_2}\mathbb{1}(T_1 \neq T_2)\right\}. \end{aligned} \tag{36}$$

Combining (34) – (36) completes the proof for the $\lambda_n = \Theta_1(\sqrt{m})$ case.

When $\lambda_n = o(\sqrt{m})$, we can readily see that

$$\mathbb{E}_{\mu_n}\mathbb{1}(\sqrt{mt_1} \geq \lambda_n\gamma) \to 1, a.s.$$

Using that both $\frac{s_{\lambda_n,\gamma}(\sqrt{mt_1})}{\sqrt{mt_1}}$ and $\mathcal{T}_{t_1,t_2}$ are bounded, we have, almost surely

$$\begin{aligned} \mathbb{E}_{\mu_n}\frac{s_{\lambda_n,\gamma}(\sqrt{mt_1})}{\sqrt{mt_1}} =& \mathbb{E}_{\mu_n}\mathbb{1}(\sqrt{mt_1} \geq \lambda_n\gamma) + \\ & \mathbb{E}_{\mu_n}\left\{\frac{s_{\lambda_n,\gamma}(\sqrt{mt_1})}{\sqrt{mt_1}}\mathbb{1}(\sqrt{mt_1} < \lambda_n\gamma)\right\} \to 1 \end{aligned}$$

and

$$\begin{aligned} \mathbb{E}_{\mu_n}(\mathcal{T}_{t_1,t_2}) =& \mathbb{E}_{\mu_n}\mathbb{1}(\sqrt{mt_1} \geq \lambda_n\gamma)\mathbb{1}(\sqrt{mt_2} \geq \lambda_n\gamma) + o(1) \\ & \to 1. \end{aligned}$$

Invoking DCT completes the proof. Similar arguments hold for the case $\lambda_n = \Theta_2(\sqrt{m})$.

6.1.2 Proof of Proposition 10

Proof Observe that R as defined in Proposition 9 can be written as:

$$R = \quad \widetilde{A}\widetilde{V}_1 - \widetilde{U}_1\widetilde{\Sigma}_1 + (A-\widetilde{A})\widetilde{V}_1 = \quad (A-\widetilde{A})\widetilde{V}_1 \tag{37}$$

where, above we have used the fact that $\widetilde{A}\widetilde{V}_1 = \widetilde{U}_1\widetilde{\Sigma}_1$, which follows from the definition of the SVD of $\widetilde{A}$. By a simple inequality it follows that

$$\|R\|_2 \leq \|(A-\widetilde{A})\|_2\|\widetilde{V}_1\|_2 = \|(A-\widetilde{A})\|_2, \tag{38}$$

where we have used the the fact that $\|\widetilde{V}_1\|_2 = 1$. Similarly, we have an analogous result for Q:

$$\|Q\|_2 \leq \|(A-\widetilde{A})\|_2\|\widetilde{U}_1\|_2 = \|(A-\widetilde{A})\|_2. \tag{39}$$

Note that (38) and (39) together imply that if $\|\widetilde{A} - A\|_2$ is small, then so are $\|R\|_2, \|Q\|_2$.

We now apply (31) (Proposition 9) with $A = X_k$ and $\widetilde{A} = X_{k+1}$ and $r_1 = p$, to arrive at the proof of Proposition 10.

6.1.3 *Proof of Proposition 11*

Proof Proof of Part (a):
Let us write the stationary conditions for every update:

$$X_{k+1} = \arg\min_X \; F_\ell(X; X_k).$$

We set the subdifferential of the map $X \mapsto F_\ell(X; X_k)$ to zero at $X = X_{k+1}$:

$$\begin{aligned}&\big(X_{k+1} - \big(\mathcal{P}_\Omega(Y) + \mathcal{P}_\Omega^\perp(X_k)\big)\big)\\ &+ \ell(X_{k+1} - X_k) + U_{k+1}\nabla_{k+1}V'_{k+1} = 0,\end{aligned} \tag{40}$$

where $X_{k+1} = U_{k+1}\mathrm{diag}(\boldsymbol{\sigma}_{k+1})V'_{k+1}$ is the SVD of X_{k+1}. Note that the term: $U_{k+1}\nabla_{k+1}V'_{k+1}$ in (40), is a subdifferential (Lewis, 1995) of the spectral function:

$$X \mapsto \sum_i P(\sigma_i(X); \lambda, \gamma),$$

where ∇_{k+1} is a diagonal matrix with the ith diagonal entry being a derivative of the map $\sigma_i \mapsto P(\sigma_i; \lambda, \gamma)$ (on $\sigma_i \geq 0$), denoted by $\partial P(\sigma_{k+1,i}; \lambda, \gamma)/\partial\sigma_i$ for all i. Note that (40) can be rewritten as:

$$\begin{aligned}&\mathcal{P}_\Omega(X_{k+1}) - \mathcal{P}_\Omega(Y) + U_{k+1}\nabla_{k+1}V'_{k+1}\\ &+ \underbrace{\big(\mathcal{P}_\Omega^\perp(X_{k+1} - X_k) + \ell(X_{k+1} - X_k)\big)}_{(a)} = 0.\end{aligned}$$

