Inference and uncertainty quantification for noisy matrix completion

Yuxin Chen[a,1], Jianqing Fan[b], Cong Ma[b], and Yuling Yan[b]

[a]Department of Electrical Engineering, Princeton University, Princeton, NJ 08544; and [b]Department of Operations Research and Financial Engineering, Princeton University, Princeton, NJ 08544

Edited by Peter J. Bickel, University of California, Berkeley, CA, and approved October 8, 2019 (received for review June 11, 2019)

Noisy matrix completion aims at estimating a low-rank matrix given only partial and corrupted entries. Despite remarkable progress in designing efficient estimation algorithms, it remains largely unclear how to assess the uncertainty of the obtained estimates and how to perform efficient statistical inference on the unknown matrix (e.g., constructing a valid and short confidence interval for an unseen entry). This paper takes a substantial step toward addressing such tasks. We develop a simple procedure to compensate for the bias of the widely used convex and nonconvex estimators. The resulting debiased estimators admit nearly precise nonasymptotic distributional characterizations, which in turn enable optimal construction of confidence intervals/regions for, say, the missing entries and the low-rank factors. Our inferential procedures do not require sample splitting, thus avoiding unnecessary loss of data efficiency. As a byproduct, we obtain a sharp characterization of the estimation accuracy of our debiased estimators in both rate and constant. Our debiased estimators are tractable algorithms that provably achieve full statistical efficiency.

confidence intervals | convex relaxation | nonconvex optimization

Low-rank matrix completion is concerned with recovering a low-rank matrix, when only a small fraction of its entries are revealed (1–3). The importance of this problem cannot be overstated, due to its broad applications in, e.g., recommendation systems, sensor network localization, magnetic resonance imaging, computer vision, large covariance estimation, and latent factor learning to name just a few. Tackling this problem in large-scale applications is computationally challenging, resulting from the intrinsic nonconvexity incurred by the low-rank structure. To further complicate matters, another inevitable challenge stems from the imperfectness of data acquisition mechanisms, wherein the acquired samples are usually contaminated by a certain amount of noise.

Fortunately, if the entries of the unknown matrix are sufficiently delocalized and randomly revealed, this problem may not be as hard as it seems. Substantial progress has been made over the past several years in designing computationally tractable algorithms—including both convex and nonconvex approaches—that allow one to fill in unseen entries faithfully from partial noisy samples (4–13). Nevertheless, modern decision making would often require one step further. It not merely anticipates a faithful estimate, but also seeks to quantify the uncertainty or "confidence" of the provided estimate, ideally in a reasonably accurate fashion. For instance, given an estimate returned by the convex approach, how does one use it to identify a short interval that is likely to contain a missing entry?

Conducting effective uncertainty quantification for noisy matrix completion is, however, far from straightforward. For the most part, the state-of-the-art matrix completion algorithms require solving highly complex optimization problems, which often do not admit closed-form solutions. Of necessity, it is extremely challenging to pin down the distributions of the estimates returned by these algorithms. The lack of distributional characterizations presents a major roadblock to performing valid, yet efficient, statistical inference on the unknown matrix of interest.

It is worth noting that a number of recent papers have been dedicated to inference and uncertainty quantification for various high-dimensional problems, including Lasso (14–18), generalized linear models (17, 19), and graphical models (20, 21), among others. Very little work, however, has looked into noisy matrix completion along this direction. While nonasymptotic statistical guarantees for noisy matrix completion have been derived in prior theory, the existing estimation error bounds are supplied only at an order-wise level. Such order-wise error bounds either lose a significant factor relative to the optimal guarantees or come with an unspecified (but often enormous) preconstant. Viewed in this light, a confidence region constructed directly based on such results is bound to be overly conservative, resulting in substantial overcoverage.

A Glimpse of Our Main Contributions

This paper takes a substantial step toward statistically optimal inference and uncertainty quantification for noisy matrix completion. Specifically, we develop a simple procedure to compensate for the bias of the widely used convex and nonconvex estimators. The resulting debiased estimators admit nearly accurate nonasymptotic distributional guarantees. Such distributional characterizations in turn allow us to reason about the uncertainty of the obtained estimates vis-à-vis the unknown matrix. For instance, we can construct 1) confidence intervals for each entry—either observed or missing—of the unknown matrix and

Significance

Matrix completion finds numerous applications in data science, ranging from information retrieval to medical imaging. While substantial progress has been made in designing estimation algorithms, it remains unknown how to perform optimal statistical inference on the unknown matrix given the obtained estimates—a task at the core of modern decision making. We propose procedures to debias the popular convex and nonconvex estimators and derive distributional characterizations for the resulting debiased estimators. This distributional theory enables valid inference on the unknown matrix. Our procedures 1) yield optimal construction of confidence intervals for missing entries and 2) achieve optimal estimation accuracy in a sharp manner.

Author contributions: Y.C., J.F., C.M., and Y.Y. designed research; Y.C., J.F., C.M., and Y.Y. performed research; C.M. and Y.Y. analyzed data; and Y.C., J.F., C.M., and Y.Y. wrote the paper.

The authors declare no competing interest.

This article is a PNAS Direct Submission.

[1]To whom correspondence may be addressed. Email: yuxin.chen@princeton.edu.

This article contains supporting information online at www.pnas.org/lookup/suppl/doi:10.1073/pnas.1910053116/-/DCSupplemental.

First published October 30, 2019.

2) confidence regions for the low-rank factors of interest (modulo some global ambiguity), both of which are provably optimal. As a byproduct, we characterize the Euclidean estimation errors of the proposed debiased estimators, which match statistical efficiency precisely (including the preconstant). This theory demonstrates that a computationally feasible algorithm can achieve the best possible statistical efficiency (including the preconstant) for noisy matrix completion.

Models and Notation

To cast noisy matrix completion in concrete statistical settings, we adopt a model commonly studied in the literature.

Ground Truth. We are interested in estimating an unknown rank-r matrix $\boldsymbol{M}^\star \in \mathbb{R}^{n\times n}$,[†] whose rank-$r$ singular-value decomposition (SVD) is given by $\boldsymbol{M}^\star = \boldsymbol{U}^\star \boldsymbol{\Sigma}^\star \boldsymbol{V}^{\star\top}$. For convenience, let $\boldsymbol{X}^\star \triangleq \boldsymbol{U}^\star \boldsymbol{\Sigma}^{\star 1/2} \in \mathbb{R}^{n\times r}$ and $\boldsymbol{Y}^\star \triangleq \boldsymbol{V}^\star \boldsymbol{\Sigma}^{\star 1/2} \in \mathbb{R}^{n\times r}$ be the balanced low-rank factors of $\boldsymbol{M}^\star$, which obey

$$\boldsymbol{X}^{\star\top}\boldsymbol{X}^\star = \boldsymbol{Y}^{\star\top}\boldsymbol{Y}^\star = \boldsymbol{\Sigma}^\star \quad \text{and} \quad \boldsymbol{M}^\star = \boldsymbol{X}^\star \boldsymbol{Y}^{\star\top}. \tag{1}$$

Denote by $\sigma_i(\boldsymbol{M}^\star)$ the ith largest singular value of $\boldsymbol{M}^\star$. Set

$$\sigma_{\max} \triangleq \sigma_1(\boldsymbol{M}^\star),\ \sigma_{\min} \triangleq \sigma_r(\boldsymbol{M}^\star),\ \text{and}\ \kappa \triangleq \sigma_{\max}/\sigma_{\min}. \tag{2}$$

Observation Models. What we observe is a randomly subsampled and corrupted subset of the entries of $\boldsymbol{M}^\star$; namely,

$$M_{ij} = M_{ij}^\star + E_{ij}, \quad E_{ij} \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0,\sigma^2), \quad \text{for all } (i,j)\in\Omega, \tag{3}$$

where $\Omega \subseteq \{1,\cdots,n\}\times\{1,\cdots,n\}$ is a small set of indexes, and E_{ij} denotes independently generated noise at the location (i,j). From now on, we assume the random sampling model where each index (i,j) is included in Ω independently with probability p (i.e., data are missing uniformly at random). We use $\mathcal{P}_\Omega(\cdot)$ to represent the orthogonal projection onto the subspace of matrices that vanish outside the index set Ω.

