Received: 16 September 2019 | Accepted: 3 September 2020

DOI: 10.1111/rssb.12400

ROYAL STATISTICAL SOCIETY
Series B Statistical Methodology

ORIGINAL ARTICLE

Statistical inferences of linear forms for noisy matrix completion

Dong Xia[1] | **Ming Yuan**[2]

[1]Hong Kong University of Science and Technology, Hong Kong, China

[2]Columbia University, New York, NY, USA

Correspondence
Dong Xia, Hong Kong University of Science and Technology, Hong Kong, China.
Email: madxia@ust.hk

Abstract

We introduce a flexible framework for making inferences about general linear forms of a large matrix based on noisy observations of a subset of its entries. In particular, under mild regularity conditions, we develop a universal procedure to construct asymptotically normal estimators of its linear forms through double-sample debiasing and low-rank projection whenever an entry-wise consistent estimator of the matrix is available. These estimators allow us to subsequently construct confidence intervals for and test hypotheses about the linear forms. Our proposal was motivated by a careful perturbation analysis of the empirical singular spaces under the noisy matrix completion model which might be of independent interest. The practical merits of our proposed inference procedure are demonstrated on both simulated and real-world data examples.

KEYWORDS

confidence interval, linear forms, matrix completion

1 | INTRODUCTION

Noisy matrix completion (NMC) refers to the reconstruction of a low-rank matrix $M \in \mathbb{R}^{d_1 \times d_2}$ after observing a small subset of M's entries with random noise. Problems of this nature arise naturally in various applications. For the sake of generality, we shall cast it in the framework of trace regression where each observation is a random pair (X, Y) with $X \in \mathbb{R}^{d_1 \times d_2}$ and $Y \in \mathbb{R}$. The random matrix X is sampled uniformly from the orthonormal basis $\mathcal{E} = \{e_{j_1} e_{j_2}^{\mathsf{T}} : j_1 \in [d_1], j_2 \in [d_2]\}$, where $[d] = \{1, \cdots, d\}$ and $\{e_{j_1}\}_{j_1 \in [d_1]}$ and $\{e_{j_2}\}_{j_2 \in [d_2]}$ are the canonical basis vectors in $\mathbb{R}^{d_1}$ and $\mathbb{R}^{d_2}$, respectively. It is worth pointing out that, while we shall focus on the canonical basis in this work, our framework can be easily extended to general product basis where $\{e_{j_1}\}_{j_1 \in [d_1]}$ and $\{e_{j_2}\}_{j_2 \in [d_2]}$ are arbitrary orthonormal

basis in $\mathbb{R}^{d_1}$ and $\mathbb{R}^{d_2}$, respectively. Without loss of generality, we shall assume that $d_1 \geq d_2$ and denote $\alpha_d = d_1/d_2$ the aspect ratio of M. The response variable Y is related to X via

$$Y = \langle M, X \rangle + \xi \tag{1}$$

where $\langle M, X \rangle = \mathrm{Tr}(M^{\mathsf{T}} X)$, and the independent measurement error ξ is assumed to be a centred sub-Gaussian random variable. Our goal is to infer M from n i.i.d. copies $\{(X_i, Y_i)\}_{i=1}^n$ obeying (1) when, in particular, M is of (approximately) low rank and n is much smaller than $d_1 d_2$.

In the absence of measurement error (e.g. $\xi = 0$), Candès and Recht (2009) first discovered that exact matrix completion can be solved efficiently by relaxing the non-convex and non-smooth rank constraint of a matrix to its nuclear norm. Following the pioneering work, nuclear-norm penalized least squares estimators (Cai & Zhou, 2016; Cai et al., 2010; Candès & Tao, 2009; Candes & Plan, 2010; Gross, 2011; Negahban & Wainwright, 2011; Rohde and Tsybakov, 2011) and numerous other variants (Cai & Zhang, 2015; Gao et al., 2016; Klopp, 2014; Koltchinskii et al., 2011; Liu, 2011; Recht et al., 2010; Sun & Zhang, 2012) have been studied. It is now understood, from these developments, that even when the observations are contaminated with noise, statistically optimal convergence rates are attainable by efficiently computable convex methods. For instance, Koltchinskii et al. (2011) proved that a modified matrix LASSO estimator, denoted by $\widehat{M}^{\mathsf{KLT}}$, achieves the convergence rate:

$$\|\widehat{M}^{\mathsf{KLT}} - M\|_{\mathsf{F}}^2 = O_P\left((\sigma_\xi + \|M\|_{\mathsf{max}})^2 \cdot \frac{r d_1^2 d_2 \log d_1}{n} \right) \tag{2}$$

as long as $n \gg d_1 \log d_1$, where r is the rank of M and σ_ξ^2 is the variance of ξ. Here, $\|\cdot\|_{\mathsf{F}}$ denotes the matrix Frobenius norm and $\|\cdot\|_{\mathsf{max}}$ denotes the max-norm defined as $\|A\|_{\mathsf{max}} = \max_{j_1 \in [d_1], j_2 \in [d_2]} |A(j_1, j_2)|$. It is worth noting that (2) was established without additional assumptions on M. As a result, the rate given on the right-hand side of (2) depends on $\|M\|_{\mathsf{max}}$ and does not vanish even when $\sigma_\xi = 0$.

In addition to convex methods, non-convex approaches such as those based on matrix factorization have also been developed. For instance, Keshavan, Montanari and Oh (2010b) proposed a non-convex estimator based on the thin singular value decomposition (SVD), denoted by $\widehat{M}^{\mathsf{KMO}}$, and show that

$$\|\widehat{M}^{\mathsf{KMO}} - M\|_{\mathsf{F}}^2 = O_P(\sigma_\xi^2 \cdot \frac{r d_1^2 d_2 \log d_1}{n}) \tag{3}$$

assuming that $n \gg r d_1 (r + \log d_1)$ and M satisfies the so-called incoherent condition. See also, for example Zhao et al. (2015), Chen and Wainwright (2015), Cai et al. (2016b) and references therein. The rate (3) is optimal up to the logarithmic factors, see, for example Koltchinskii et al. (2011) and Ma and Wu (2015), for a comparable minimax lower bound. More recently, an alternative scheme of matrix factorization attracted much attention. See, for example Wang et al. (2016); Ge et al. (2016); Zheng and Lafferty (2016); Chen et al. (2019c, 2019b); Ma et al. (2017); Chen et al. (2019a). In particular, Ma et al. (2017) showed this approach yields an estimator, denoted by $\widehat{M}^{\mathsf{MWC}}$, that is statistically optimal not only in matrix Frobenius norm but also in entry-wise max-norm, that is

$$\|\widehat{M}^{\mathsf{MWC}} - M\|_{\mathsf{max}}^2 = O_P(\sigma_\xi^2 \cdot \frac{r d_1 \log d_1}{n}) \tag{4}$$

provided that $n \gg r^3 d_1 \log^3 d_1$.

While there is a rich literature on statistical estimation for NMC, results about its statistical inferences are relatively scarce. In Carpentier et al. (2015), a debiasing procedure, based on sample splitting, was proposed for the nuclear norm penalized least squares estimator which enables constructing confidence region for M with respect to matrix Frobenius norm when $n \gg rd_1 \log d_1$. Their technique, however, cannot be directly used to make inferences about individual entries or linear forms as confidence regions for M with respect to matrix Frobenius norm can be too wide for such purposes. To this end, Carpentier et al. (2018) proposed another procedure to construct entry-wise confidence intervals. However, their procedure requires that the design, namely the underlying distribution of X satisfy the so-called *restricted isometry property* which is violated when X is sampled uniformly from $\mathcal{E}$. Another proposal introduced by Cai et al. (2016a) can be used to construct confidence intervals for M's entries. However, it requires that the sample size $n \gg d_1 d_2$ which is significantly larger than the optimal sample size requirement for estimation. In addition, during the preparation of the current work, Chen et al. (2019c) announced a different approach to constructing confidence intervals for the entries of M.

The present article aims to further expand this line of research by introducing a flexible framework for constructing confidence intervals and testing hypotheses about general linear forms of M, with its entries as special cases, under optimal sample size requirement. In a nutshell, we develop a procedure that, given any entry-wise consistent estimator $\widehat{M}^{\text{init}}$ in that $\|\widehat{M}^{\text{init}} - M\|_{\max} = o_P(\sigma_\xi)$, can yield valid statistical inferences for $m_T := \mathrm{Tr}(M^\top T)$ under mild regularity conditions. More specifically, we show that, through double-sample debiasing and spectral projection, we can obtain from the initial estimator a new one, denoted by $\widehat{M}$, so that

$$\frac{\mathrm{Tr}(\widehat{M}^\top T) - \mathrm{Tr}(M^\top T)}{\sigma_\xi (\|U^\top T\|_{\mathsf{F}}^2 + \|TV\|_{\mathsf{F}}^2)^{1/2} \sqrt{d_1 d_2 / n}} \xrightarrow{\mathrm{d}} \mathcal{N}(0,1), \tag{5}$$

provided that

$$\|U^\top T\|_{\mathsf{F}} + \|TV\|_{\mathsf{F}} \gg \|T\|_{\ell_1} \sqrt{\frac{r}{d_1}} \cdot max \left\{ \sqrt{\frac{r \log d_1}{d_2}},\ \frac{\sigma_\xi}{\lambda_r} \sqrt{\frac{\alpha_d r d_1^2 d_2 \log^2 d_1}{n}} \right\}$$

where U and V are M's left and right singular vectors and λ_r is its r-th singular value, and $\|\cdot\|_{\ell_1}$ stands for the vectorized ℓ_1 norm. We not only show that (5) holds under optimal sample size (independent of T) but also derive its non-asymptotic convergence rate explicitly. Note that condition for $\|U^\top T\|_{\mathsf{F}} + \|TV\|_{\mathsf{F}}$ in a certain sense is necessary to avoid non-regular asymptotic behaviour when $\|U^\top T\|_{\mathsf{F}} + \|TV\|_{\mathsf{F}} = 0$. Moreover, we show that under similar conditions, (5) continues to hold when we replace σ_ξ, $\|U^\top T\|_{\mathsf{F}}$ and $\|TV\|_{\mathsf{F}}$ by suitable estimates, denoted by $\widehat{\sigma}_\xi$, $\|\widehat{U}^\top T\|_{\mathsf{F}}$ and $\|T\widehat{V}\|_{\mathsf{F}}$, respectively:

$$\frac{\mathrm{Tr}(\widehat{M}^\top T) - \mathrm{Tr}(M^\top T)}{\widehat{\sigma}_\xi (\|\widehat{U}^\top T\|_{\mathsf{F}}^2 + \|T\widehat{V}\|_{\mathsf{F}}^2)^{1/2} \sqrt{d_1 d_2 / n}} \xrightarrow{\mathrm{d}} \mathcal{N}(0,1). \tag{6}$$

The statistic on the left-hand side is now readily applicable for making inferences about the linear form $\mathrm{Tr}(M^\top T)$.