As $k \to \infty$, term (a) converges to zero (See Proposition 7) and thus, we have:

$$\mathcal{P}_\Omega(X_{k+1}) - \mathcal{P}_\Omega(Y) + U_{k+1}\nabla_{k+1}V'_{k+1} \to 0.$$

Let us denote the ith column of U_k by $u_{k,i}$, and use a similar notation for V_k and $v_{k,i}$. Let r_{k+1} denote the rank of X_{k+1}. Hence, we have:

$$\sum_{i=1}^{r_{k+1}} \sigma_{k+1,i}\mathcal{P}_\Omega(u_{k+1,i}v'_{k+1,i}) - \mathcal{P}_\Omega(Y) + U_{k+1}\nabla_{k+1}V'_{k+1} \to 0.$$

Multiplying the left and right hand sides of the above by $u'_{k+1,j}$ and $v_{k+1,j}$, we have the following:

$$\begin{aligned}&\sum_{i=1}^{r_{k+1}} \sigma_{k+1,i}u'_{k+1,j}\mathcal{P}_\Omega(u_{k+1,i}v'_{k+1,i})v_{k+1,j}\\ &- u'_{k+1,j}\mathcal{P}_\Omega(Y)v_{k+1,j} + \nabla_{k+1,j} \to 0,\end{aligned}$$

for $j = 1, \ldots, r_{k+1}$. Let $\{\bar{U}, \bar{V}\}$ denote a limit point of the sequence $\{U_k, V_k\}$ (which exists since the sequence is bounded); and let r be the rank of $\bar{U}$ and $\bar{V}$. Let us now study the following equations[8]:

$$\sum_{i=1}^{r} \bar{\sigma}_j\bar{u}'_j\mathcal{P}_\Omega(\bar{u}_i\bar{v}'_i)\bar{v}_j - \bar{u}'_j\mathcal{P}_\Omega(Y)\bar{v}_j + \bar{\nabla}_j = 0, \quad j = 1, \ldots, r. \tag{41}$$

Using the notation $\bar{\theta}_j = \mathrm{vec}\left(\mathcal{P}_\Omega(\bar{u}_j\bar{v}'_j)\right)$ and $\bar{y} = \mathrm{vec}(\mathcal{P}_\Omega(Y))$, we note that (41) are the first order stationary conditions for a point $\bar{\boldsymbol{\sigma}}$ for the following penalized regression problem:

$$\min_{\boldsymbol{\sigma}} \;\; \frac{1}{2}\|\sum_{j=1}^{r} \sigma_j\bar{\theta}_j - \bar{y}\|_2^2 + \sum_{j=1}^{r} P(\sigma_j; \lambda, \gamma), \tag{42}$$

with $\boldsymbol{\sigma} \geq \mathbf{0}$.

If the matrix $\bar{\Theta} = [\bar{\theta}_1, \ldots, \bar{\theta}_r]$ (note that $\bar{\Theta} \in \mathbb{R}^{mn \times r}$) has rank r, then any $\boldsymbol{\sigma}$ that satisfies (41) is finite — in particular, the sequence $\boldsymbol{\sigma}_k$ is bounded and has a limit point: $\bar{\boldsymbol{\sigma}}$ which satisfies the first order stationary condition (41).

Proof of Part (b):
Furthermore, if we assume that

$$\lambda_{\min}(\bar{\Theta}'\bar{\Theta}) + \phi_P > 0,$$

then (42) admits a unique solution $\bar{\boldsymbol{\sigma}}$, which implies that $\boldsymbol{\sigma}_k$ has a unique limit point, and hence the sequence $\boldsymbol{\sigma}_k$ necessarily converges.