Incoherence Conditions. Clearly, not all matrices can be reliably estimated from a highly incomplete set of measurements. To address this issue, we impose a standard incoherence condition (2) on the singular subspaces of $\boldsymbol{M}^\star$ (i.e., $\boldsymbol{U}^\star$ and $\boldsymbol{V}^\star$),

$$\max\{\|\boldsymbol{U}^\star\|_{2,\infty}, \|\boldsymbol{V}^\star\|_{2,\infty}\} \le \sqrt{\mu r/n}, \tag{4}$$

where μ is termed the incoherence parameter and $\|\boldsymbol{A}\|_{2,\infty}$ denotes the largest ℓ_2 norm of all rows in $\boldsymbol{A}$. A small μ implies that the energies of $\boldsymbol{U}^\star$ and $\boldsymbol{V}^\star$ are reasonably spread out across all of their rows.

Asymptotic Notation. $f(n) \lesssim h(n)$ (or $f(n) = O(h(n))$) means $|f(n)| \le c_1|h(n)|$ for some constant $c_1 > 0$, $f(n) \gtrsim h(n)$ means $|f(n)| \ge c_2|h(n)|$ for some constant $c_2 > 0$, $f(n) \asymp h(n)$ means $c_2|h(n)| \le |f(n)| \le c_1|h(n)|$ for some constants $c_1, c_2 > 0$, and $f(n) = o(h(n))$ means $\lim_{n\to\infty} f(n)/h(n) = 0$.

Inferential Procedures and Main Results

The proposed inferential procedure has its basis on 2 of the most popular matrix completion paradigms: convex relaxation and nonconvex optimization. Recognizing the complicated bias of these 2 highly nonlinear estimators and motivated by refs. 14, 15, and 17, we first illustrate how to perform bias correction, followed by a theory that establishes the near-Gaussianity and optimality of the proposed debiased estimators.

Algorithm 1. Gradient descent for solving Eq. 7

Suitable initialization: $\boldsymbol{X}^0$, $\boldsymbol{Y}^0$ (*SI Appendix*)
Gradient updates: for $t = 0, 1, \ldots, t_0 - 1$ **do**

$$\boldsymbol{X}^{t+1} = \boldsymbol{X}^t - \frac{\eta}{p}[\mathcal{P}_\Omega(\boldsymbol{X}^t\boldsymbol{Y}^{t\top} - \boldsymbol{M})\boldsymbol{Y}^t + \lambda\boldsymbol{X}^t], \tag{6a}$$

$$\boldsymbol{Y}^{t+1} = \boldsymbol{Y}^t - \frac{\eta}{p}[[\mathcal{P}_\Omega(\boldsymbol{X}^t\boldsymbol{Y}^{t\top} - \boldsymbol{M})]^\top\boldsymbol{X}^t + \lambda\boldsymbol{Y}^t], \tag{6b}$$

where $\eta > 0$ determines the step size or the learning rate.

Background: Convex and Nonconvex Estimation Algorithms. We first review in passing 2 computationally feasible estimation algorithms that are arguably the most widely used in practice. They serve as the starting point for us to design inferential procedures for noisy low-rank matrix completion.

Convex Relaxation. Recall that the rank function is highly nonconvex, which often prevents us from computing a rank-constrained estimator in polynomial time. For the sake of computational feasibility, prior works suggest relaxing the rank function into its convex surrogate (22); for example, one can consider a penalized least-squares convex program

$$\underset{\boldsymbol{Z}\in\mathbb{R}^{n\times n}}{\text{minimize}} \quad \frac{1}{2}\|\mathcal{P}_\Omega(\boldsymbol{Z} - \boldsymbol{M})\|_{\mathrm{F}}^2 + \lambda\|\boldsymbol{Z}\|_*. \tag{5}$$

Here, $\|\cdot\|_*$ is the nuclear norm (the sum of singular values, which is a convex surrogate of the rank function), and $\lambda > 0$ is some regularization parameter. Under mild conditions, the solution to the convex program Eq. **5** attains appealing estimation accuracy (in an order-wise sense), provided that a proper regularization parameter λ is adopted (4, 13).

Nonconvex Optimization. It is recognized that the convex approach, which typically relies on solving a semidefinite program, is still expensive and not scalable to large dimensions. This motivates an alternative route, which represents the matrix variable via 2 low-rank factors $\boldsymbol{X}, \boldsymbol{Y} \in \mathbb{R}^{n\times r}$ and attempts solving the following nonconvex program directly:

$$\underset{\boldsymbol{X},\boldsymbol{Y}\in\mathbb{R}^{n\times r}}{\text{minimize}} \quad \frac{1}{2}\left\|\mathcal{P}_\Omega\left(\boldsymbol{X}\boldsymbol{Y}^\top - \boldsymbol{M}\right)\right\|_{\mathrm{F}}^2 + \frac{\lambda}{2}\|\boldsymbol{X}\|_{\mathrm{F}}^2 + \frac{\lambda}{2}\|\boldsymbol{Y}\|_{\mathrm{F}}^2. \tag{7}$$

Here, we choose a regularizer of the form $0.5\lambda(\|\boldsymbol{X}\|_{\mathrm{F}}^2 + \|\boldsymbol{Y}\|_{\mathrm{F}}^2)$ primarily to mimic the nuclear norm $\lambda\|\boldsymbol{Z}\|_*$ (23, 24). A variety of optimization algorithms have been proposed to tackle the nonconvex program Eq. **7** or its variants (7, 10, 11, 25); readers are referred to ref. 26 for a recent overview. As a prominent example, a 2-stage algorithm—gradient descent following suitable initialization—provably enjoys fast convergence for a wide range of scenarios (11, 13). The present paper focuses on this simple yet powerful algorithm, as documented in *Algorithm 1* and detailed in *SI Appendix*.

Intimate Connections between Convex and Nonconvex Estimates. Denote by $\boldsymbol{Z}^{\text{cvx}}$ any minimizer of the convex program Eq. **5** and $(\boldsymbol{X}^{\text{ncvx}}, \boldsymbol{Y}^{\text{ncvx}})$ the estimate returned by *Algorithm 1* aimed at solving Eq. **7**. As was recently shown in ref. 13, when the regularization parameter λ is properly chosen, these 2 estimates provably obey (see *SI Appendix* for a precise statement)

$$\boldsymbol{X}^{\text{ncvx}}\boldsymbol{Y}^{\text{ncvx}\top} \approx \boldsymbol{Z}^{\text{cvx}}. \tag{8}$$

[†]We restrict attention to squared matrices for simplicity of presentation. Most findings extend immediately to the more general rectangular case $\boldsymbol{M}^\star \in \mathbb{R}^{n_1\times n_2}$ with different n_1 and n_2.

In truth, the 2 matrices in Eq. **8** are exceedingly close to, if not identical with, each other. This salient feature paves the way for a unified treatment of convex and nonconvex approaches: Most inferential procedures and guarantees developed for the nonconvex estimate can be readily transferred to perform inference for the convex one, and vice versa.