Our proposal greatly generalizes the scope of earlier works on inferences for entries of M in several crucial aspects. First, unlike earlier approaches that focus on a specific estimator of M, our procedure

can be applied to any entry-wise consistent estimator. This not only brings potential practical benefits but also helps us better understand the fundamental differences between estimation and testing in the context of NMC. For instance, our results suggest that, perhaps surprisingly, when it comes to make valid inferences with optimal sample sizes, the rate of convergence of the initial estimate is irrelevant as long as it is consistent; therefore, a suboptimal estimator may be used for making optimal inferences.

Second, our approach can be applied in general when T is sparse, and depending on its alignment with the singular spaces of M, even to cases where it is dense and $\|T\|_{\ell_1}^2/\|T\|_{\mathsf{F}}^2$ is of the order $O(d_2)$. Entry-wise inferences correspond to the special case when T takes the form $e_i e_j^{\mathsf{T}}$. Extensions to more general linear forms could prove useful in many applications. For example, in recommender systems, it may be of interest to decide between items j_1 and j_2 which should we recommend to user i. This can obviously be formulated as a testing problem:

$$H_0: M(i,j_1) = M(i,j_2) \quad \text{v. s.} \quad H_1: M(i,j_1) > M(i,j_2), \tag{7}$$

which can be easily solved within our framework by taking $T = e_i e_{j_1}^{\mathsf{T}} - e_i e_{j_2}^{\mathsf{T}}$. More generally, if the target is a group of users $\mathcal{G} \subset [d_1]$, we might take a linear form $T = \sum_{i\in\mathcal{G}} e_i (e_{j_1} - e_{j_2})^{\mathsf{T}}$. At a technical level, inferences about general linear forms as opposed to entries of M present nontrivial challenges because of the complex dependence structure among the estimated entries. As our theoretical analysis shows, the variance of the plug-in estimator for the linear form depends on the alignment of the linear form with respect to the singular space of M rather than the sparsity of the linear form.

An essential part of our technical development is the characterization of the distribution of the empirical singular vectors for NMC where we take advantage of the recently developed spectral representation for empirical singular vectors. Similar tools have been used earlier to derive confidence regions for singular subspaces with respect to ℓ_2-norm for low-rank matrix regression (LMR) when the linear measurement matrix Xs are Gaussian (Xia, 2019a), and the planted low-rank matrix (PLM) model where every entry of M is observed with i.i.d. Gaussian noise (Xia, 2019b). In both cases, Gaussian assumption plays a critical role, and furthermore, it was observed that first-order approximation may lead to suboptimal performances. In the absence of the Gaussian assumption, the treatment of NMC is technically more challenging and requires us to derive sharp bounds for the $(2,\max)$-norm for the higher order perturbation terms. Interestingly, it turns out that, unlike LMR or PLM, a first-order approximation actually suffices for NMC.

Even though our framework applies to any max-norm consistent matrix estimator, for concreteness, we introduce a novel rotation calibrated gradient descent algorithm on Grassmannians that yields such an initial estimator. The rotation calibration promotes fast convergence on Grassmannians so that constant stepsize can be selected to guarantee geometric convergence. We note that existing results on max-norm convergence rates are established for sampling without replacement (Ma et al., 2017). It is plausible that (4) may continue to hold under our assumption of independent sampling given the close connection between the two sampling schemes, but an actual proof is likely much more involved and therefore we opted for the proposed alternative for illustration as it is more amenable for analysis.

The rest of our paper is organized as follows. In next section, we present a general framework for estimating $m_T = \mathrm{Tr}(M^{\mathsf{T}}T)$ given an initial estimator through double-sample debiasing and spectral projection. In Section 3, we establish the asymptotic normality of the estimate obtained. In Section 4, we propose data-driven estimates for the noise variance and the true singular vectors, based on which confidence intervals of m_T are constructed. In Section 5, we introduce a rotation calibrated gradient

descent algorithm on Grassmannians, which, under mild conditions, provides the initial estimator $\widehat{M}^{\text{init}}$ so that $\|\widehat{M}^{\text{init}} - M\|_{\max} = o_P(\sigma_\xi)$. Numerical experiments on both synthetic and real-world data sets presented in Section 6 further demonstrate the merits of the proposed methodology. All proofs are presented in the online supplement.

2 | ESTIMATING LINEAR FORMS

We are interested in making inferences about $m_T = \text{Tr}(M^\top T)$ for a given T based on observations $\mathcal{D} = \{(X_i, Y_i): 1 \le i \le n\}$ satisfying model (1), assuming that M has low rank. To this end, we first need to construct an appropriate estimate of m_T which we shall do in this section.

Without loss of generality, we assume n is an even number with $n = 2n_0$, and split $\mathcal{D}$ into two subsamples:

$$\mathcal{D}_1 = \{(X_i, Y_i)\}_{i=1}^{n_0} \quad \text{and} \quad \mathcal{D}_2 = \{(X_i, Y_i)\}_{i=n_0+1}^{n}.$$

In what follows, we shall denote M's thin SVD by $M = U\Lambda V^\top$, where $U \in \mathbb{O}^{d_1 \times r}$, $V \in \mathbb{O}^{d_2 \times r}$ and $\Lambda = \text{diag}(\lambda_1, \cdots, \lambda_r)$ represent M's singular vectors and singular values, respectively. The Stiefel manifold $\mathbb{O}^{d \times r}$ is defined as $\mathbb{O}^{d \times r} := \{A \in \mathbb{R}^{d \times r}: A^\top A = I\}$. We arrange M's positive singular values non-increasingly, that is $\lambda_1 \ge \cdots \ge \lambda_r > 0$.

Assuming the availability of an initial estimator, our procedure consists of four steps as follows:

- *Step 1 (Initialization)*: By utilizing the first and second data subsample $\mathcal{D}_1, \mathcal{D}_2$ separately, we apply the initial estimating procedure on noisy matrix completion to yield initial (biased in general) estimates $\widehat{M}_1^{\text{init}}$ and $\widehat{M}_2^{\text{init}}$, respectively.
- *Step 2 (Debiasing)*: Using the second data subsample $\mathcal{D}_2$, we debias $\widehat{M}_1^{\text{init}}$:

$$\widehat{M}_1^{\text{unbs}} = \widehat{M}_1^{\text{init}} + \frac{d_1 d_2}{n_0} \sum_{i=n_0+1}^{n} (Y_i - \langle \widehat{M}_1^{\text{init}}, X_i \rangle) X_i.$$

Similarly, we use the first data subsample $\mathcal{D}_1$ to debias $\widehat{M}_2^{\text{init}}$ and obtain

$$\widehat{M}_2^{\text{unbs}} = \widehat{M}_2^{\text{init}} + \frac{d_1 d_2}{n_0} \sum_{i=1}^{n_0} (Y_i - \langle \widehat{M}_2^{\text{init}}, X_i \rangle) X_i.$$

- *Step 3 (Projection)*: Compute the top-r left and right singular vectors of $\widehat{M}_1^{\text{unbs}}$, denoted by $\widehat{U}_1$ and $\widehat{V}_1$. Similarly, compute the top-r left and right singular vectors of $\widehat{M}_2^{\text{unbs}}$, denoted by $\widehat{U}_2$ and $\widehat{V}_2$. Then, we calculate the (averaged) projection estimate

$$\widehat{M} = \frac{1}{2}\widehat{U}_1\widehat{U}_1^\top \widehat{M}_1^{\text{unbs}} \widehat{V}_1\widehat{V}_1^\top + \frac{1}{2}\widehat{U}_2\widehat{U}_2^\top \widehat{M}_2^{\text{unbs}} \widehat{V}_2\widehat{V}_2^\top.$$

- *Step 4 (Plug-in)*: Finally, we estimate m_T by $\widehat{m}_T = \text{Tr}(\widehat{M}^\top T)$.

We now discuss each of the steps in further details.

2.1 | Initialization

Apparently, our final estimate depends on the initial estimates $\widehat{M}_1^{\text{init}}, \widehat{M}_2^{\text{init}}$. However, as we shall show in the next section, such dependence is fairly weak and the resulting estimate $\widehat{m}_T$ is asymptotically equivalent as long as the estimation error of $\widehat{M}_1^{\text{init}}$ and $\widehat{M}_2^{\text{init}}$, in terms of max-norm, is of a smaller order than σ_ξ. More specifically, we shall assume that

Assumption 1 There exists a sequence $\gamma_{n,d_1,d_2} \to 0$ as $n, d_1, d_2 \to \infty$ so that with probability at least $1 - d_1^{-2}$,

$$\|\widehat{M}_1^{\text{init}} - M\|_{\max} + \|\widehat{M}_2^{\text{init}} - M\|_{\max} \le C\gamma_{n,d_1,d_2} \cdot \sigma_\xi \tag{8}$$

for an absolute constant $C > 0$.