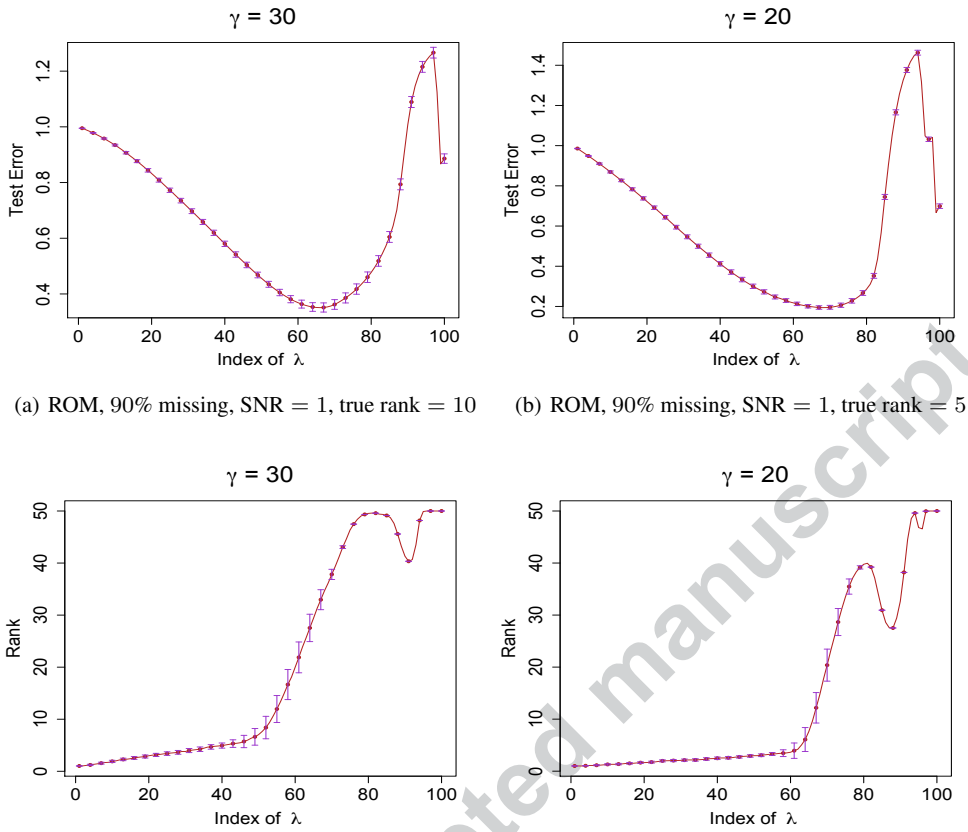
6.2 Additional Simulation Results

This section contains additional numerical results from the simulation study in Section 4.1.

– To demonstrate the variation of the procedures in the experiments, we plot the averaged value and standard error of both test error and rank for some representative nonconvex penalty functions. Specifically, under each scenario considered in Section 4.1, we pick the nonconvex penalty that yields the best prediction and rank estimation performance. For each picked penalty, we plot the averaged value of test error and rank along with the associated standard error, against the tuning parameter λ. The results are shown in Figures 9, 10, and 11. As is clear form the figures, the standard error is typically (at least) one order of magnitude smaller than the average. Moreover, the general patterns of test error and rank on the solution path are expected, except for a few points corresponding to very small values of λ. The irregularity of these few points occurs probably because the solutions are getting unstable as the nonconvex regularization becomes weak when λ is significantly small.

[8] Note that we do not assume that the sequence $\boldsymbol{\sigma}_k$ has a limit point.

Example-A (Low SNR, less missing entries)



(a) ROM, 90% missing, SNR $= 1$, true rank $= 10$ (b) ROM, 90% missing, SNR $= 1$, true rank $= 5$

(c) ROM, 90% missing, SNR $= 1$, true rank $= 10$ (d) ROM, 90% missing, SNR $= 1$, true rank $= 5$

Fig. 9 Random Orthogonal Model (ROM) simulations with SNR $= 1$. The optimal nonconvex penalties are obtained at $\gamma = 30$ and $\gamma = 20$ under the two scenarios respectively. The integers from 1 to 100 on the x-axis index the grid of 100 values of λ (from largest to smallest) as described in Section 4.1.

– To examine the rank dynamics of the updates in NC-IMPUTE, we compute the number of iterations that the algorithm takes for the convergence of the rank. We choose the same six non-convex penalties as above and evaluate the rank stabilization for several values of λ. The results are summarized in Figure 12. One clearly observes that except for few instances, it takes less than 10 iterations for the rank to stabilize. Moreover, when the penalty is more "nonconvex" (i.e., γ is smaller), the rank stabilization occurs earlier. These empirical results provide complementary information on rank stabilization that has been theoretically investigated in 3.1.1.