Constructing Debiased Estimators. We are now well equipped to describe how to construct estimators based on the convex estimate $\boldsymbol{Z}^{\mathrm{cvx}}$ and the nonconvex estimate $(\boldsymbol{X}^{\mathrm{ncvx}}, \boldsymbol{Y}^{\mathrm{ncvx}})$, to enable efficient inference. Motivated by the proximity of the convex and nonconvex estimates and for the sake of conciseness, we abuse notation by using $\boldsymbol{Z}, \boldsymbol{X}, \boldsymbol{Y}$ for both convex and nonconvex estimates; see Table 1 and *SI Appendix* for precise definitions. This allows us to unify the presentation for both convex and nonconvex estimators.

Given that Eqs. **5** and **7** are both regularized least-squares problems, they behave effectively like shrinkage estimators, indicating that the provided estimates necessarily suffer from nonnegligible bias. To enable the desired statistical inference, it is natural to first correct the estimation bias.

A debiased estimator for the matrix. A natural debiasing strategy that immediately comes to mind is the simple linear transformation (recall the notation in Table 1)

$$\begin{aligned}\boldsymbol{Z}^0 &\triangleq \boldsymbol{Z} - p^{-1}\mathcal{P}_\Omega\left(\boldsymbol{Z}-\boldsymbol{M}\right)\\ &= \underbrace{p^{-1}\mathcal{P}_\Omega\left(\boldsymbol{M}^\star\right)}_{\text{mean: } \boldsymbol{M}^\star} + \underbrace{p^{-1}\mathcal{P}_\Omega\left(\boldsymbol{E}\right)}_{\text{mean: } \boldsymbol{0}} + \underbrace{\boldsymbol{Z} - p^{-1}\mathcal{P}_\Omega\left(\boldsymbol{Z}\right)}_{\text{mean: } \boldsymbol{0} \text{ (heuristically)}}, \end{aligned} \quad [9]$$

where we identify $\mathcal{P}_\Omega(\boldsymbol{M})$ with $\mathcal{P}_\Omega(\boldsymbol{M}^\star) + \mathcal{P}_\Omega(\boldsymbol{E})$. Heuristically, if Ω and $\boldsymbol{Z}$ are statistically independent, then $\boldsymbol{Z}^0$ serves as an unbiased estimator of $\boldsymbol{M}^\star$, i.e., $\mathbb{E}[\boldsymbol{Z}^0] = \boldsymbol{M}^\star$; this arises since the noise $\boldsymbol{E}$ has zero mean and $\mathbb{E}[\mathcal{P}_\Omega] = p\mathcal{I}$ under the uniform random sampling model, with $\mathcal{I}$ the identity operator. Despite its (near) unbiasedness nature at a heuristic level, however, the matrix $\boldsymbol{Z}^0$ is typically full rank, with nonnegligible energy spread across its entire spectrum. This results in dramatically increased variability in the estimate, which is undesirable for inferential purposes.

To remedy this issue, we propose to further project $\boldsymbol{Z}^0$ onto the set of rank-r matrices[‡] , leading to the estimator

$$\boldsymbol{M}^{\mathrm{d}} \triangleq \mathcal{P}_{\text{rank-}r}\left[\boldsymbol{Z} - \frac{1}{p}\mathcal{P}_\Omega\left(\boldsymbol{Z}-\boldsymbol{M}\right)\right], \quad [10]$$

where $\mathcal{P}_{\text{rank-}r}(\boldsymbol{B}) \triangleq \arg\min_{\boldsymbol{A}:\mathrm{rank}(\boldsymbol{A})\le r} \|\boldsymbol{A}-\boldsymbol{B}\|_{\mathrm{F}}$, and $\boldsymbol{Z}$ can again be found in Table 1. This projection step effectively suppresses the variability outside the r-dimensional principal subspace. As we shall demonstrate, the proposed estimator Eq. **10** provably debiases the provided estimate $\boldsymbol{Z}$, while optimally controlling the extent of variability.

An equivalent perspective on the low-rank factors. As it turns out, the debiased estimator Eq. **10** admits another almost equivalent representation that offers further insights. Specifically, we consider the following estimator for the low-rank factors,

$$\boldsymbol{X}^{\mathrm{d}} \triangleq \boldsymbol{X}\left(\boldsymbol{I}_r + p^{-1}\lambda\left(\boldsymbol{X}^\top\boldsymbol{X}\right)^{-1}\right)^{1/2}, \quad [11a]$$

$$\boldsymbol{Y}^{\mathrm{d}} \triangleq \boldsymbol{Y}\left(\boldsymbol{I}_r + p^{-1}\lambda\left(\boldsymbol{Y}^\top\boldsymbol{Y}\right)^{-1}\right)^{1/2}, \quad [11b]$$

where we recall the definition of $\boldsymbol{X}$ and $\boldsymbol{Y}$ in Table 1. To develop some intuition, let us look at a simple scenario where $\boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^\top$ is the rank-r SVD of $\boldsymbol{X}\boldsymbol{Y}^\top$ and $\boldsymbol{X} = \boldsymbol{U}\boldsymbol{\Sigma}^{1/2}$, $\boldsymbol{Y} = \boldsymbol{V}\boldsymbol{\Sigma}^{1/2}$. It is then self-evident that $\boldsymbol{X}^{\mathrm{d}} = \boldsymbol{U}(\boldsymbol{\Sigma} + (\lambda/p)\boldsymbol{I}_r)^{1/2}$ and $\boldsymbol{Y}^{\mathrm{d}} = \boldsymbol{V}(\boldsymbol{\Sigma} + (\lambda/p)\boldsymbol{I}_r)^{1/2}$. In words, $\boldsymbol{X}^{\mathrm{d}}$ and $\boldsymbol{Y}^{\mathrm{d}}$ are obtained by deshrinking the spectrum of $\boldsymbol{X}$ and $\boldsymbol{Y}$ properly.

As we formally establish in *SI Appendix*, the estimator Eq. **11** for the low-rank factors is extremely close to the debiased estimator Eq. **10** for the whole matrix, in the sense that

$$\boldsymbol{M}^{\mathrm{d}} \approx \boldsymbol{X}^{\mathrm{d}}\boldsymbol{Y}^{\mathrm{d}\top}. \quad [12]$$

Main Results: Distributional Guarantees. The proposed estimators admit tractable distributional characterizations in the large-n regime, which facilitates the construction of confidence regions for many quantities of interest. In particular, this paper centers around 2 types of inferential problems:

1) Each entry of the matrix $\boldsymbol{M}^\star$: The entry can be either missing (i.e., predicting an unseen entry) or observed (i.e., denoising an observed entry). For example, in the problem of sensor localization (27), one wants to infer the distance between any 2 sensors, given partially revealed distances. Mathematically, this seeks to determine the distribution of
$$M_{ij}^{\mathrm{d}} - M_{ij}^\star, \qquad \text{for all } 1 \le i,j \le n. \quad [13]$$
2) The low-rank factors $\boldsymbol{X}^\star, \boldsymbol{Y}^\star \in \mathbb{R}^{n\times r}$: The low-rank factors often reveal critical information about the applications of interest [e.g., community memberships of each individual in the community detection problem (28), angles between each object and a global reference point in the angular synchronization problem (29), or factor loadings and latent factors in factor analysis (30)]. Recognizing the global rotational ambiguity issue,[§] we aim to pin down the distributions of $\boldsymbol{X}^{\mathrm{d}}$ and $\boldsymbol{Y}^{\mathrm{d}}$ up to global rotational ambiguity. More precisely, we intend to characterize the distributions of
$$\boldsymbol{X}^{\mathrm{d}}\boldsymbol{H}^{\mathrm{d}} - \boldsymbol{X}^\star \qquad \text{and} \qquad \boldsymbol{Y}^{\mathrm{d}}\boldsymbol{H}^{\mathrm{d}} - \boldsymbol{Y}^\star \quad [14]$$
for the global rotation matrix $\boldsymbol{H}^{\mathrm{d}} \in \mathbb{R}^{r\times r}$ that best "aligns" $(\boldsymbol{X}^{\mathrm{d}}, \boldsymbol{Y}^{\mathrm{d}})$ and $(\boldsymbol{X}^\star, \boldsymbol{Y}^\star)$, i.e.,
$$\boldsymbol{H}^{\mathrm{d}} \triangleq \arg\min_{\boldsymbol{R}\in\mathcal{O}^{r\times r}} \left\|\boldsymbol{X}^{\mathrm{d}}\boldsymbol{R} - \boldsymbol{X}^\star\right\|_{\mathrm{F}}^2 + \left\|\boldsymbol{Y}^{\mathrm{d}}\boldsymbol{R} - \boldsymbol{Y}^\star\right\|_{\mathrm{F}}^2. \quad [15]$$
Here, $\mathcal{O}^{r\times r}$ is the set of orthonormal matrices in $\mathbb{R}^{r\times r}$.

