Uncertainty quantification for nonconvex tensor completion: Confidence intervals, heteroscedasticity and optimality

Changxiao Cai [1] **H. Vincent Poor** [1] **Yuxin Chen** [1]

Abstract

We study the distribution and uncertainty of nonconvex optimization for noisy tensor completion — the problem of estimating a low-rank tensor given incomplete and corrupted observations of its entries. Focusing on a two-stage nonconvex estimation algorithm proposed by (Cai et al., 2019), we characterize the distribution of this estimator down to fine scales. This distributional theory in turn allows one to construct valid and short confidence intervals for both the unseen tensor entries and its underlying tensor factors. The proposed inferential procedure enjoys several important features: (1) it is fully adaptive to noise heteroscedasticity, and (2) it is data-driven and adapts automatically to unknown noise distributions. Furthermore, our findings unveil the statistical optimality of nonconvex tensor completion: it attains un-improvable estimation accuracy — including both the rates and the pre-constants — under i.i.d. Gaussian noise.

1. Introduction

1.1. Noisy low-rank tensor completion

Tensor data are routinely employed in data and information sciences to model (structured) multi-dimensional objects (Kolda & Bader, 2009; Anandkumar et al., 2014; Sidiropoulos et al., 2017; Zhang, 2019). In many practical scenarios of interest, however, we do not have full access to a large-dimensional tensor of interest, as only a sampling of its entries are revealed to us; yet we would still wish to reliably infer all missing data. This task, commonly referred to as *tensor completion*, finds applications in numerous domains including medical imaging (Semerci et al., 2014), visual data analysis (Liu et al., 2013), seismic data reconstruction (Kreimer et al., 2013), to name just a few. In order to make meaningful inference about the unseen entries, additional information about the unknown tensor plays a pivotal role (otherwise one is in the position with fewer equations than unknowns). A common type of such prior information is low-rank structure, which hypothesizes that the unknown tensor is decomposable into the superposition of a small number of rank-one tensors. Substantial attempts have been made in the past few years to tackle this low-rank tensor completion problem.

To set the stage for a formal discussion, suppose we are interested in reconstructing a third-order tensor $\boldsymbol{T}^\star = [T_{i,j,k}]_{1\le i,j,k\le d} \in \mathbb{R}^{d\times d\times d}$, which is *a priori* known to have low canonical polyadic (CP) rank. This means that $\boldsymbol{T}^\star$ admits the following CP decomposition[1]

$$\boldsymbol{T}^\star = \sum\nolimits_{l=1}^{r} \boldsymbol{u}_l^\star \otimes \boldsymbol{u}_l^\star \otimes \boldsymbol{u}_l^\star =: \sum\nolimits_{l=1}^{r} (\boldsymbol{u}_l^\star)^{\otimes 3}, \qquad (1)$$

where $\boldsymbol{u}_l^\star \in \mathbb{R}^d$ $(1 \le l \le r)$ represents the unknown tensor factor, and the rank r is considerably smaller than the ambient dimension d. What we have obtained is a highly incomplete collection of noisy observations about the entries of $\boldsymbol{T}^\star \in \mathbb{R}^{d\times d\times d}$; more precisely, suppose we observe

$$T_{i,j,k}^{\mathsf{obs}} = T_{i,j,k}^\star + E_{i,j,k}, \qquad (i,j,k) \in \Omega, \qquad (2)$$

where $\Omega \subseteq [d]^3$ with $[d] := \{1, \cdots, d\}$ is a subset of entries, $T_{i,j,k}^{\mathsf{obs}}$ denotes the observed entry in the (i,j,k)-th position, and we use $E_{i,j,k}$ to represent the noise, in an attempt to model more realistic scenarios. The presence of missing data and noise, as well as the "notorious" tensor structure (which is not as computationally friendly as its matrix analog (Hillar & Lim, 2013)), poses severe computational and statistical challenges for reliable tensor reconstruction.

1.2. Review: a nonconvex optimization approach

A natural reconstruction strategy based on the partial data at hand is to resort to the following least-squares problem:

$$\underset{\boldsymbol{U}\in\mathbb{R}^{d\times r}}{\text{minimize}}\ f(\boldsymbol{U}) = \sum_{(i,j,k)\in\Omega} \left[\Big(\sum_{l=1}^{r} \boldsymbol{u}_l^{\otimes 3} \Big)_{i,j,k} - T_{i,j,k}^{\mathsf{obs}} \right]^2. \qquad (3)$$

[1]Electrical Engineering, Princeton University. Correspondence to: Yuxin Chen <yuxin.chen@princeton.edu>.

Proceedings of the 37^{th} International Conference on Machine Learning, Online, PMLR 119, 2020.

[1]For any $\boldsymbol{u}, \boldsymbol{v}, \boldsymbol{w} \in \mathbb{R}^d$, we denote by $\boldsymbol{u}\otimes\boldsymbol{v}\otimes\boldsymbol{w} \in \mathbb{R}^{d\times d\times d}$ such that $(\boldsymbol{u}\otimes\boldsymbol{v}\otimes\boldsymbol{w})_{i,j,k} = u_i v_j w_k$ for all $1 \le i,j,k \le d$.