In particular, bounds similar to (8) have recently been established by Ma et al. (2017), Chen et al. (2019c). See Equation (4). Assumption 1 was motivated by their results. However, as noted earlier, (4) was obtained under sampling without replacement and for positively semidefinite matrices. While it is plausible that it also holds under independent sampling as considered here, an actual proof is lacking at this point. For concreteness, we shall present a simple algorithm in Section 5 capable of producing an initial estimate that satisfies Assumption 1.

2.2 | Debiasing

The initial estimate is only assumed to be consistent. It may not necessarily be unbiased or optimal. To ensure good quality of our final estimate $\widehat{m}_T$, it is important that we first debias it which allows for sharp spectral perturbation analysis. Debiasing is an essential technique in statistical inferences of high-dimensional sparse linear regression (see, e.g. Cai & Guo, 2017; Javanmard & Montanari, 2014; Van de Geer et al, 2014; Zhang & Zhang, 2014) and low-rank matrix regression (see, e.g. Cai et al., 2016a; Carpentier & Kim, 2018; Carpentier et al., 2018; Xia, 2019a). Oftentimes, debiasing is done in the absence of the knowledge of $\mathbb{E}\text{vec}(X)\text{vec}(X)^\top$ and a crucial step is to construct an appropriate decorrelating matrix. In our setting, it is clear that $\mathbb{E}\text{vec}(X)\text{vec}(X)^\top = (d_1 d_2)^{-1} I_{d_1 d_2}$. This allows for a much simplified treatment via sample splitting, in the same spirit as earlier works including Carpentier et al. (2015), Xia (2019a), among others. The particular double-sample-splitting technique we employ was first proposed by Chernozhukov et al. (2018) and avoids the loss of statistical efficiency associated with the sample splitting. It is worth noting that if the entries are not sampled uniformly, the debiasing procedure needs to be calibrated accordingly.

In addition to reducing possible bias of the initial estimate, the sample splitting also enables us to extend the recently developed spectral representation for empirical singular vectors under Gaussian assumptions to general sub-Gaussian distributions.

2.3 | Spectral projection

Since M have low rank, it is natural to apply spectral truncation to a matrix estimate to yield an improved estimate. To this end, we project $\widehat{M}_1^{\text{unbs}}$ and $\widehat{M}_2^{\text{unbs}}$ to their respective leading singular subspaces. Note that, while $\widehat{M}_1^{\text{unbs}}, \widehat{M}_2^{\text{unbs}}$ are unbiased, their empirical singular vectors $\widehat{U}_1, \widehat{U}_2, \widehat{V}_1$ and $\widehat{V}_2$

are typically not. The spectral projection serves the purpose of reducing entry-wise variances at the cost of negligible biases.

It is worth noting that the estimate $\widehat{M}$ may not be of rank r. If an exact rank-r estimator is desired, it suffices to obtain the best rank-r approximation of $\widehat{M}$ via singular value decomposition and all our development remains valid under such a modification. In general, getting the initial estimates is the most computational expensive step as the other steps involving fairly standard operation without incurring any challenging optimization. This is noteworthy because it suggests that as long as we can compute a good estimate, it does not cost much more computationally to make inferences.

3 | ASYMPTOTIC NORMALITY OF $\widehat{m}_T$

We now show the estimate $\widehat{m}_T$ we derived in the previous section is indeed suitable for inferences about m_T by establishing its asymptotic normality.

3.1 | General results

For brevity, let e_j denote the j-th canonical basis in $\mathbb{R}^d$ where d might be d_1 or d_2 or d_1+d_2 at different appearances. With slight abuse of notation, denote by $\|\cdot\|$ the matrix operator norm or vector ℓ_2-norm depending on the dimension of its argument. Denote the condition number of M by

$$\kappa(M)=\lambda_1(M)/\lambda_r(M)=\lambda_1/\lambda_r. \tag{9}$$

As is conventional in the literature, we shall assume implicitly that rank r is known with $r \ll d_2$ and M is well-conditioned so that $\kappa(M)\le\kappa_0$. In practice, r is usually not known in advance and needs to be estimated from the data. Our experience with numerical experiments such as those reported in Section 6 suggests that our procedure is generally robust to reasonable estimate of r. Although a more rigorous justification of such a phenomenon has thus far eluded us, these promising empirical observations nonetheless indicate a more careful future investigation is warranted.

In addition, we shall assume that U and V are incoherent, a standard condition for matrix completion.

Assumption 2 Let $\|U\|_{2,\max}=\max_{j\in[d_1]}\|e_j^{\mathsf{T}}U\|$ and there exists $\mu_{\max}>0$ so that

$$\max\{\sqrt{\frac{d_1}{r}}\|U\|_{2,\max},\sqrt{\frac{d_2}{r}}\|V\|_{2,\max}\}\le\mu_{\max}.$$

We also assume that the noise ξ is independent with X and sub-Gaussian such that

Assumption 3 The noise ξ is independent with X and

$$\mathbb{E}\xi=0,\quad \mathbb{E}\xi^2=\sigma_\xi^2,\quad \text{and}\quad \mathbb{E}e^{s\xi}\le e^{s^2\sigma_\xi^2},\ \forall s\in\mathbb{R} \tag{10}$$

Let $\alpha_d=d_1/d_2$. There exists a large enough absolute constant $C_1>0$ so that

$$\lambda_r \geq C_1 \mu_{\max} \kappa_0^2 \sigma_\xi \sqrt{\frac{\alpha_d r d_1^2 d_2 \log^2 d_1}{n}}. \tag{11}$$

The SNR condition (11) is optimal up to the logarithmic factors if $\alpha_d, \mu_{\max}, \kappa_0 = O(1)$. Indeed, the consistent estimation of singular subspaces requires $\lambda_r \gg \sigma_\xi \sqrt{r d_1^2 d_2 / n}$. This condition is common for non-convex methods of NMC. However, when $\alpha_d \gg 1$, that is M is highly rectangular, condition (11) is significantly stronger than the optimal SNR condition even if $\mu_{\max}, \kappa_0 = O(1)$. It is unclear to us whether this suboptimality is due to technical issues or reflection of more fundamental differences between statistical estimation and inference.

To avoid the non-regular asymptotics, we focus on the case when T does not lie entirely in the null space of M. More specifically, we assume that

Assumption 4 There exists a constant $\alpha_T > 0$ such that

$$\|U^{\mathsf{T}} T\|_{\mathsf{F}} \geq \alpha_T \|T\|_{\mathsf{F}} \cdot \sqrt{\frac{r}{d_1}} \quad \text{or} \quad \|TV\|_{\mathsf{F}} \geq \alpha_T \|T\|_{\mathsf{F}} \cdot \sqrt{\frac{r}{d_2}}.$$

The alignment parameter α_T in Assumption 4 is allowed to vanish as $d_1, d_2, n \to \infty$. Indeed, as we show below, the asymptotic normality of $\widehat{m}_T - m_T$ only requires that

$$\alpha_T \geq \frac{\|T\|_{\ell_1}}{\|T\|_{\mathsf{F}}} \cdot \max\left\{ \mu_{\max}^2 \sqrt{\frac{r \log d_1}{d_2}},\ \frac{\kappa_0 \mu_{\max}^2 \sigma_\xi}{\lambda_r} \sqrt{\frac{\alpha_d r d_1^2 d_2 \log^2 d_1}{n}} \right\}. \tag{12}$$

We are now in position to establish the asymptotic normality of $\widehat{m}_T$.

Theorem 1 *Under Assumptions 1–4, there exist absolute constants $C_1, C_2, C_3, C_4, C_5, C_6 > 0$ so that if $n \geq C_1 \mu_{\max}^2 r d_1 \log d_1$, then*

$$\begin{aligned}
&\sup_{x \in \mathbb{R}} \left| \mathbb{P}\left(\frac{\widehat{m}_T - m_T}{\sigma_\xi (\|TV\|_{\mathsf{F}}^2 + \|U^{\mathsf{T}} T\|_{\mathsf{F}}^2)^{1/2} \cdot \sqrt{d_1 d_2 / n}} \leq x \right) - \Phi(x) \right| \\
&\quad \leq C_2 \frac{\mu_{\max}^2 \|T\|_{\ell_1}}{\alpha_T \|T\|_{\mathsf{F}}} \sqrt{\frac{\log d_1}{d_2}} + C_3 \kappa_0 \frac{\mu_{\max}^2 \|T\|_{\ell_1}}{\alpha_T \|T\|_{\mathsf{F}}} \cdot \frac{\sigma_\xi}{\lambda_r} \sqrt{\frac{\alpha_d r d_1^2 d_2 \log^2 d_1}{n}} \\
&\quad + C_4 \frac{\mu_{\max}^4 \|T\|_{\ell_1}^2}{\alpha_T^2 \|T\|_{\mathsf{F}}^2} \cdot \frac{r \sqrt{\log d_1}}{d_2} + \frac{6 \log d_1}{d_1^2} + C_5 \gamma_{n, d_1, d_2} \sqrt{\log d_1} + C_6 \mu_{\max} \sqrt{\frac{r d_1}{n}}.
\end{aligned}$$

where $\Phi(x)$ denotes the c.d.f. of the standard normal distribution.

By Theorem 1, if $\mu_{\max}, \alpha_d, \kappa_0 = O(1)$ and

$$max\left\{ \frac{\|T\|_{\ell_1}}{\alpha_T \|T\|_{\mathsf{F}}} \sqrt{\frac{r \log d_1}{d_2}}, \frac{\|T\|_{\ell_1}}{\alpha_T \|T\|_{\mathsf{F}}} \cdot \frac{\sigma_\xi}{\lambda_r} \sqrt{\frac{r d_1^2 d_2 \log^2 d_1}{n}}, \gamma_{n, d_1, d_2} \sqrt{\log d_1} \right\} \to 0, \tag{13}$$

then

$$\frac{\widehat{m}_T - m_T}{\sigma_\xi(\|TV\|_{\mathsf{F}}^2 + \|U^{\mathsf{T}}T\|_{\mathsf{F}}^2)^{1/2} \cdot \sqrt{d_1 d_2/n}} \xrightarrow{\mathrm{d}} \mathcal{N}(0,1),$$

as $n, d_1, d_2 \to \infty$.