Clearly, the above 2 inferential problems are tightly related: An accurate distributional characterization for the low-rank factors (Eq. **14**) often results in a distributional guarantee for the entries (Eq. **13**).

Distributional guarantees for low-rank factors. We begin with our distributional characterizations of the low-rank factors. Here, $\boldsymbol{e}_i$ denotes the ith standard basis vector in $\mathbb{R}^n$.

Theorem 1. *Suppose that the sample complexity meets* $n^2 p \ge C\kappa^8\mu^3 r^2 n\log^3 n$ *for some sufficiently large constant* $C > 0$ *and the noise obeys* $\sigma/\sigma_{\min} \le c\sqrt{p/(\kappa^8\mu r n\log^2 n)}$ *for some sufficiently small constant* $c > 0$. *Then one can write*

$$\boldsymbol{X}^{\mathrm{d}}\boldsymbol{H}^{\mathrm{d}} - \boldsymbol{X}^\star = \boldsymbol{Z}_{\boldsymbol{X}} + \boldsymbol{\Psi}_{\boldsymbol{X}}, \quad [16a]$$

$$\boldsymbol{Y}^{\mathrm{d}}\boldsymbol{H}^{\mathrm{d}} - \boldsymbol{Y}^\star = \boldsymbol{Z}_{\boldsymbol{Y}} + \boldsymbol{\Psi}_{\boldsymbol{Y}}, \quad [16b]$$

[‡]The true rank r can often be reliably estimated in a data-dependent manner. For instance, according to ref. 13, theorem 1, one can employ a rank estimator $\hat{r} \triangleq \min_i\left\{\sigma_{i+1}(\boldsymbol{Z})/\sigma_i(\boldsymbol{Z}) \le n^{-1/2}\right\}$, which recovers the true rank with high probability under our assumptions.

[§]For any $r\times r$ rotation matrix $\boldsymbol{H}$, we cannot possibly distinguish $(\boldsymbol{X}^\star, \boldsymbol{Y}^\star)$ from $(\boldsymbol{X}^\star\boldsymbol{H}, \boldsymbol{Y}^\star\boldsymbol{H})$, if only pairwise measurements are available.

Table 1. Notation used to unify the convex estimate Z^{cvx} and the nonconvex estimate $(X^{\mathrm{ncvx}}, Y^{\mathrm{ncvx}})$

$Z \in \mathbb{R}^{n\times n}$	Either Z^{cvx} or $X^{\mathrm{ncvx}} Y^{\mathrm{ncvx}\top}$
$X, Y \in \mathbb{R}^{n\times r}$	For the nonconvex case, we take $X = X^{\mathrm{ncvx}}$ and $Y = Y^{\mathrm{ncvx}}$; for the convex case, let $X = X^{\mathrm{cvx}}$ and $Y = Y^{\mathrm{cvx}}$, which are the balanced low-rank factors of $Z^{\mathrm{cvx},r}$ obeying $Z^{\mathrm{cvx},r} = X^{\mathrm{cvx}} Y^{\mathrm{cvx}\top}$ and $X^{\mathrm{cvx}\top} X^{\mathrm{cvx}} = Y^{\mathrm{cvx}\top} Y^{\mathrm{cvx}}$.
$M^{\mathrm{d}} \in \mathbb{R}^{n\times n}$	The proposed debiased estimator as in Eq. **10**.
$X^{\mathrm{d}}, Y^{\mathrm{d}} \in \mathbb{R}^{n\times r}$	The proposed estimator as in Eq. **11**.

Here, $Z^{\mathrm{cvx},r} = \mathcal{P}_{\text{rank-}r}(Z^{\mathrm{cvx}})$ is the best rank-r approximation of Z^{cvx}. See *SI Appendix* for a complete summary.

with $(X^\star, Y^\star)$ defined in Eq. **1**, *$(X^{\mathrm{d}}, Y^{\mathrm{d}})$ defined in Table 1, and H^{d} defined in Eq.* **15**. *Here, the rows of $Z_X \in \mathbb{R}^{n\times r}$ (resp. $Z_Y \in \mathbb{R}^{n\times r}$) are independent and obey*

$$Z_X^\top e_j \overset{\text{i.i.d.}}{\sim} \mathcal{N}\left(\mathbf{0}, \frac{\sigma^2}{p}(\Sigma^\star)^{-1}\right), \quad \text{for} \quad 1 \le j \le n; \qquad \textbf{[17a]}$$

$$Z_Y^\top e_j \overset{\text{i.i.d.}}{\sim} \mathcal{N}\left(\mathbf{0}, \frac{\sigma^2}{p}(\Sigma^\star)^{-1}\right), \quad \text{for} \quad 1 \le j \le n. \qquad \textbf{[17b]}$$

In addition, the residual matrices $\Psi_X, \Psi_Y \in \mathbb{R}^{n\times r}$ satisfy, with probability at least $1 - O(n^{-3})$, that

$$\max\left\{\|\Psi_X\|_{2,\infty}, \|\Psi_Y\|_{2,\infty}\right\} = o\left(\frac{\sigma\sqrt{r}}{\sqrt{p\sigma_{\max}}}\right). \qquad \textbf{[18]}$$

In words, *Theorem 1* decomposes the estimation error $X^{\mathrm{d}}H^{\mathrm{d}} - X^\star$ (resp. $Y^{\mathrm{d}}H^{\mathrm{d}} - Y^\star$) into a Gaussian component Z_X (resp. Z_Y) and a residual term Ψ_X (resp. Ψ_Y). If the sample size is sufficiently large and the noise size is sufficiently small, then the residual terms are much smaller in size compared to Z_X and Z_Y. To see this, it is helpful to leverage the Gaussianity (Eq. **17a**) to compute that for each $1 \le j \le n$, the jth row of Z_X obeys

$$\mathbb{E}\left[\left\|Z_X^\top e_j\right\|_2^2\right] = \mathsf{Tr}\left(\frac{\sigma^2}{p}(\Sigma^\star)^{-1}\right) \ge \frac{\sigma^2 r}{p\sigma_{\max}};$$

in other words, the typical size of the jth row of Z_X is no smaller than the order of $\sigma\sqrt{r/(p\sigma_{\max})}$. In comparison, the size of each row of Ψ_X (Eq. **18**) is much smaller than $\sigma\sqrt{r/(p\sigma_{\max})}$ (and hence smaller than the size of the corresponding row of Z_X) with high probability.