References

Alquier P (2015) A bayesian approach for noisy matrix completion: Optimal rate under general sampling distribution. Electronic Journal of Statistics 9(1):823–841

Bai Z, Silverstein JW (2010) Spectral Analysis of Large Dimensional Random Matrices. Springer

Bertsimas D, King A, Mazumder R (2016) Best subset selection via a modern optimization lens. Annals of Statistics (to appear)

Bhojanapalli S, Neyshabur B, Srebro N (2016) Global optimality of local search for low rank matrix recovery. In: Advances in Neural Information Processing Systems, pp 3873–3881

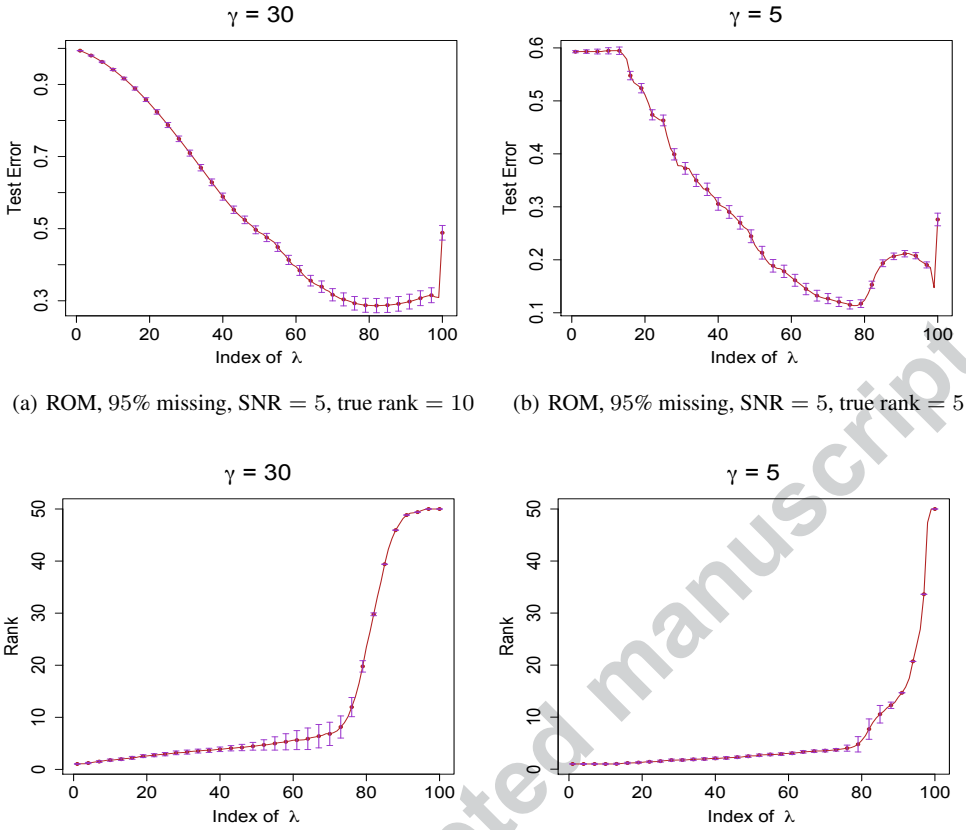
Borwein J, Lewis A (2006) Convex Analysis and Nonlinear Optimization. Springer

Cai JF, Candès EJ, Shen Z (2010) A singular value thresholding algorithm for matrix completion. SIAM Journal on Optimization 20:1956–1982

Candès E, Sing-Long C, Trzasko JD (2013) Unbiased risk estimates for singular value thresholding and spectral estimators. Signal Processing, IEEE Transactions on 61(19):4643–4657

Candès EJ, Plan Y (2010) Matrix completion with noise. Proceedings of the IEEE 98:925–936

Example-A (High SNR, more missing entries)



(a) ROM, 95% missing, SNR $= 5$, true rank $= 10$ (b) ROM, 95% missing, SNR $= 5$, true rank $= 5$

(c) ROM, 95% missing, SNR $= 5$, true rank $= 10$ (d) ROM, 95% missing, SNR $= 5$, true rank $= 5$

Fig. 10 Random Orthogonal Model (ROM) simulations with SNR $= 5$. The optimal nonconvex penalties are obtained at $\gamma = 30$ and $\gamma = 5$ under the two scenarios respectively. The integers from 1 to 100 on the x-axis index the grid of 100 values of λ (from largest to smallest) as described in Section 4.1.