3.2 | Specific examples

We now consider several specific linear forms to further illustrate the implications of Theorem 1.

Example 1 As noted before, among the simplest linear forms are entries of M. In particular, $M(i,j) = \langle M, T\rangle$ with $T = e_i e_j^{\mathsf{T}}$. It is clear that $\|T\|_{\ell_1} = \|T\|_{\mathsf{F}} = 1$ and Assumption 4 is equivalent to

$$\|e_i^{\mathsf{T}}U\| + \|e_j^{\mathsf{T}}V\| \geq \alpha_T \sqrt{\frac{r}{d_1}}. \tag{14}$$

Theorem 1 immediately implies that

$$\frac{\widehat{M}(i,j) - M(i,j)}{(\|e_i^{\mathsf{T}}U\|^2 + \|e_j^{\mathsf{T}}V\|^2)^{1/2} \cdot \sigma_\xi \sqrt{d_1 d_2/n}} \xrightarrow{\mathrm{d}} \mathcal{N}(0,1),$$

provided that

$$\max\{\frac{\mu_{\max}^2}{\alpha_T}\sqrt{\frac{r\log d_1}{d_2}},\ \frac{\kappa_0\mu_{\max}^2}{\alpha_T}\cdot\frac{\sigma_\xi}{\lambda_r}\sqrt{\frac{\alpha_d r d_1^2 d_2 \log^2 d_1}{n}},\ \gamma_{n,d_1,d_2}\sqrt{\log d_1}\} \to 0 \tag{15}$$

as $n, d_1, d_2 \to \infty$.

We can also infer from the entry-wise asymptotic normality that

$$\mathbb{E}\|\widehat{M} - M\|_{\mathsf{F}}^2 = (1 + o(1)) \cdot \frac{\sigma_\xi^2 r d_1 d_2 (d_1 + d_2)}{n}. \tag{16}$$

The mean squared error on the right-hand side is sharply optimal and matches the minimax lower bound in Koltchinskii et al. (2011).

Example 2 In the case when we want to compare $M(i,j_1)$ and $M(i,j_2)$, we can take $T = e_i e_{j_1}^{\mathsf{T}} - e_i e_{j_2}^{\mathsf{T}}$. Because $\|T\|_{\ell_1}/\|T\|_{\mathsf{F}} = \sqrt{2}$, Assumption 4 then becomes

$$\|TV\|_{\mathsf{F}}^2 + \|U^{\mathsf{T}}T\|_{\mathsf{F}}^2 = 2\|U^{\mathsf{T}}e_i\|^2 + \|V^{\mathsf{T}}(e_{j_1} - e_{j_2})\|^2 \geq \frac{2\alpha_T^2 r}{d_1}. \tag{17}$$

Theorem 1 therefore implies that

$$\frac{(\widehat{M}(i,j_1) - \widehat{M}(i,j_2)) - (M(i,j_1) - M(i,j_2))}{(2\|U^{\mathsf{T}}e_i\|^2 + \|V^{\mathsf{T}}(e_{j_1} - e_{j_2})\|^2)^{1/2} \cdot \sigma_\xi \sqrt{d_1 d_2/n}} \xrightarrow{\mathrm{d}} \mathcal{N}(0,1),$$

provided that

$$\max\left\{\frac{\mu_{\max}^2}{\alpha_T}\sqrt{\frac{r\log d_1}{d_2}},\ \frac{\kappa_0\mu_{\max}^2}{\alpha_T}\cdot\frac{\sigma_\xi}{\lambda_r}\sqrt{\frac{\alpha_d r d_1^2 d_2\log^2 d_1}{n}},\ \gamma_{n,d_1,d_2}\sqrt{\log d_1}\right\}\to 0. \tag{18}$$

Example 3 More generally, we can consider the case when T is sparse in that it has up to s_0 nonzero entries. By Cauchy–Schwartz inequality, $\|T\|_{\ell_1}/\|T\|_{\mathsf{F}}\le\sqrt{s_0}$ so that Assumption 4 holds. By Theorem 1,

$$\frac{\widehat{m}_T - m_T}{\sigma_\xi(\|TV\|_{\mathsf{F}}^2+\|U^{\mathsf{T}}T\|_{\mathsf{F}}^2)^{1/2}\cdot\sqrt{d_1d_2/n}}\xrightarrow{\mathrm{d}}\mathcal{N}(0,1),$$

as long as

$$\max\left\{\frac{\mu_{\max}^2}{\alpha_T}\sqrt{\frac{s_0 r\log d_1}{d_2}},\ \frac{\kappa_0\mu_{\max}^2}{\alpha_T}\cdot\frac{\sigma_\xi}{\lambda_r}\sqrt{\frac{s_0\alpha_d r d_1^2 d_2\log^2 d_1}{n}},\ \gamma_{n,d_1,d_2}\sqrt{\log d_1}\right\}\to 0. \tag{19}$$

It is of interest to consider the effect of alignment of T with respect to the singular spaces of M. Note that

$$\|T\|_{\mathsf{F}}^2=\|U^{\mathsf{T}}T\|_{\mathsf{F}}^2+\|U_\perp^{\mathsf{T}}T\|_{\mathsf{F}}^2=\|TV\|_{\mathsf{F}}^2+\|TV_\perp\|_{\mathsf{F}}^2,$$

where $U_\perp\in\mathbb{O}^{d_1\times(d_1-r)}$ and $V_\perp\in\mathbb{O}^{d_2\times(d_2-r)}$ are the basis of the orthogonal complement of U and V, respectively. In the case that T is not dominated by its projection onto $U_\perp$ or $V_\perp$ in that $\|U^{\mathsf{T}}T\|_{\mathsf{F}}+\|TV\|_{\mathsf{F}}$ is of the same order as $\|T\|_{\mathsf{F}}$, we can allow T to have as many as $O(d_2)$ non-zero entries.

4 | INFERENCES ABOUT LINEAR FORMS

The asymptotic normality of $\widehat{m}_T$ we showed in the previous section forms the basis for making inferences about m_T. To derive confidence intervals of or testing hypotheses about m_T, however, we need to also estimate the variance of $\widehat{m}_T$. To this end, we shall estimate the noise variance by

$$\widehat{\sigma}_\xi^2=\frac{1}{2n_0}\sum_{i=n_0+1}^{n}(Y_i-\langle\widehat{M}_1^{\mathsf{init}},X_i\rangle)^2+\frac{1}{2n_0}\sum_{i=1}^{n_0}(Y_i-\langle\widehat{M}_2^{\mathsf{init}},X_i\rangle)^2. \tag{20}$$

and $\|TV\|_{\mathsf{F}}^2+\|U^{\mathsf{T}}T\|_{\mathsf{F}}^2$ by

$$\widehat{s}_T^2:=\frac{1}{2}\left(\|T\widehat{V}_1\|_{\mathsf{F}}^2+\|\widehat{U}_1^{\mathsf{T}}T\|_{\mathsf{F}}^2+\|T\widehat{V}_2\|_{\mathsf{F}}^2+\|\widehat{U}_2^{\mathsf{T}}T\|_{\mathsf{F}}^2\right).$$

The following theorem shows that the asymptotic normality remains valid if we replace the variance of $\widehat{m}_T$ with these estimates:

Theorem 2 *Under Assumptions 1–4, if* $n\ge C\mu_{\max}^2 r d_1\log d_1$ *for some absolute constant* $C>0$ *and*

$$\max\{\frac{\mu_{\max}^2\|T\|_{\ell_1}}{\alpha_T\|T\|_{\mathsf{F}}}\sqrt{\frac{r\log d_1}{d_2}},\frac{\kappa_0\mu_{\max}^2\|T\|_{\ell_1}}{\alpha_T\|T\|_{\mathsf{F}}}\cdot\frac{\sigma_\xi}{\lambda_r}\sqrt{\frac{\alpha_d r d_1^2 d_2\log^2 d_1}{n}},\gamma_{n,d_1,d_2}\sqrt{\log d_1}\}\to 0,$$

then

$$\frac{\widehat{m}_T-m_T}{\widehat{\sigma}_\xi\widehat{s}_T\cdot\sqrt{d_1d_2/n}}\xrightarrow{\mathrm{d}}\mathcal{N}(0,1),$$

as $n,d_1,d_2\to\infty$.

Theorem 2 immediately allows for constructing confidence intervals for m_T. More specifically, we can define the $100(1-\theta)\%$-th confidence interval as

$$\widehat{\mathrm{CI}}_{\theta,T}=[\widehat{m}_T-z_{\theta/2}\cdot\widehat{\sigma}_\xi\widehat{s}_T\sqrt{\frac{d_1d_2}{n}},\quad\widehat{m}_T+z_{\theta/2}\cdot\widehat{\sigma}_\xi\widehat{s}_T\sqrt{\frac{d_1d_2}{n}}] \tag{21}$$

for any $\theta\in(0,1)$, where $z_\theta=\Phi^{-1}(1-\theta)$ is the upper θ quantile of the standard normal. In light of Theorem 2, we have

$$\lim_{n,d_1,d_2\to\infty}\mathbb{P}(m_T\in\widehat{\mathrm{CI}}_{\theta,T})=1-\theta,$$

for any $\theta\in(0,1)$.