Remark 1. *Another interesting feature—which we make precise in the proof of Theorem 1—is that for any given $1 \le i, j \le n$, the two random vectors $Z_X^\top e_i$ and $Z_Y^\top e_j$ are nearly statistically independent. This is crucial for deriving inferential guarantees for the entries of the matrix.*

Distributional guarantees for matrix entries. Equipped with the above theory for low-rank factors and *Remark 1*, we are ready to characterize the distribution of $M_{ij}^{\mathrm{d}} - M_{ij}^\star$.

Theorem 2. *For each $1 \le i, j \le n$, define the variance $v_{ij}^\star$ as*

$$v_{ij}^\star \triangleq \frac{\sigma^2}{p}\left(\left\|U_{i,\cdot}^\star\right\|_2^2 + \left\|V_{j,\cdot}^\star\right\|_2^2\right), \qquad \textbf{[19]}$$

where $U_{i,\cdot}^\star$ (resp. $V_{j,\cdot}^\star$) denotes the ith (resp. jth) row of $U^\star$ (resp. $V^\star$). Suppose that

$$np \gtrsim \kappa^8\mu^3 r^3 \log^3 n, \quad \sigma\sqrt{(\kappa^8\mu r n \log^2 n)/p} \lesssim \sigma_{\min}, \qquad \textbf{[20a]}$$

$$\text{and } \left\|U_{i,\cdot}^\star\right\|_2 + \left\|V_{j,\cdot}^\star\right\|_2 \gtrsim \sqrt{\frac{r}{n}}\frac{\sigma}{\sigma_{\min}}\sqrt{\frac{\kappa^6\mu^2 r n \log^3 n}{p}}. \qquad \textbf{[20b]}$$

Then the matrix M^{d} defined in Table 1 satisfies

$$M_{ij}^{\mathrm{d}} - M_{ij}^\star = g_{ij} + \Delta_{ij}, \qquad \textbf{[21]}$$

where $g_{ij} \sim \mathcal{N}(0, v_{ij}^\star)$ and the residual obeys $|\Delta_{ij}| = o(\sqrt{v_{ij}^\star})$ with probability exceeding $1 - O(n^{-3})$.

Several remarks are in order. First, we develop some intuition regarding where the formula $v_{ij}^\star$ comes from. By virtue of *Theorem 1*, one has the following Gaussian approximation

$$X^{\mathrm{d}}H^{\mathrm{d}} - X^\star \approx Z_X \qquad \text{and} \qquad Y^{\mathrm{d}}H^{\mathrm{d}} - Y^\star \approx Z_Y.$$

Assuming that the first-order expansion is tight, one has

$$\begin{aligned} M_{ij}^{\mathrm{d}} - M_{ij}^\star &= \left[X^{\mathrm{d}}H^{\mathrm{d}}\left(Y^{\mathrm{d}}H^{\mathrm{d}}\right)^\top - X^\star Y^{\star\top}\right]_{ij} \\ &\approx e_i^\top\left(X^{\mathrm{d}}H^{\mathrm{d}} - X^\star\right)Y^{\star\top}e_j + e_i^\top X^\star\left(Y^{\mathrm{d}}H^{\mathrm{d}} - Y^\star\right)^\top e_j \\ &\approx e_i^\top Z_X Y^{\star\top} e_j + e_i^\top X^\star Z_Y^\top e_j. \end{aligned} \qquad \textbf{[22]}$$

According to *Remark 1*, $Z_X^\top e_i$ and $Z_Y^\top e_j$ are nearly independent. One can thus compute the variance of Eq. **22** as

$$\begin{aligned} \mathsf{Var}\left(M_{ij}^{\mathrm{d}} - M_{ij}^\star\right) &\overset{\text{(i)}}{\approx} \mathsf{Var}\left(e_i^\top Z_X Y^{\star\top} e_j\right) + \mathsf{Var}\left(e_i^\top X^\star Z_Y^\top e_j\right) \\ &\overset{\text{(ii)}}{=} p^{-1}\sigma^2\left\{e_j^\top Y^\star(\Sigma^\star)^{-1}Y^{\star\top}e_j + e_i^\top X^\star(\Sigma^\star)^{-1}X^{\star\top}e_i\right\} \\ &\overset{\text{(iii)}}{=} p^{-1}\sigma^2\left(\left\|U_{i,\cdot}^\star\right\|_2^2 + \left\|V_{j,\cdot}^\star\right\|_2^2\right) = v_{ij}^\star. \end{aligned}$$

Here, (i) relies on Eq. **22** and the near independence between $Z_X^\top e_i$ and $Z_Y^\top e_j$, (ii) uses the variance formula in *Theorem 1*, and (iii) arises from the definitions of $X^\star$ and $Y^\star$ (cf. Eq. **1**). This explains (heuristically) the variance formula $v_{ij}^\star$.

Given that *Theorem 2* reveals the tightness of Gaussian approximation under conditions in Eq. **20**, it in turn allows us to construct nearly accurate confidence intervals for each matrix entry $M_{ij}^\star$. This is formally summarized in the following corollary. Here, $[a \pm b]$ denotes the interval $[a - b, a + b]$.

Table 2. Empirical coverage rates of $M_{ij}^\star$ for different (r, p, σ) over 200 Monte Carlo trials

(r, p, σ)	Mean($\widehat{\mathrm{Cov}}_{\mathsf{E}}$)	Std($\widehat{\mathrm{Cov}}_{\mathsf{E}}$)
$(2, 0.2, 10^{-6})$	0.9380	0.0200
$(2, 0.2, 10^{-3})$	0.9392	0.0196
$(2, 0.4, 10^{-6})$	0.9455	0.0164
$(2, 0.4, 10^{-3})$	0.9456	0.0164
$(5, 0.2, 10^{-6})$	0.9226	0.0247
$(5, 0.2, 10^{-3})$	0.9271	0.0228
$(5, 0.4, 10^{-6})$	0.9410	0.0173
$(5, 0.4, 10^{-3})$	0.9417	0.0172

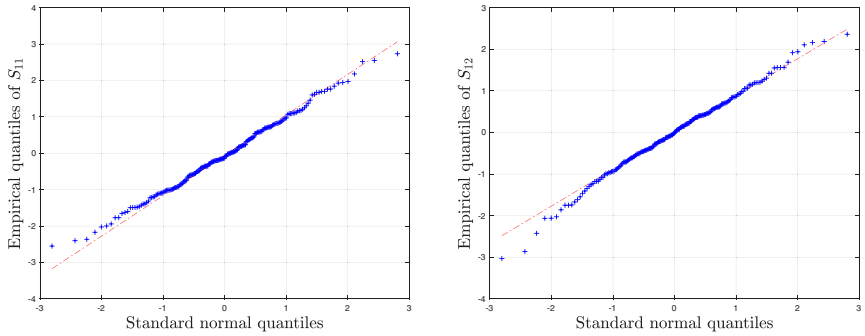

Fig. 1. Q-Q plots of S_{11} (*Left*) and S_{12} (*Right*) vs. the standard normal distribution. The results are reported over 200 independent trials for $r = 5$, $p = 0.4$, and $\sigma = 10^{-3}$.