Candès EJ, Recht B (2009) Exact matrix completion via convex optimization. Foundations of Computational mathematics 9:717–772

Candès EJ, Tao T (2010) The power of convex relaxation: near-optimal matrix completion. IEEE Transactions on Information Theory 56:2053–2080

Candes EJ, Wakin MB, Boyd SP (2008) Enhancing sparsity by reweighted ℓ_1 minimization. Journal of Fourier analysis and applications 14(5-6):877–905

Chen J, Liu D, Li X (2019a) Nonconvex rectangular matrix completion via gradient descent without $\ell_{2,\infty}$ regularization. arXiv preprint arXiv:190106116

Chen Y (2015) Incoherence-optimal matrix completion. IEEE Transactions on Information Theory 61(5):2909–2923

Chen Y, Wainwright MJ (2015) Fast low-rank estimation by projected gradient descent: General statistical and algorithmic guarantees. arXiv preprint arXiv:150903025

Chen Y, Bhojanapalli S, Sanghavi S, Ward R (2014) Coherent matrix completion. In: Proceedings of the 31st International Conference on Machine Learning, JMLR, pp 674–682

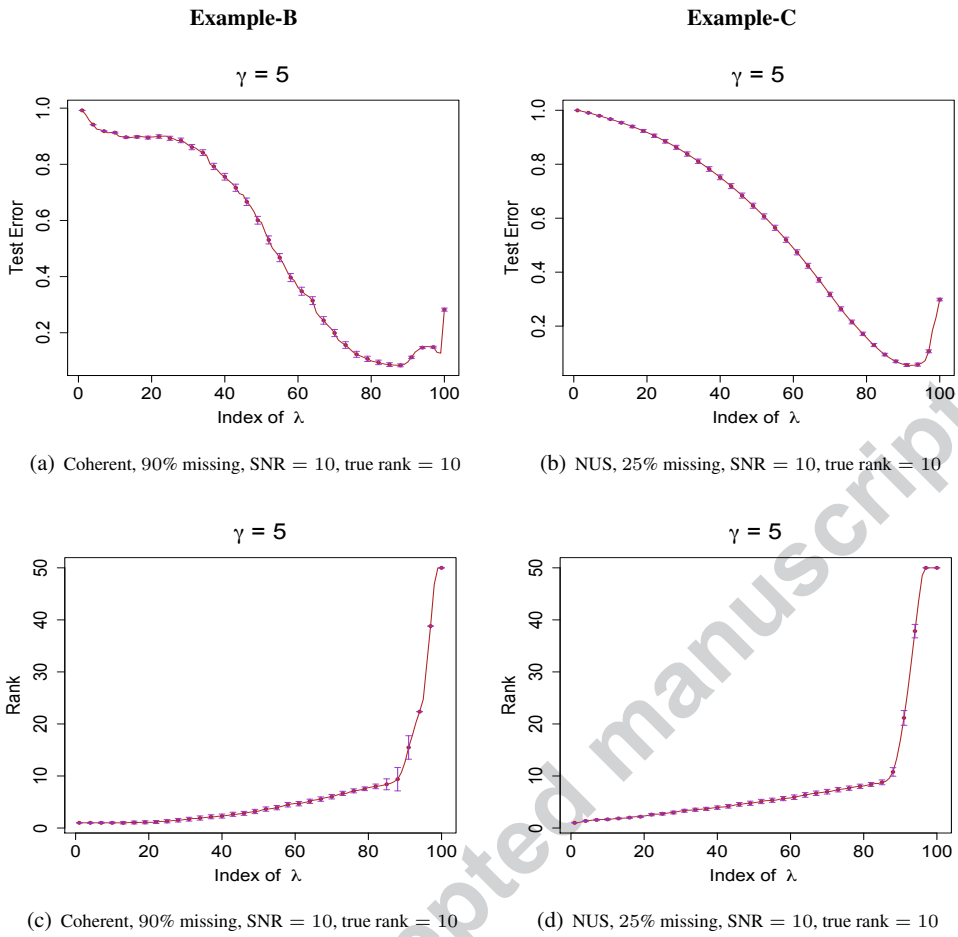
Chen Y, Chi Y, Fan J, Ma C, Yan Y (2019b) Noisy matrix completion: Understanding statistical guarantees for convex relaxation via nonconvex optimization. arXiv preprint arXiv:190207698

Chi Y, Lu YM, Chen Y (2019) Nonconvex optimization meets low-rank matrix factorization: An overview. IEEE Transactions on Signal Processing 67(20):5239–5269

Chistov AL, Grigor'ev DY (1984) Complexity of quantifier elimination in the theory of algebraically closed fields. In: Mathematical Foundations of Computer Science 1984, Springer, pp 17–31

Daubechies I, DeVore R, Fornasier M, Güntürk CS (2010) Iteratively reweighted least squares minimization for sparse recovery. Communications on Pure and Applied Mathematics: A Journal Issued by the Courant Institute of Mathematical Sciences 63(1):1–38