Similarly, we can also consider using Theorem 2 for the purpose of hypothesis test. Consider, for example testing linear hypothesis

$$H_0:\langle M,T\rangle=0\qquad\text{against}\qquad H_1:\langle M,T\rangle\neq 0.$$

Then we can proceed to reject H_0 if $|\widehat{z}|>z_{\theta/2}$ and accept H_0 otherwise, where

$$\widehat{z}=\frac{\widehat{m}_T}{\widehat{\sigma}_\xi\widehat{s}_T\cdot\sqrt{d_1d_2/n}}.$$

Following Theorem 2, this is a test with asymptotic level θ. For example, in the particular case of comparing two entries of M:

$$H_0:M(i,j_1)=M(i,j_2)\quad\text{v. s.}\quad H_1:M(i,j_1)>M(i,j_2), \tag{22}$$

the test statistic can be expressed as

$$\widehat{z}=\frac{\sqrt{2}(\widehat{M}(i,j_1)-\widehat{M}(i,j_2))}{\widehat{\sigma}_\xi(\|\widehat{V}_1^{\mathsf{T}}(e_{j_2}-e_{j_1})\|_{\mathsf{F}}^2+2\|\widehat{U}_1^{\mathsf{T}}e_i\|_{\mathsf{F}}^2+\|\widehat{V}_2^{\mathsf{T}}(e_{j_2}-e_{j_1})\|_{\mathsf{F}}^2+2\|\widehat{U}_2^{\mathsf{T}}e_i\|_{\mathsf{F}}^2)^{1/2}\sqrt{d_1d_2/n}}$$

and we shall proceed to reject the null hypothesis if and only if $\widehat{z}>z_\theta$ to account for the one-sided alternative.

5 | INITIAL ESTIMATE

Thus far, our development has assumed a generic max-norm consistent matrix estimate as initial estimator. For concreteness, we now introduce a rotation calibrated gradient descent algorithm on Grassmannians which, under mild conditions, produces such an estimate.

Any rank r matrix of dimension $d_1 \times d_2$ can be written as UGV^{T} where $U \in \mathbb{O}^{d_1 \times r}$, $V \in \mathbb{O}^{d_2 \times r}$ and $G \in \mathbb{R}^{r \times r}$. The loss of the triplet (U, G, V) over $\mathcal{D}$ is given by

$$L(\mathcal{D}, (U, G, V)) = \sum_{(X,Y) \in \mathcal{D}} (Y - \langle UGV^{\mathsf{T}}, X \rangle)^2. \tag{23}$$

Given (U, V), we can easily minimize (23) to solve for G. This allows us to reduce the problem of minimizing (23) to a minimization over the product space of two Grassmannians $\mathsf{Gr}(d_1, r) \times \mathsf{Gr}(d_2, r)$ as $\mathsf{Gr}(d, r) = \mathbb{O}^{d_1 \times r}/\mathbb{O}^{r \times r}$. In particular, we can do so via a rotation calibrated gradient descent algorithm on Grassmannians as detailed in Algorithm 1 where, for simplicity, we resort to data-splitting. It is plausible that a more elaborative analysis via the leave-one-out (LOO) framework introduced by Ma et al. (2017) can be applied to show that our algorithm continues to produce estimates of similar quality without data-splitting, as we observe empirically. An actual proof, however, is likely much more involved under our setting. For brevity, we opted here for data-splitting.

Algorithm 1 Rotation Calibrated Gradient descent on Grassmannians

Let $\widehat{U}^{(1)}$ and $\widehat{V}^{(1)}$ be the top-r left and right singular vectors of $d_1 d_2 N_0^{-1} \sum_{j \in \mathfrak{D}_1} Y_j X_j$.

Compute $\widehat{G}^{(1)} = \arg\min_{G \in \mathbb{R}^{r \times r}} L(\mathfrak{D}_2, (\widehat{U}^{(1)}, G, \widehat{V}^{(1)}))$ and its SVD $\widehat{G}^{(1)} = \widehat{L}_G^{(1)} \widehat{\Lambda}^{(1)} \widehat{R}_G^{(1)\mathsf{T}}$.

for $t = 1, 2, 3, \cdots, m-1$ **do**

Update by rotation calibrated gradient descent

$$\widehat{U}^{(t+0.5)} = \widehat{U}^{(t)} \widehat{L}_G^{(t)} - \eta \cdot \frac{d_1 d_2}{N_0} \sum_{j \in \mathfrak{D}_{2t+1}} \big(\langle \widehat{U}^{(t)} \widehat{G}^{(t)} \widehat{V}^{(t)}, X_j \rangle - Y_j \big) X_j \widehat{V}^{(t)} \widehat{R}_G^{(t)} (\widehat{\Lambda}^{(t)})^{-1}$$

$$\widehat{V}^{(t+0.5)} = \widehat{V}^{(t)} \widehat{R}_G^{(t)} - \eta \cdot \frac{d_1 d_2}{N_0} \sum_{j \in \mathfrak{D}_{2t+1}} \big(\langle \widehat{U}^{(t)} \widehat{G}^{(t)} \widehat{V}^{(t)}, X_j \rangle - Y_j \big) X_j^{\mathsf{T}} \widehat{U}^{(t)} \widehat{L}_G^{(t)} (\widehat{\Lambda}^{(t)})^{-1}$$

Compute the top-r left singular vectors

$$\widehat{U}^{(t+1)} = \mathrm{SVD}(\widehat{U}^{(t+0.5)}) \quad \text{and} \quad \widehat{V}^{(t+1)} = \mathrm{SVD}(\widehat{V}^{(t+0.5)})$$

Compute $\widehat{G}^{(t+1)}$ by

$$\widehat{G}^{(t+1)} = \underset{G \in \mathbb{R}^{r \times r}}{\arg\min}\, L\big(\mathfrak{D}_{2t+2}, (\widehat{U}^{(t+1)}, G, \widehat{V}^{(t+1)})\big) \text{ and its SVD } \widehat{G}^{(t+1)} = \widehat{L}_G^{(t+1)} \widehat{\Lambda}^{(t+1)} \widehat{R}_G^{(t+1)\mathsf{T}}$$

end for

Output: $(\widehat{U}^{(m)}, \widehat{G}^{(m)}, \widehat{V}^{(m)})$ and $\widehat{M}^{(m)} = \widehat{U}^{(m)} \widehat{G}^{(m)} (\widehat{V}^{(m)})^{\mathsf{T}}$.

Let $m = C_1 \log(d_1 + d_2)$ for some positive integer $C_1 \geq 1$. We shall partition the data $\mathcal{D} = \{(X_i, Y_i)\}_{i=1}^n$ into $2m$ subsets:

$$\mathcal{D}_t = \{(X_j, Y_j)\}_{j=(t-1)N_0+1}^{tN_0}, \quad \forall\ t = 1, \cdots, 2m$$

where, without loss of generality, we assumed $n = 2mN_0$ for some positive integer N_0.

The algorithm presented here is similar in spirit to those developed earlier by Keshavan et al. (2010a, 2010b), Xia and Yuan (2019). A key difference is that we introduce an explicit rule of gradient descent update where each iteration on Grassmannians is calibrated with orthogonal rotations. The rotation calibrations are necessary to guarantee the contraction property for the (2, max)-norm accuracy of empirical singular vectors. Indeed, we show that the algorithm converges geometrically with constant stepsizes.

To this end, write

$$\widehat{O}_U^{(1)} = \arg O\mathbb{O}^{rr} \times \widehat{\min} \|\widehat{U}^{(1)} - UO\| \quad \text{and} \quad \widehat{O}_V^{(1)} = \arg O\mathbb{O}^{rr} \times \widehat{\min} \|\widehat{V}^{(1)} - VO\|$$

and, for all $t = 1, \cdots, m-1$, denote the SVDs

$$\widehat{U}^{(t+0.5)} = \widehat{U}^{(t+1)} \widehat{\Sigma}_U^{(t+1)} \widehat{K}_U^{(t+1)\mathsf{T}} \quad \text{and} \quad \widehat{V}^{(t+0.5)} = \widehat{V}^{(t+1)} \widehat{\Sigma}_V^{(t+1)} \widehat{K}_V^{(t+1)\mathsf{T}}.$$

For all $t = 1, \cdots, m-1$, define the orthogonal matrices

$$\widehat{O}_U^{(t+1)} = \widehat{O}_U^{(t)} \widehat{L}_G^{(t)} \widehat{K}_U^{(t+1)} \quad \text{and} \quad \widehat{O}_V^{(t+1)} = \widehat{O}_V^{(t)} \widehat{R}_G^{(t)} \widehat{K}_V^{(t+1)}.$$

Then we have

Theorem 3 *Under Assumptions 2 and 3, if* $\eta \in [0.25, 0.75]$ *and*

$$n \geq C_1 \alpha_d \kappa_0^6 \mu_{\max}^6 r^3 d_1 \log^2 d_1 \quad \text{and} \quad C_2 \kappa_0^2 \mu_{\max} \frac{\sigma_\xi}{\lambda_r} \cdot \sqrt{\frac{\alpha_d r d_1^2 d_2 \log^2 d_1}{n}} \leq 1$$

for some large enough constants $C_1, C_2 > 0$, then for all $t = 1, \cdots, m-1$, with probability at least $1 - 4md_1^{-2}$,

$$\begin{aligned} \|\widehat{U}^{(t+1)} - U\widehat{O}_U^{(t+1)}\|_{2,\max} + \|\widehat{V}^{(t+1)} - V\widehat{O}_V^{(t+1)}\|_{2,\max} &\leq C_3 \eta \frac{\sigma_\xi}{\lambda_r} \sqrt{\frac{r d_1 d_2 \log^2 d_1}{n}} \\ &+ (1 - \frac{2\eta}{3}) \cdot (\|\widehat{U}^{(t)} - U\widehat{O}_U^{(t)}\|_{2,\max} + \|\widehat{V}^{(t)} - V\widehat{O}_V^{(t)}\|_{2,\max}) \end{aligned}$$

where $C_3 > 0$ is an absolute constant. Moreover, if in addition $\|M\|_{\max}/\sigma_\xi \leq d_1^{C_4}$ for some constant $C_4 > 0$, then, by setting $m = 2C_4 \log d_1$ and $\eta = 0.75$, with probability at least $1 - C_5 d_1^{-2} \log d_1$,

$$\|\widehat{M}^{(m)} - M\|_{\max} \leq C_6 \mu_{\max} \kappa_0 \sigma_\xi \sqrt{\frac{r^2 d_1 \log^2 d_1}{n}}$$

for some absolute constants $C_5, C_6 > 0$.