Corollary 1 (Confidence Intervals for the Entries $\{M_{ij}^\star\}$). *Let $\boldsymbol{X}^{\mathrm{d}}$, $\boldsymbol{Y}^{\mathrm{d}}$, and $\boldsymbol{M}^{\mathrm{d}}$ be as defined in Table 1. For any given $1 \le i, j \le n$, suppose that Eq.* **20a** *holds and that*

$$\left\|\boldsymbol{U}_{i,\cdot}^\star\right\|_2 + \left\|\boldsymbol{V}_{j,\cdot}^\star\right\|_2 \gtrsim \sqrt{\frac{r}{n}}\frac{\sigma}{\sigma_{\min}}\sqrt{\frac{\kappa^{10}\mu^2 rn\log^3 n}{p}}. \qquad [23]$$

Denote by $\Phi(t)$ the CDF of a standard Gaussian random variable and by $\Phi^{-1}(\cdot)$ its inverse function. Let

$$v_{ij} \triangleq \frac{\sigma^2}{p}\left(\boldsymbol{X}_{i,\cdot}^{\mathrm{d}}\left(\boldsymbol{X}^{\mathrm{d}\top}\boldsymbol{X}^{\mathrm{d}}\right)^{-1}(\boldsymbol{X}_{i,\cdot}^{\mathrm{d}})^\top + \boldsymbol{Y}_{j,\cdot}^{\mathrm{d}}\left(\boldsymbol{Y}^{\mathrm{d}\top}\boldsymbol{Y}^{\mathrm{d}}\right)^{-1}(\boldsymbol{Y}_{j,\cdot}^{\mathrm{d}})^\top\right) \qquad [24]$$

be the empirical estimate of $v_{ij}^\star$. Then one has

$$\sup_{0<\alpha<1}\left|\mathbb{P}\left\{M_{ij}^\star \in \left[M_{ij}^{\mathrm{d}} \pm \Phi^{-1}(1-\alpha/2)\sqrt{v_{ij}}\right]\right\} - (1-\alpha)\right| = o(1).$$

In words, *Corollary 1* tells us that for any fixed significance level $0 < \alpha < 1$, the interval

$$\left[M_{ij}^{\mathrm{d}} \pm \Phi^{-1}(1-\alpha/2)\sqrt{v_{ij}}\right] \qquad [25]$$

is a nearly accurate $(1-\alpha)$ confidence interval of $M_{ij}^\star$.

In addition, we remark that when $\left\|\boldsymbol{U}_{i,\cdot}^\star\right\|_2 = \left\|\boldsymbol{V}_{j,\cdot}^\star\right\|_2 = 0$ (and hence $V_{ij}^\star = 0$), the above Gaussian approximation is completely off. In this case, one can still leverage *Theorem 1* to show that

$$M_{ij}^{\mathrm{d}} - M_{ij}^\star = M_{ij}^{\mathrm{d}} \approx \boldsymbol{u}^\top\boldsymbol{v}, \qquad [26]$$

where $\boldsymbol{u}, \boldsymbol{v} \in \mathbb{R}^r$ are independent and identically distributed according to $\mathcal{N}(\boldsymbol{0}, \sigma^2(\boldsymbol{\Sigma}^\star)^{-1}/p)$. However, it is nontrivial to determine whether $\left\|\boldsymbol{U}_{i,\cdot}^\star\right\|_2 + \left\|\boldsymbol{V}_{j,\cdot}^\star\right\|_2$ is vanishingly small or not based on the observed data, which makes it challenging to conduct efficient inference for entries with small (but a priori unknown) $\left\|\boldsymbol{U}_{i,\cdot}^\star\right\|_2 + \left\|\boldsymbol{V}_{j,\cdot}^\star\right\|_2$.

Last but not least, the careful reader might wonder how to interpret our conditions on the sample complexity and the signal-to-noise ratio. Take the case with $r, \mu, \kappa = O(1)$ for example: Our conditions read

$$n^2 p \gtrsim n\log^3 n; \qquad \sigma/\sigma_{\min} \lesssim \sqrt{p/(n\log^2 n)}. \qquad [27]$$

The first condition matches the sample complexity limit (up to some log factor), while the second one coincides with the regime (up to log factor) in which popular algorithms (like spectral methods or nonconvex algorithms) work better than a random guess (7, 10, 11). The take-away message is this: Once we are able to compute a reasonable estimate in an overall ℓ_2 sense, then we can reinforce it to conduct entrywise inference in a statistically efficient fashion.

Lower Bounds and Optimality for Inference. It is natural to ask how well our inferential procedures perform compared to other algorithms. Encouragingly, the debiased estimator is optimal in some sense; for instance, it attains the minimum covariance among all unbiased estimators. To formalize this claim, we 1) quantify the performance of 2 ideal estimators with the assistance of an oracle and 2) demonstrate that the performance of our debiased estimators is arbitrarily close to that of the ideal estimators. We remark in passing such results here; see *SI Appendix* for precise statements. Below, we denote by $\boldsymbol{X}_{i,\cdot}^\star$ (resp. $\boldsymbol{Y}_{i,\cdot}^\star$) the ith row of $\boldsymbol{X}^\star$ (resp. $\boldsymbol{Y}^\star$).

Lower bound for estimating $\boldsymbol{X}_{i,\cdot}^\star$ ($1 \le i \le n$). Suppose there is an oracle informing us of $\boldsymbol{Y}^\star$ and we observe the same set of data as in Eq. **3**. Under such an idealistic setting and under our sample complexity condition, one has, with high probability, that any unbiased estimator $\widehat{\boldsymbol{X}}_{i,\cdot}$ of $\boldsymbol{X}_{i,\cdot}^\star$ satisfies

$$\mathrm{Cov}\left(\widehat{\boldsymbol{X}}_{i,\cdot} - \boldsymbol{X}_{i,\cdot}^\star \,|\, \Omega\right) \succeq (1-o(1))\, p^{-1}\sigma^2\,(\boldsymbol{\Sigma}^\star)^{-1}.$$

This reveals that the covariance of the estimator $\boldsymbol{X}_{i,\cdot}^{\mathrm{d}}$ (cf. *Theorem 1*) attains the Cramér–Rao lower bound with high probability. The same conclusion applies to $\boldsymbol{Y}_{j,\cdot}^{\mathrm{d}}$ too.

Lower bound for estimating $M_{ij}^\star$ ($1 \le i, j \le n$). Suppose there is another oracle informing us of $\{\boldsymbol{X}_{k,\cdot}^\star\}_{k:k\neq i}$ and $\{\boldsymbol{Y}_{k,\cdot}^\star\}_{k:k\neq j}$, that is, everything about $\boldsymbol{X}^\star$ except $\boldsymbol{X}_{i,\cdot}^\star$ and everything about $\boldsymbol{Y}^\star$ except $\boldsymbol{Y}_{j,\cdot}^\star$. In addition, we observe the same set of data as in Eq. **3**, except that we do not get to see M_{ij}.[¶] Under this idealistic model, one can show that with high probability, any unbiased estimator of $M_{ij}^\star$ must have variance no smaller than $(1-o(1))v_{ij}^\star$, where $v_{ij}^\star$ is defined in *Theorem 2*. This indicates that the variance of our debiased estimator M_{ij}^{d} (cf. *Theorem 2*)—which certainly does not have access to the side information provided by the oracle—is arbitrarily close to the Cramér–Rao lower bound aided by an oracle.

Back to Estimation: The Debiased Estimator Is Optimal. While the emphasis herein is on inference, we nevertheless single out an important consequence that informs the estimation step. To be specific, the distributional guarantees derived in *Theorems 1* and *2* allow us to track the estimation accuracy of $\boldsymbol{M}^{\mathrm{d}}$, as stated below.

[¶]The exclusion of M_{ij} is merely for ease of presentation. One can consider the model where all M_{ij} with $(i, j) \in \Omega$ are observed with a slightly more complicated argument.