Dempster AP, Laird NM, Rubin DB (1977) Maximum likelihood from incomplete data via the em algorithm. Journal



(a) Coherent, 90% missing, SNR = 10, true rank = 10

(b) NUS, 25% missing, SNR = 10, true rank = 10

(c) Coherent, 90% missing, SNR = 10, true rank = 10

(d) NUS, 25% missing, SNR = 10, true rank = 10

Fig. 11 Coherent and Nonuniform Sampling (NUS) simulations with SNR = 10. The optimal nonconvex penalties are both obtained at $\gamma = 5$ under the two scenarios respectively. The integers from 1 to 100 on the x-axis index the grid of 100 values of λ (from largest to smallest) as described in Section 4.1.

of the Royal Statistical Society, Series B 39:1–38

Efron B, Hastie T, Johnstone I, Tibshirani R (2004) Least angle regression (with discussion). Annals of Statistics 32(2):407–499

Fan J, Li R (2001) Variable selection via nonconcave penalized likelihood and its oracle properties. Journal of the American Statistical Association 96:1348–1360

Fazel M (2002) Matrix rank minimization with applications. PhD thesis, Stanford University

Feng L, Zhang CH (2017) Sorted concave penalized regression. arXiv preprint arXiv:171209941

Fornasier M, Rauhut H, Ward R (2011) Low-rank matrix recovery via iteratively reweighted least squares minimization. SIAM Journal on Optimization 21(4):1614–1640

Frank IE, Friedman JH (1993) A statistical view of some chemometrics regression tools. Technometrics 35:109–135

Freund RM, Grigas P, Mazumder R (2015) An Extended Frank-Wolfe Method with "In-Face" Directions, and its Application to Low-Rank Matrix Completion. ArXiv e-prints 1511.02204

Ge R, Lee JD, Ma T (2016) Matrix completion has no spurious local minimum. In: Advances in Neural Information Processing Systems, pp 2973–2981

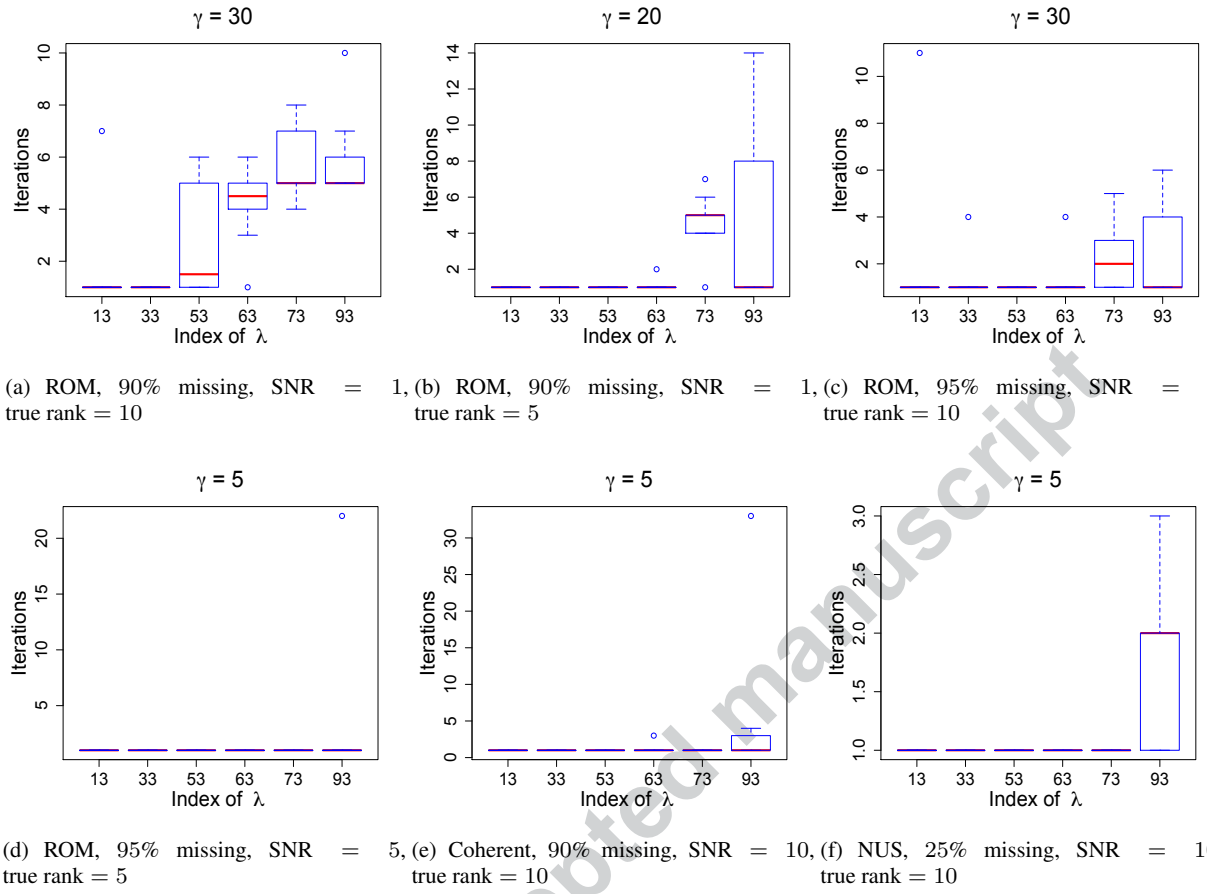
Ge R, Jin C, Zheng Y (2017) No spurious local minima in nonconvex low rank problems: A unified geometric analysis. In: Proceedings of the 34th International Conference on Machine Learning-Volume 70, JMLR. org, pp 1233–1242