We can then apply Algorithm 1 to produce initial estimates suitable for inferences about linear forms of M. With this particular choice of initial estimate, Assumption 1 is satisfied with

$$\gamma_{n,d_1,d_2} = \mu_{\max} \kappa_0 \sqrt{\frac{r^2 d_1 \log^2 d_1}{n}}$$

when the sample size $n \geq C_1 \alpha_d \kappa_0^6 \mu_{\max}^6 r^3 d_1 \log^2 d_1$. We note that this sample size requirement in general is not optimal and the extra logarithmic factor is due to data splitting. As this is not the main focus of the current work, no attempt is made here to further improve it.

6 | NUMERICAL EXPERIMENTS

We now present several sets of numerical studies to further illustrate the practical merits of the proposed methodology, and complement our theoretical developments.

6.1 | Simulations

We first consider several sets of simulation studies. Throughout the simulations, the true matrix M has rank $r=3$ and dimension $d_1=d_2=d=2000$. M's singular values were set to be $\lambda_i = d$ for $i=1,2,3$. In addition, M's singular vectors were generated from the SVD of $d \times r$ Rademacher random matrices. The noise standard deviation was set at $\sigma_\xi = 0.6$.

First, we show the convergence performance of the proposed Algorithm 1 where both Frobenius norm and max-norm convergence rates are recorded. Even though the algorithm we presented in the previous section uses sample splitting for technical convenience, in the simulation, we did not split the sample. Figure 1 shows a typical realization under Gaussian noise, which suggest the fast convergence of Algorithm 1. In particular, $\log \frac{\|\widehat{M}^{\text{init}} - M\|_{\max}}{\sigma_\xi}$ becomes negative after 3 iterations when the stepsize is $\eta = 0.6$. Recall that our double-sample debiasing approach requires $\|\widehat{M}^{\text{init}} - M\|_{\max} = o_P(\sigma_\xi)$ for the initial estimates, that is $\widehat{M}_1^{\text{init}}, \widehat{M}_2^{\text{init}}$ in Assumption 1.

Next, we investigate how the proposed inference tools behave under Gaussian noise and for four different linear forms corresponding to $T_1 = e_1 e_1^\top, T_2 = e_1 e_1^\top - e_1 e_2^\top, T_3 = e_1 e_1^\top - e_1 e_2^\top + e_2 e_1^\top$ and

$$T_4 = e_1 e_1^\top - e_1 e_2^\top + 2 e_2 e_1^\top + 3 e_2 e_2^\top.$$

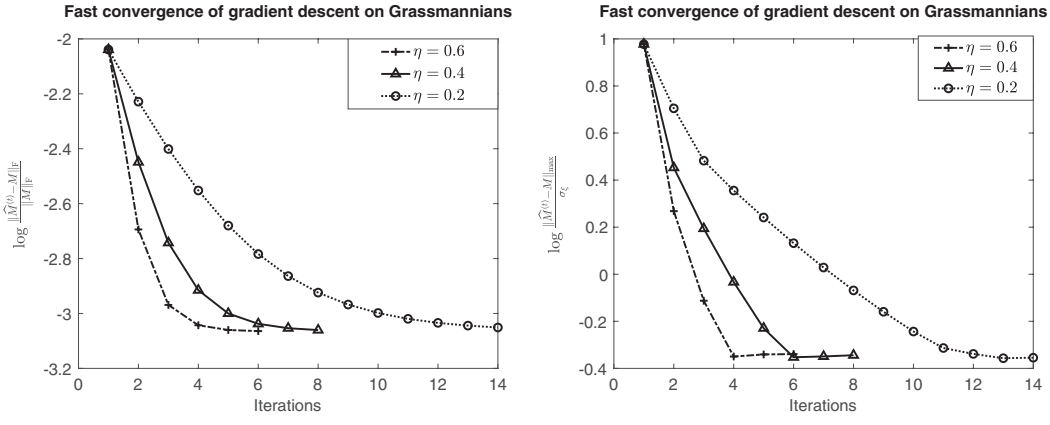

FIGURE 1 Convergence of Algorithm 1 in relative matrix Frobenius norm and the max-norm, with respect to step size η and the number of iterations. The parameters are $d_1 = d_2 = d = 2000$, $r=3$, $\lambda_i = d$, $\sigma_\xi = 0.6$ and U, V are generated from the SVD of $d \times r$ Rademacher random matrices. The sample size is $n = 4r^2 d \log(d)$ and the noise is Gaussian

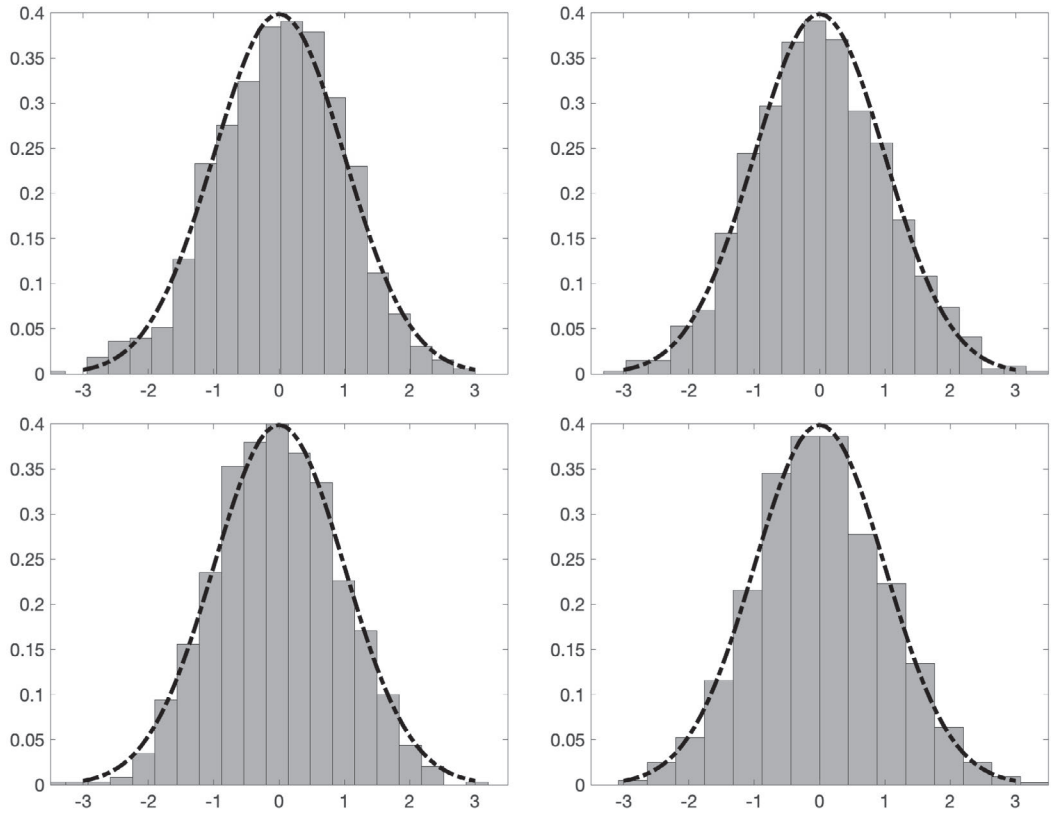

FIGURE 2 Normal approximation of $\frac{\widehat{m}_T - m_T}{\widehat{\sigma}_\xi \widehat{s}_T \sqrt{d_1 d_2/n}}$. The parameters are $d_1 = d_2 = d = 2000$, $r = 3$, $\lambda_i = d$, $\sigma_\xi = 0.6$ and U, V are generated from the SVD of $d \times r$ Rademacher random matrices. The sample size is $n = 4r^2 d\log(d)$ for the top two and $n = 5r^2 d\log(d)$ for bottom two. The noise is Gaussian. Each density histogram is based on 1000 independent simulations and the dashed black curve represents the p.d.f. of standard normal distributions. Top left: $T = e_1 e_1^\top$, top right: $T = e_1 e_1^\top - e_1 e_2^\top$. Bottom left: $T = e_1 e_1^\top - e_1 e_2^\top + e_2 e_1^\top$, bottom right: $T = e_1 e_1^\top - e_1 e_2^\top + 2e_2 e_1^\top + 3e_2 e_2^\top$

For each T, we drew the density histogram of $(\widehat{m}_T - m_T)/(\widehat{\sigma}_\xi \widehat{s}_T \sqrt{d_1 d_2/n})$ based on 1000 independent simulation runs. The density histograms are displayed in Figure 2 where the dashed black curve represents the p.d.f. of standard normal distributions. The sample size was $n = 4r^2 d\log(d)$ for T_1, T_2 and $n = 5r^2 d\log(d)$ for T_3, T_4. The empirical observation agrees fairly well with our theoretical results.

Finally, we examine the performance of the proposed approach under non-Gaussian noise. In particular, we repeated the last set of experiments with noise $(\xi/\sqrt{3}\sigma_\xi) \in \text{Unif}([-1, 1])$. The density histograms are displayed in Figure 3 where the dashed black curve represents the p.d.f. of standard normal distributions. Again the empirical evidences support the asymptotic normality of the proposed statistic.