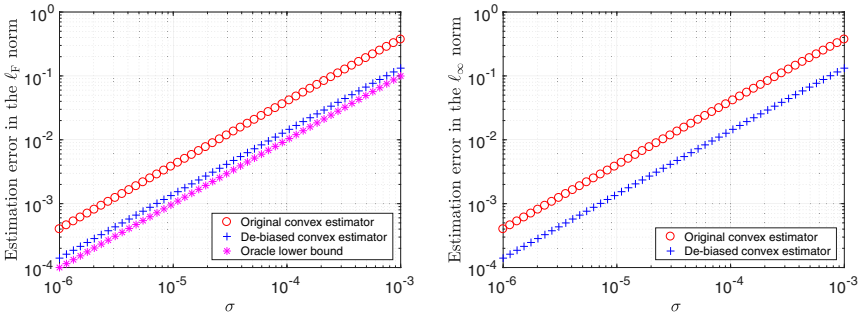

Fig. 2. (*Left*) Estimation error of $\boldsymbol{Z}^{\mathrm{cvx}}$ vs. $\boldsymbol{M}^{\mathrm{d}}$ measured in the Frobenius norm. (*Right*) Estimation error of $\boldsymbol{Z}^{\mathrm{cvx}}$ vs. $\boldsymbol{M}^{\mathrm{d}}$ measured in the ℓ_∞ norm. The results are averaged over 20 independent trials for $r=5$, $p=0.2$, and $n=1,000$.

Theorem 3 (Estimation Accuracy of M^{d}). *Let M^{d} be the debiased estimator as defined in Table 1. Instate the conditions in Eq.* **20a**. *Then with probability at least $1-O(n^{-3})$, one has*

$$\left\|\boldsymbol{M}^{\mathrm{d}}-\boldsymbol{M}^{\star}\right\|_{\mathrm{F}}^{2}=(2+o(1))\,nr\sigma^{2}/p. \qquad [28]$$

In stark contrast to prior statistical estimation guarantees (e.g., refs. 4–6 and 13), *Theorem 3* pins down the estimation error of the proposed debiased estimator in a sharp manner (namely, even the preconstant is fully determined). Encouragingly, there is a sense in which the proposed debiased estimator achieves the best possible statistical estimation accuracy. In fact, a lower bound has already been derived in ref. 4, section III.B, asserting that one cannot beat the mean-square estimation error $(2-o(1))\,nr\sigma^2/p$ even with the help of an oracle. See *SI Appendix* for a precise statement.

The implication of *Theorems 1* to *3* is remarkable: The debiasing step not merely facilitates uncertainty assessment, but also proves crucial in minimizing estimation errors. It achieves optimal statistical efficiency in terms of both the rate and the preconstant. This theory about a polynomial time algorithm matches the statistical limit in terms of the preconstant. This intriguing finding is further corroborated by numerical experiments (see Fig. 2).

Numerical Experiments. We conduct numerical experiments on synthetic data to verify the distributional characterizations provided in *Theorem 2*. The verification of *Theorem 1* is left to *SI Appendix*. Note that our main results hold for the debiased estimators built upon $\boldsymbol{Z}^{\mathrm{cvx}}$ and $\boldsymbol{X}^{\mathrm{ncvx}}\boldsymbol{Y}^{\mathrm{ncvx}\top}$. As we formalize in *SI Appendix*, these 2 debiased estimators are extremely close to each other. Therefore, to save space, we use the debiased estimator built upon the convex estimate $\boldsymbol{Z}^{\mathrm{cvx}}$ throughout the experiments.

Fix the dimension $n=1,000$ and the regularization parameter $\lambda=2.5\sigma\sqrt{np}$ throughout the experiments. We generate a rank-r matrix $\boldsymbol{M}^{\star}=\boldsymbol{X}^{\star}\boldsymbol{Y}^{\star\top}$, where $\boldsymbol{X}^{\star},\boldsymbol{Y}^{\star}\in\mathbb{R}^{n\times r}$ are random orthonormal matrices, and apply the proximal gradient method to solve the convex program Eq. **5**.

Denote $S_{ij}\triangleq v_{ij}^{-1/2}\left(M_{ij}^{\mathrm{d}}-M_{ij}^{\star}\right)$, where v_{ij} is the empirical variance defined in Eq. **24**. In view of the 95% confidence interval predicted by *Corollary 1*, for each (i,j), we define $\widehat{\mathrm{Cov}}_{\mathrm{E},(i,j)}$ to be the empirical coverage rate of $M_{ij}^{\star}$ over 200 Monte Carlo simulations. Correspondingly, denote by $\mathrm{Mean}(\widehat{\mathrm{Cov}}_{\mathrm{E}})$ (resp. $\mathrm{Std}(\widehat{\mathrm{Cov}}_{\mathrm{E}})$) the average (resp. the SD) of $\widehat{\mathrm{Cov}}_{\mathrm{E},(i,j)}$ over indexes $1\le i,j\le n$. As before, Table 2 gathers the empirical coverage rates for $M_{ij}^{\star}$ and Fig. 1 displays the quantile–quantile (Q-Q) plots of S_{11} and S_{12} vs. the standard Gaussian random variable over 200 Monte Carlo trials for $r=5$, $p=0.4$, and $\sigma=10^{-3}$. It is evident that the distribution of S_{ij} matches that of $\mathcal{N}(0,1)$ reasonably well.

In addition to the tractable distributional guarantees, the debiased estimator $\boldsymbol{M}^{\mathrm{d}}$ also exhibits superior estimation accuracy compared to the original estimator $\boldsymbol{Z}^{\mathrm{cvx}}$ (cf. *Theorem 3*). Fig. 2 reports the estimation error of $\boldsymbol{M}^{\mathrm{d}}$ vs. $\boldsymbol{Z}^{\mathrm{cvx}}$ measured in both the Frobenius norm and the ℓ_∞ norm across different noise levels. The results are averaged over 20 Monte Carlo simulations for $r=5$, $p=0.2$. It can be seen that the errors of the debiased estimator are uniformly smaller than that of the original estimator and are much closer to the oracle lower bound. As a result, we recommend using $\boldsymbol{M}^{\mathrm{d}}$ even for the purpose of estimation.

We conclude this section with experiments on real data. Similar to ref. 4, we use the daily temperature data (31) for 1,400 stations across the world in 2018, which results in a $1,400\times 365$ data matrix. Inspection of the singular values reveals that the data matrix is nearly low rank. We vary the observation probability p from 0.5 to 0.9 and randomly subsample the data accordingly. Based on the observed temperatures, we then apply the proposed methodology to obtain 95% confidence intervals for all of the entries. Table 3 reports the empirical coverage probabilities and the average length of the confidence intervals as well as the estimation error of both $\boldsymbol{Z}^{\mathrm{cvx}}$ and $\boldsymbol{M}^{\mathrm{d}}$ over 20 independent experiments. It can be seen that the average coverage probabilities are reasonably close to 95% and the confidence intervals are also quite short. In addition, the estimation error of $\boldsymbol{M}^{\mathrm{d}}$ is smaller than that of $\boldsymbol{Z}^{\mathrm{cvx}}$, which corroborates our theoretical prediction. The discrepancy between the nominal coverage probability and the actual one might arise from the facts that 1) the underlying true temperature matrix is only approximately low rank and 2) the noise in the temperature might not be independent.