Golub G, Van Loan C (1983) Matrix Computations. Johns Hopkins University Press, Baltimore.

Gross D (2011) Recovering low-rank matrices from few coefficients in any basis. IEEE Transactions on Information Theory 57(3):1548–1566

Gu S, Xie Q, Meng D, Zuo W, Feng X, Zhang L (2017) Weighted nuclear norm minimization and its applications to low level vision. International journal of computer vision 121(2):183–208

Hardt M (2014) Understanding alternating minimization for matrix completion. In: 2014 IEEE 55th Annual Sym-



(a) ROM, 90% missing, SNR = 1, true rank = 10 (b) ROM, 90% missing, SNR = 1, true rank = 5 (c) ROM, 95% missing, SNR = 5, true rank = 10

(d) ROM, 95% missing, SNR = 5, true rank = 5 (e) Coherent, 90% missing, SNR = 10, true rank = 10 (f) NUS, 25% missing, SNR = 10, true rank = 10

Fig. 12 The y-axis denotes the number of iterations NC-IMPUTE takes to stabilize the rank. The integers on the x-axis index some values on a grid of λ (from largest to smallest) as described in Section 4.1. The six plots represent the six scenarios considered in Section 4.1: (a)-(d) correspond to the four scenarios of Example-A; (e) covers Example-B; (f) is for Example-C. Each procedure is repeated 10 times.

posium on Foundations of Computer Science, IEEE, pp 651–660

Hardt M, Wootters M (2014) Fast matrix completion without the condition number. In: Conference on learning theory, pp 638–678

Hastie T, Tibshirani R, Friedman J (2009) The Elements of Statistical Learning: Prediction, Inference and Data Mining (Second Edition). Springer Verlag, New York

Hastie T, Mazumder R, Lee JD, Zadeh R (2016) Matrix completion and low-rank svd via fast alternating least squares. Journal of Machine Learning Research to appear

Hazimeh H, Mazumder R (2019) Fast best subset selection: Coordinate descent and local combinatorial optimization algorithms. Operations Research, to appear

Horn RA, Johnson CR (2012) Matrix analysis. Cambridge university press

Jaggi M, Sulovský M (2010) A simple algorithm for nuclear norm regularized problems. In: Proceedings of the 27th International Conference on Machine Learning (ICML-10), pp 471–478

Jain P, Meka R, Dhillon IS (2010) Guaranteed rank minimization via singular value projection. In: Advances in Neural Information Processing Systems, pp 937–945

Jain P, Netrapalli P, Sanghavi S (2013) Low-rank matrix completion using alternating minimization. In: Proceedings of the forty-fifth annual ACM symposium on Theory of computing, ACM, pp 665–674

Keshavan RH, Montanari A, Oh S (2010) Matrix completion from noisy entries. Journal of Machine Learning Research 11:2057–2078

Klopp O (2014) Noisy low-rank matrix completion with general sampling distribution. Bernoulli 20(1):282–303

Koltchinskii V, Lounici K, Tsybakov AB (2011) Nuclear-norm penalization and optimal rates for noisy low-rank matrix completion. The Annals of Statistics 39(5):2302–2329

Larsen R (2004) Propack-software for large and sparse svd calculations. Available at `http://sun.stanford.edu/~rmunk/PROPACK`

Lecué G, Mendelson S (2018) Regularization and the small-ball method i: sparse recovery. The Annals of Statistics 46(2):611–641