6.2 | Real-world data examples

We now turn our attention to two real-world data examples—the Jester and MovieLens data sets. The Jester data set is downloadable from http://eigentaste.berkeley.edu/dataset/. The Jester data set contains ratings of 100 jokes from ∼70K users Goldberg et al. (2001). The data set consists of three

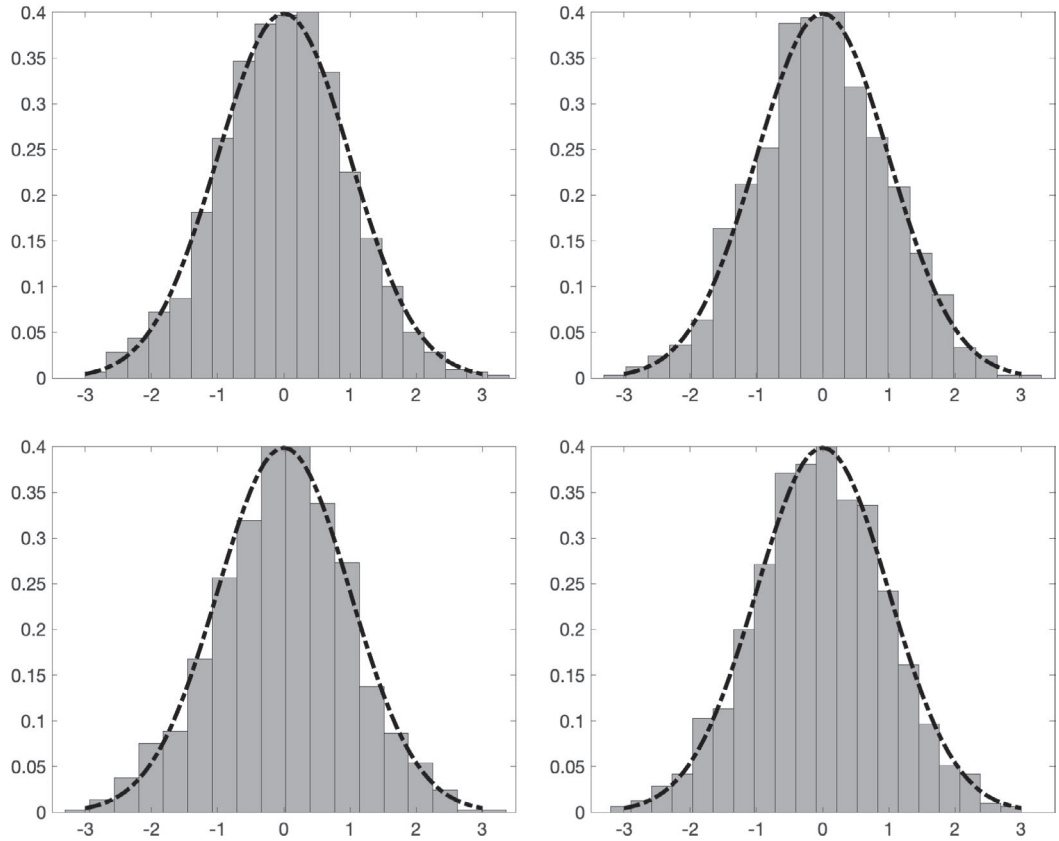

FIGURE 3 Normal approximation of $\frac{\widehat{m}_T - m_T}{\widehat{\sigma}_\xi \widehat{s}_T \sqrt{d_1 d_2 / n}}$. The parameters are $d_1 = d_2 = d = 2000$, $r = 3$, $\lambda_i = d$, $\sigma_\xi = 0.6$ and U, V are generated from the SVD of $d \times r$ Rademacher random matrices. The sample size is $n = 4r^2 d\log(d)$ for the top two and $n = 5r^2 d\log(d)$ for the bottom two. The non-Gaussian noise $(\xi/\sqrt{3}\sigma_\xi) \in \text{Unif}([-1, 1])$. Each density histogram is based on 1000 independent simulations and the dashed black curve represents the p.d.f. of standard normal distributions. Top left: $T = e_1 e_1^\top$, top right: $T = e_1 e_1^\top - e_1 e_2^\top$. Bottom left: $T = e_1 e_1^\top - e_1 e_2^\top + e_2 e_1^\top$, bottom right: $T = e_1 e_1^\top - e_1 e_2^\top + 2e_2 e_1^\top + 3e_2 e_2^\top$

subsets of data with different characteristics as summarized in Table 1. Note that for each subset, we randomly sample a fixed number of ratings for each user so that the numbers of ratings of all users are equal (see Table 1). MovieLens was a recommender system created by GroupLens that recommends movies for its users. We use three data sets released by MovieLens (Harper & Konstan, 2016) whose details are summarized in Table 1. These data sets are downloadable from https://grouplens.org/datasets/movielens/. In these three data sets, each user rates at least 20 movies. The Matlab codes used for preprocessing the raw data sets are included in the supplementary files on the journal's websites.

For illustration, we consider the task of recommending jokes or movies to a particular users. Because of the lack of ground truth, we resort to resampling. For the Jester data set, we randomly sample ∼2000 users, and for each user 2 ratings that at least $\zeta \in \{0, 2, 6, 10, 14\}$ apart. We removed these ratings from the training and used the proposed procedure to infer, for each user (i), between these two jokes (j_1 or j_2) with ratings which one should be recommended. This amounts to the following one-sided tests:

$$H_0: M(i, j_1) \le M(i, j_2) \quad \text{v. s.} \quad H_1: M(i, j_1) > M(i, j_2).$$

TABLE 1 Summary of data sets

Data set	#users	#jokes	#ratings per user	rating values
Jester-1	24983	100	29	[–10, 10]
Jester-2	23500	100	34	[–10, 10]
Jester-3	24938	100	14	[–10, 10]
Data set	#users	#movies	total #ratings	rating values
ml-100k	943	1682	$\sim 10^5$	{1, 2, 3, 4, 5}
ml-1m	6040	3952	$\sim 10^6$	{1, 2, 3, 4, 5}
ml-10m	71567	10681	$\sim 10^7$	{0.5, 1.0, ..., 4.5, 5.0}

We ran the proposed procedure on the training data and evaluate the test statistic $\widehat{z}$ for each user from the testing set. In particular, we fixed the rank $r=2$ corresponding to the smallest estimate $\widehat{\sigma}_{\xi}$. Note that we do not know the true value of $M(i,j)$ and only observe $Y(i,j)=M(i,j)+\xi(i,j)$. We therefore use $\mathbb{I}(Y(i,j_1)>Y(i,j_2))$ as a proxy to differentiate between H_0 and H_1. Assuming that the ξ has a distribution symmetric about 0, then $\mathbb{I}(Y(i,j_1)>Y(i,j_2))$ is more likely to take value 0 under H_0, and 1 under H_1. We shall evaluate the performance of our procedure based on its discriminant power in predicting $\mathbb{I}(Y(i,j_1)>Y(i,j_2))$. In particular, we record the ROC curve of $\widehat{z}$ for all users from the testing set. The results, averaged over 10 simulation runs for each value of ζ, are reported in Figure 4. Clearly, we can observe an increase in predictive power as ζ increases suggesting $\widehat{z}$ as a reasonable statistic for testing H_0 against H_1.

We ran a similar experiment on the MovieLens data sets. In each simulation run, we randomly sampled ~ 800 users and 2 ratings each as the test data. These ratings are sampled such that $|Y(i,j_1)-Y(i,j_2)|\geq\zeta$ for $\zeta=0,1,2,3,4$. The false-positive rates and true-positive rates of our proposed procedure were again recorded. The ROC curves, averaged again over 10 runs for each value of ζ, are shown in Figure 5. This again indicates a reasonable performance of the proposed testing procedure. Empirically, we observe a better debiasing approach on these data sets which is $\widehat{M}_1^{\text{unbs}}=\widehat{M}_1^{\text{init}}+\sum_{i\in\mathcal{D}_2}(Y_i-\langle\widehat{M}_1^{\text{init}},X_i\rangle)X_i$. The rationale is to partially replace $\widehat{M}_1^{\text{init}}$'s entries with the observed training ratings. This improvement might be due to the severe heterogeneity in the numbers of observed ratings from distinct users, or due to the unknown noise distributions.

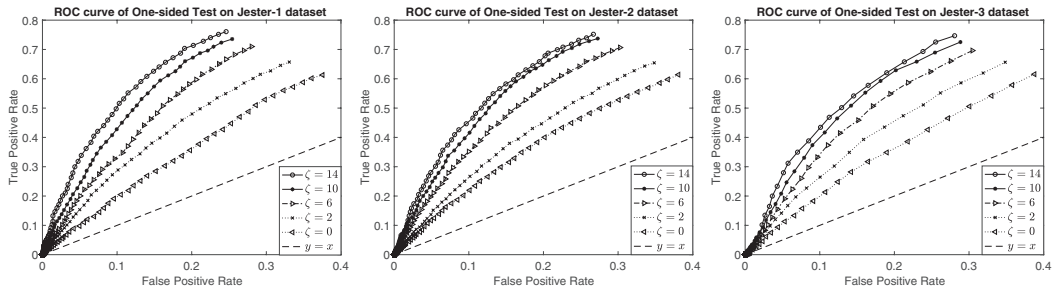

FIGURE 4 ROC curves for one-sided tests $H_0: M(i,j_1)\leq M(i,j_2)$ v. s. $H_1: M(i,j_1)>M(i,j_2)$ on Jester data sets. The testing data are sampled such that $|Y(i,j_1)-Y(i,j_2)|\geq\zeta$. The estimated noise level $\widehat{\sigma}_{\xi}=4.5160$ on Jester-1, $\widehat{\sigma}_{\xi}=4.4843$ on Jester-2 and $\widehat{\sigma}_{\xi}=5.1152$ on Jester-3. The rightmost point of each ROC curve corresponds to the significance level $\theta=0.5$ so that $z_\theta=0$

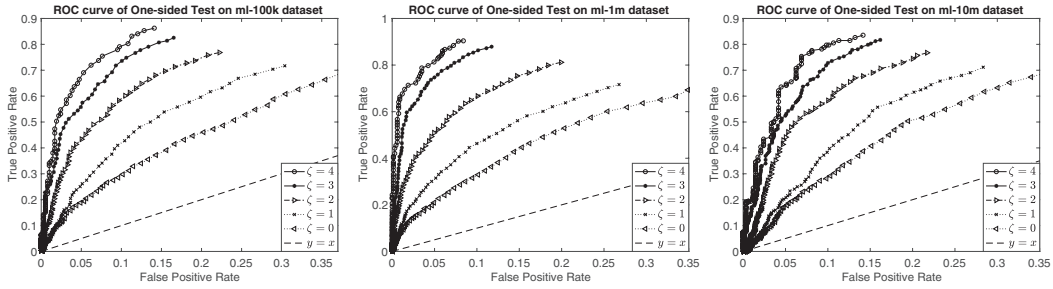

FIGURE 5 ROC curves for one-sided tests $H_0: M(i,j_1) \le M(i,j_2)$ v. s. $H_1: M(i,j_1) > M(i,j_2)$ on MovieLens data sets. The testing data are sampled such that $|Y(i,j_1) - Y(i,j_2)| \ge \zeta$. The estimated noise level $\widehat{\sigma}_\xi = 0.9973$ on ml-100k, $\widehat{\sigma}_\xi = 0.8936$ on ml-1m and $\widehat{\sigma}_\xi = 0.9151$ on ml-10m. The rightmost point of each ROC curve corresponds to the significance level $\theta = 0.5$ so that $z_\theta = 0$

ACKNOWLEDGEMENTS

Dong Xia's research is partially supported by Hong Kong RGC Grant ECS 26302019. Ming Yuan's research is supported in part by NSF Grant DMS-1803450.