Discussion

The present paper makes progress toward inference and uncertainty quantification for noisy matrix completion, by developing

Table 3. Empirical coverage rates and average lengths of the confidence intervals of the entries as well as the estimation error vs. observation probability p

	Coverage		CI length		$\|\widehat{\boldsymbol{Z}}-\boldsymbol{M}^{\star}\|_{\mathrm{F}}/\|\boldsymbol{M}^{\star}\|_{\mathrm{F}}$	
p	Mean	SD	Mean	SD	Convex $\boldsymbol{Z}^{\mathrm{cvx}}$	Debiased $\boldsymbol{M}^{\mathrm{d}}$
0.5	0.8265	0.0016	3.6698	0.0209	0.029	0.028
0.6	0.8268	0.0011	2.8774	0.0098	0.025	0.023
0.7	0.8431	0.0006	2.3426	0.0054	0.022	0.019
0.8	0.8725	0.0003	2.0234	0.0052	0.020	0.015
0.9	0.9093	0.0003	1.8296	0.0072	0.018	0.011

The results are averaged over 20 Monte Carlo trials.

simple debiased estimators that admit tractable and accurate distributional characterizations. While we have achieved some early success in accomplishing this, our results are likely suboptimal in the dependency on the rank r and the condition number κ. Also, our theory operates under the moderate-to-high signal-to-noise ratio (SNR) regime, where $\sigma_{\min}^2/\sigma^2$ (which is proportional to the SNR) is required to exceed the order of n/p; see the conditions in *Theorem 1*. How to conduct inference in the low SNR regime is an important future direction.

More broadly, this paper uncovers that computational feasibility and full statistical efficiency can sometimes be simultaneously achieved despite a high degree of nonconvexity. The analysis and insights herein might shed light on inference for a broader class of nonconvex statistical problems.

ACKNOWLEDGMENTS. Y.C. is supported in part by the Air Force Office of Scientific Research Young Investigator Program award FA9550-19-1-0030, by the Office of Naval Research (ONR) grant N00014-19-1-2120, by the Army Research Office grant W911NF-18-1-0303, by the NSF grants CCF-1907661 and IIS-1900140, and by the Princeton School of Engineering and Applied Science innovation award. J.F. is supported in part by NSF grants DMS-1662139 and DMS-1712591, ONR grant N00014-19-1-2120, and NIH grant 2R01-GM072611-13. C.M. is supported in part by Hudson River Trading AI Labs (HAIL) Fellowship. This work was done in part while Y.C. was visiting the Kavli Institute for Theoretical Physics (supported in part by the NSF grant PHY-1748958). We thank Weijie Su for helpful discussions.

1. N. Srebro, "Learning with matrix factorizations," PhD thesis, Massachusetts Institute of Technology, Cambridge, MA (2004).
2. E. Candès, B. Recht, Exact matrix completion via convex optimization. *Found. Comput. Math.* **9**, 717–772 (2009).
3. R. H. Keshavan, A. Montanari, S. Oh, Matrix completion from a few entries. *IEEE Trans. Inf. Theory* **56**, 2980–2998 (2010).
4. E. Candès, Y. Plan, Matrix completion with noise. *Proc. IEEE* **98**, 925–936 (2010).
5. S. Negahban, M. Wainwright, Restricted strong convexity and weighted matrix completion: Optimal bounds with noise. *J. Mach. Learn. Res.* **13**, 1665–1697 (2012).
6. V. Koltchinskii, K. Lounici, A. B. Tsybakov, Nuclear-norm penalization and optimal rates for noisy low-rank matrix completion. *Ann. Stat.* **39**, 2302–2329 (2011).
7. R. H. Keshavan, A. Montanari, S. Oh, Matrix completion from noisy entries. *J. Mach. Learn. Res.* **11**, 2057–2078 (2010).
8. D. Gross, Recovering low-rank matrices from few coefficients in any basis. *IEEE Trans. Inf. Theory* **57**, 1548–1566 (2011).
9. R. Foygel, N. Srebro, "Concentration-based guarantees for low-rank matrix reconstruction" in *Conference on Learning Theory*, S. M. Kakade, U. von Luxburg, Eds. (Proceedings of Machine Learning Research, Budapest, Hungary, 2011), pp. 315–340.
10. Y. Chen, M. J. Wainwright, Fast low-rank estimation by projected gradient descent: General statistical and algorithmic guarantees. arXiv:1509.03025 (10 September 2015).
11. C. Ma, K. Wang, Y. Chi, Y. Chen, Implicit regularization in nonconvex statistical estimation: Gradient descent converges linearly for phase retrieval, matrix completion, and blind deconvolution. *Found. Comput. Math.,* 10.1007/s10208-019-09429-9 (2019).
12. A. Carpentier, O. Klopp, M. Löffler, R. Nickl, Adaptive confidence sets for matrix completion. *Bernoulli* **24**, 2429–2460 (2018).
13. Y. Chen, Y. Chi, J. Fan, C. Ma, Y. Yan, Noisy matrix completion: Understanding statistical guarantees for convex relaxation via nonconvex optimization. arXiv:1902.07698 (20 February 2019).
14. C. H. Zhang, S. S. Zhang, Confidence intervals for low dimensional parameters in high dimensional linear models. *J. R. Stat. Soc. Ser. B* **76**, 217–242 (2014).
15. A. Javanmard, A. Montanari, Confidence intervals and hypothesis testing for high-dimensional regression. *J. Mach. Learn. Res.* **15**, 2869–2909 (2014).
16. R. Lockhart, J. Taylor, R. J. Tibshirani, R. Tibshirani, A significance test for the lasso. *Ann. Stat.* **42**, 413–468 (2014).
17. S. van de Geer, P. Bühlmann, Y. Ritov, R. Dezeure, On asymptotically optimal confidence regions and tests for high-dimensional models. *Ann. Stat.* **42**, 1166–1202 (2014).
18. T. T. Cai, Z. Guo, Confidence intervals for high-dimensional linear regression: Minimax rates and adaptivity. *Ann. Stat.* **45**, 615–646 (2017).
19. Y. Ning, H. Liu, A general theory of hypothesis tests and confidence regions for sparse high dimensional models. *Ann. Stat.* **45**, 158–195 (2017).
20. J. Jankova, S. Van De Geer, Confidence intervals for high-dimensional inverse covariance estimation. *Electron. J. Stat.* **9**, 1205–1229 (2015).
21. Z. Ren, T. Sun, C. Zhang, H. Zhou, Asymptotic normality and optimalities in estimation of large Gaussian graphical models. *Ann. Stat.* **43**, 991–1026 (2015).
22. B. Recht, M. Fazel, P. A. Parrilo, Guaranteed minimum-rank solutions of linear matrix equations via nuclear norm minimization. *SIAM Rev.* **52**, 471–501 (2010).
23. N. Srebro, A. Shraibman, "Rank, trace-norm and max-norm" in *International Conference on Computational Learning Theory*, P. Auer, R. Meir, Eds. (Springer, Berlin, Germany, 2005), pp. 545–560.
24. R. Mazumder, T. Hastie, R. Tibshirani, Spectral regularization algorithms for learning large incomplete matrices. *J. Mach. Learn. Res.* **11**, 2287–2322 (2010).
25. R. Sun, Z. Q. Luo, Guaranteed matrix completion via non-convex factorization. *IEEE Trans. Inf. Theory* **62**, 6535–6579 (2016).
26. Y. Chi, Y. M. Lu, Y. Chen, Nonconvex optimization meets low-rank matrix factorization: An overview. *IEEE Trans Signal Process* **67**, 5239–5269 (2019).
27. A. M. C. So, Y. Ye, Theory of semidefinite programming for sensor network localization. *Math. Program.* **109**, 367–384 (2007).
28. E. Abbe, J. Fan, K. Wang, Y. Zhong, Entrywise eigenvector analysis of random matrices with low expected rank. arXiv:1709.09565 (2 May 2017).
29. A. Singer, Angular synchronization by eigenvectors and semidefinite programming. *Appl. Comput. Harmon. Anal.* **30**, 20–36 (2011).
30. J. Fan, Y. Liao, M. Mincheva, Large covariance estimation by thresholding principal orthogonal complements. *J. R. Stat. Soc. Ser. B* **75**, 603–680 (2013).
31. National Climatic Data Center. https://www.ncdc.noaa.gov/ (31 August 2019).