Lewis AS (1995) The convex analysis of unitarily invariant matrix functions. Journal of Convex Analysis 2:173–183

Loh PL, Wainwright MJ (2015) Regularized m-estimators with nonconvexity: Statistical and algorithmic theory for local optima. Journal of Machine Learning Research 16:559–616

Lv J, Fan Y (2009) A unified approach to model selection and sparse recovery using regularized least squares. The Annals of Statistics 37:3498–3528

Ma C, Wang K, Chi Y, Chen Y (2017) Implicit regularization in nonconvex statistical estimation: Gradient descent converges linearly for phase retrieval, matrix completion and blind deconvolution. arXiv preprint arXiv:171110467

Mazumder R, Radchenko P (2015) The discrete dantzig selector: Estimating sparse linear models via mixed integer linear optimization. arXiv preprint arXiv:150801922

Mazumder R, Hastie T, Tibshirani R (2010) Spectral regularization algorithms for learning large incomplete matrices. Journal of Machine Learning Research 11:2287–2322

Mazumder R, Friedman JH, Hastie T (2011) Sparsenet: coordinate descent with nonconvex penalties. Journal of the American Statistical Association 106:1125–1138

Mazumder R, Radchenko P, Dedieu A (2017) Subset selection with shrinkage: Sparse linear modeling when the snr is low. arXiv preprint arXiv:170803288

Mohan K, Fazel M (2010) Reweighted nuclear norm minimization with application to system identification. In: Proceedings of the 2010 American Control Conference, IEEE, pp 2953–2959

Mohan K, Fazel M (2012) Iterative reweighted algorithms for matrix rank minimization. Journal of Machine Learning Research 13(Nov):3441–3473

Negahban SN, Wainwright MJ (2011) Estimation of (near) low-rank matrices with noise and high-dimensional scaling. The Annals of Statistics 39:1069–1097

Negahban SN, Wainwright MJ (2012) Restricted strong convexity and weighted matrix completion: optimal bounds with noise. Journal of Machine Learning Research 13:1665–1697

Nikolova M (2000) Local strong homogeneity of a regularized estimator. SIAM Journal on Applied Mathematics 61:633–658

Recht B (2011) A simpler approach to matrix completion. Journal of Machine Learning Research 12:3413–3430

Recht B, Fazel M, Parrilo PA (2010) Guaranteed minimum-rank solutions of linear matrix equations via nuclear norm minimization. SIAM Review 52:471–501

Rockafellar RT (1970) Convex Analysis. Princeton University Press, Princeton, New Jersey

Rohde A, Tsybakov AB (2011) Estimation of high-dimensional low-rank matrices. The Annals of Statistics 39:887–930

Shapiro A, Xie Y, Zhang R (2018) Matrix completion with deterministic pattern: A geometric perspective. IEEE Transactions on Signal Processing 67(4):1088–1103

SIGKDD A, Netflix (2007) Soft modelling by latent variables: the nonlinear iterative partial least squares (NIPALS) approach. In: Proceedings of KDD Cup and Workshop

Stein CM (1981) Estimation of the mean of a multivariate normal distribution. The Annals of Statistics pp 1135–1151

Stewart GW, Sun JG (1990) Matrix Perturbation Theory. Computer science and scientific computing, Academic Press

Sun R, Luo ZQ (2016) Guaranteed matrix completion via non-convex factorization. IEEE Transactions on Information Theory 62(11):6535–6579

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

Troyanskaya O, Cantor M, Sherlock G, Brown P, Hastie T, Tibshirani R, Botstein D, Altman RB (2001) Missing value estimation methods for dna microarrays. Bioinformatics 17(6):520–525

Wang S, Weng H, Maleki A (2019) Which bridge estimator is optimal for variable selection? Annals of Statistics, to appear

Zhang CH (2010) Nearly unbiased variable selection under minimax concave penalty. The Annals of Statistics 38:894–942

Zhang CH, Zhang T (2012) A general theory of concave regularization for high-dimensional sparse estimation problems. Statistical Science 27(4):576–593

Zheng L, Maleki A, Weng H, Wang X, Long T (2017) Does ℓ_p-minimization outperform ℓ_1-minimization? IEEE Transactions on Information Theory 63(11):6896–6935

Zheng Q, Lafferty J (2016) Convergence analysis for rectangular matrix completion using burer-monteiro factorization and gradient descent. arXiv preprint arXiv:160507051

Zou H (2006) The adaptive lasso and its oracle properties. Journal of the American statistical association 101(476):1418–1429

Zou H, Li R (2008) One-step sparse estimates in nonconcave penalized likelihood models. The Annals of Statistics 36(4):1509–1533