REFERENCES

Berry, A. C. (1941) The accuracy of the gaussian approximation to the sum of independent variates. *Transactions of the American Mathematical Society*, 49(1), 122–136.

Cai, J.-F., Candès, E.J. & Shen, Z. (2010) A singular value thresholding algorithm for matrix completion. *SIAM Journal on Optimization*, 20(4), 1956–1982.

Cai, T.T. & Guo, Z. (2017) Confidence intervals for high-dimensional linear regression: Minimax rates and adaptivity. *The Annals of Statistics*, 45(2), 615–646.

Cai, T.T. & Zhang, A. (2015) Rop: Matrix recovery via rank-one projections. *The Annals of Statistics*, 43(1), 102–138.

Cai, T.T. & Zhou, W.-X. (2016) Matrix completion via max-norm constrained optimization. *Electronic Journal of Statistics*, 10(1), 1493–1525.

Cai, T.T., Liang, T. & Rakhlin, A. (2016a) Geometric inference for general high-dimensional linear inverse problems. *The Annals of Statistics*, 44(4), 1536–1563.

Cai, T., Cai, T.T. & Zhang, A. (2016b) Structured matrix completion with applications to genomic data integration. *Journal of the American Statistical Association*, 111(514), 621–633.

Candes, E.J. & Plan, Y. (2010) Matrix completion with noise. *Proceedings of the IEEE*, 98(6), 925–936.

Candès, E.J. & Recht, B. (2009) Exact matrix completion via convex optimization. *Foundations of Computational Mathematics*, 9(6), 717.

Candès, E.J. & Tao, T. (2009) The power of convex relaxation: Near-optimal matrix completion. arXiv preprint arXiv:0903.1476.

Carpentier, A. & Kim, A.K.H. (2018) An iterative hard thresholding estimator for low rank matrix recovery with explicit limiting distribution. *Statistica Sinica*, 28, 1371–1393.

Carpentier, A., Eisert, J., Gross, D. & Nickl, R. (2015) Uncertainty quantification for matrix compressed sensing and quantum tomography problems. arXiv preprint arXiv:1504.03234.

Carpentier, A., Klopp, O., Löffler, M. & Nickl, R. (2018) Adaptive confidence sets for matrix completion. *Bernoulli*, 24(4A), 2429–2460.

Chen, J., Liu, D. & Li, X. (2019a) Nonconvex rectangular matrix completion via gradient descent without $\ell_{2,\infty}$ regularization. arXiv preprint arXiv:1901.06116.

Chen, Y. & Wainwright, M.J. (2015) Fast low-rank estimation by projected gradient descent: General statistical and algorithmic guarantees. arXiv preprint arXiv:1509.03025.

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:1902.07698.

Chen, Y., Fan, J., Ma, C. & Yan, Y. (2019c) Inference and uncertainty quantification for noisy matrix completion. arXiv preprint arXiv:1906.04159.

Chernozhukov, V., Chetverikov, D., Demirer, M., Duflo, E., Hansen, C., Newey, W. & Robins, J. (2018) Double/debiased machine learning for treatment and structural parameters. *The Econometrics Journal*, 21(1), C1–C68.

Davis, C. & Kahan, W.M. (1970) The rotation of eigenvectors by a perturbation iii. *SIAM Journal on Numerical Analysis*, 7(1), 1–46.

Edelman, A., Arias, T.A. & Smith, S.T. (1998) The geometry of algorithms with orthogonality constraints. *SIAM Journal on Matrix Analysis and Applications*, 20(2), 303–353.

Esseen, Carl-Gustav (1956) A moment inequality with an application to the central limit theorem. *Scandinavian Actuarial Journal*, 1956(2), 160–170.

Gao, C., Lu, Y., Ma, Z. & Zhou, H.H. (2016) Optimal estimation and completion of matrices with biclustering structures. *The Journal of Machine Learning Research*, 17(1), 5602–5630.

Ge, R., Lee, J.D. & Ma, T. (2016) Matrix completion has no spurious local minimum. *Advances in Neural Information Processing Systems*, 2973–2981.

Goldberg, K., Roeder, T., Gupta, D. & Perkins, C. (2001) Eigentaste: A constant time collaborative filtering algorithm. *Information Retrieval*, 4(2), 133–151.

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

Harper, F.M. & Konstan, J.A. (2016) The movielens datasets: History and context. *ACM Transactions on Interactive Intelligent Systems (TIIS)*, 5(4), 19.

Javanmard, A. & Montanari, A. (2014) Confidence intervals and hypothesis testing for high-dimensional regression. *The Journal of Machine Learning Research*, 15(1), 2869–2909.

Keshavan, R.H., Montanari, A. & Oh, S. (2010a) Matrix completion from a few entries. *IEEE Transactions on Information Theory*, 56(6), 2980–2998.

Keshavan, R.H., Montanari, A. & Oh, S. (2010b) Matrix completion from noisy entries. *Journal of Machine Learning Research*, 11(7), 2057–2078.

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

Koltchinskii, V. (2011) Von neumann entropy penalization and low-rank matrix estimation. *The Annals of Statistics*, 39(6), 2936–2973.

Koltchinskii, V. & Xia, D. (2015) Optimal estimation of low rank density matrices. *Journal of Machine Learning Research*, 16(53), 1757–1792.

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

Liu, Y.-K. (2011) Universal low-rank matrix recovery from pauli measurements. *Advances in Neural Information Processing Systems*, 1638–1646.

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:1711.10467.

Ma, Z. & Wu, Y. (2015) Volume ratio, sparsity, and minimaxity under unitarily invariant norms. *IEEE Transactions on Information Theory*, 61(12), 6939–6956.

Minsker, S. (2017) On some extensions of bernstein's inequality for self-adjoint operators. *Statistics & Probability Letters*, 127, 111–119.

Negahban, S. & Wainwright, M.J. (2011) Estimation of (near) low-rank matrices with noise and high-dimensional scaling. *The Annals of Statistics*, 39(2), 1069–1097.

Pajor, A. (1998) Metric entropy of the grassmann manifold. *Convex Geometric Analysis*, 34, 181–188.

Recht, B., Fazel, M. & Parrilo, P.A. (2010) Guaranteed minimum-rank solutions of linear matrix equations via nuclear norm minimization. *SIAM Review*, 52(3), 471–501.

Rohde, A. & Tsybakov, A.B. (2011) Estimation of high-dimensional low-rank matrices. *The Annals of Statistics*, 39(2), 887–930.

Sun, T. & Zhang, C.-H. (2012) Calibrated elastic regularization in matrix completion. *Advances in Neural Information Processing Systems*, 863–871.

Tropp, J.A. (2012) User-friendly tail bounds for sums of random matrices. *Foundations of Computational Mathematics*, 12(4), 389–434.

Van de Geer, S., Bühlmann, P., Ritov, Y. & Dezeure, R. (2014) On asymptotically optimal confidence regions and tests for high-dimensional models. *The Annals of Statistics*, 42(3), 1166–1202.

Wang, L., Zhang, X. & Gu, Q. (2016) A unified computational and statistical framework for nonconvex low-rank matrix estimation. arXiv preprint arXiv:1610.05275.

Wedin, P. (1972) Perturbation bounds in connection with singular value decomposition. *BIT Numerical Mathematics*, 12(1), 99–111.

Xia, D. (2019a) Confidence region of singular subspaces for high-dimensional and low-rank matrix regression. *IEEE Transactions on Information Theory*, 65(11), 1–23.

Xia, D. (2019b) Normal approximation and confidence region of singular subspaces. arXiv preprint arXiv:1901.00304.

Xia, D. & Yuan, M. (2019) On polynomial time methods for exact low-rank tensor completion. *Foundations of Computational Mathematics*, 19, 1265–1313.

Zhang, C.-H. & Zhang, S.S. (2014) Confidence intervals for low dimensional parameters in high dimensional linear models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 76(1), 217–242.

Zhao, T., Wang, Z. & Liu, H. (2015) A nonconvex optimization framework for low rank matrix estimation. *Advances in Neural Information Processing Systems*, 559–567.

Zheng, Q. & Lafferty, J. (2016) Convergence analysis for rectangular matrix completion using burer-monteiro factorization and gradient descent. arXiv preprint arXiv:1605.07051.

SUPPORTING INFORMATION

Additional supporting information may be found online in the Supporting Information section.

How to cite this article: Xia D, Yuan M. Statistical inferences of linear forms for noisy matrix completion. *J R Stat Soc Series B*. 2021;83:58–77. https://doi.org/10.1111/rssb.12400