Here and in the sequel, we define $\boldsymbol{U} := [\boldsymbol{u}_1, \cdots, \boldsymbol{u}_r]$. Unfortunately, owing to its highly nonconvex nature, the optimization problem (3) is, in general, daunting to solve.

To alleviate computational intractability, a number of polynomial-time algorithms have been proposed; partial examples include convex relaxation (Gandy et al., 2011; Romera-Paredes & Pontil, 2013; Huang et al., 2015), spectral methods (Montanari & Sun, 2018; Cai et al., 2020a), sum of squares hierarchy (Barak & Moitra, 2016; Potechin & Steurer, 2017), alternating minimization (Jain & Oh, 2014; Liu & Moitra, 2020). Nevertheless, most of these algorithms either are still computationally prohibitive for large-scale problems, or do not come with optimal statistical guarantees; see Section 4. To address the computational and statistical challenges at once, the recent work (Cai et al., 2019) proposed a two-stage nonconvex paradigm that guarantees efficient yet reliable solutions. In a nutshell, this algorithm starts by computing a rough (but reasonable) initial guess $\boldsymbol{U}^0 = [\boldsymbol{u}_1^0, \cdots, \boldsymbol{u}_r^0]$ for all tensor factors, and iteratively refines the estimate by means of the gradient descent (GD) update rule:

$$\boldsymbol{U}^{t+1} = \boldsymbol{U}^t - \eta_t \nabla f(\boldsymbol{U}^t), \qquad t = 0, 1, \cdots \tag{4}$$

See Algorithm 1 (the initialization scheme is more complex to describe, and is hence postponed to Appendix A.1). Despite the nonconvex optimization landscape, theoretical guarantees have been developed for Algorithm 1 under random sampling and random noise. Take the noiseless case for instance: this approach converges *linearly* to the ground truth under near-minimal sample complexity; furthermore, the algorithm enjoys intriguing ℓ_2 and ℓ_∞ statistical guarantees under a broad family of noise models.

Algorithm 1 A nonconvex algorithm for tensor completion.

1: **Initialize** $\boldsymbol{U}^0 = [\boldsymbol{u}_1^0, \cdots, \boldsymbol{u}_r^0]$ via Algorithm 2.
2: **Gradient updates**: **for** $t = 0, 1, \ldots, t_0 - 1$ **do**

$$\boldsymbol{U}^{t+1} = \boldsymbol{U}^t - \eta_t \nabla f(\boldsymbol{U}^t). \tag{5}$$

3: **Output** $\boldsymbol{U} = [\boldsymbol{u}_1, \cdots, \boldsymbol{u}_r] := \boldsymbol{U}^{t_0}$.

1.3. Uncertainty quantification for tensor completion

In various decision making scenarios (e.g. medical imaging), it is crucial not only to provide the users with the reconstruction outcome, but also to inform them of the uncertainty or risk underlying the reconstruction. The latter task, often termed *uncertainty quantification,* can be accomplished by characterizing the distribution of our reconstruction, which can be further employed to construct valid confidence intervals for the unknowns. Two questions deserve particular attention: given an estimate returned by the above nonconvex algorithm, how to identify a confidence interval when predicting an unseen entry, and how to produce a confidence region that is likely to contain the tensor factors of interest?

Unfortunately, classical distributional theory available in the statistics literature, which typically operates in the large-sample regime (with a fixed number of unknowns and a sample size tending to infinity), is not applicable to assess the uncertainty of the above nonconvex algorithm in high dimension. In fact, due to the nonconvex nature of the algorithm, it becomes remarkably challenging to track the distribution of the solution returned by Algorithm 1. The absence of distributional characterization prevents us from offering a trustworthy uncertainty estimate to the users. While the statistical performance of Algorithm 1 has been investigated in (Cai et al., 2019), existing statistical guarantees — which hide the (potentially huge) pre-constants — can only yield confidence intervals that are overly wide and, as a result, practically uninformative. In principle, we should aim for valid confidence intervals that are as short as possible.

Further, an ideal uncertainty quantification procedure should be adaptive to unknown noise distributions. Accomplishing this goal becomes particularly challenging when the noise variance is not only unknown but also location-varying — a scenario commonly referred to as *heteroscedasticity*. In fact, there is no shortage of realistic scenarios in which the data heteroscedasticity makes it infeasible to pre-estimate local variability in a uniformly reliable manner. Addressing this challenge calls for the design of model-agnostic data-driven procedures that are fully adaptive to noise heteroscedasticity.

1.4. Main contributions and insights

We now give an informal overview of the main contributions and insights of this paper. To the best of our knowledge, results of this kind were previously unavailable.

A distributional theory for nonconvex tensor completion. Despite its nonconvex nature, a distributional representation of the estimate returned by Algorithm 1 can be established down to quite fine scales (i.e. down to the entrywise level). Under mild conditions, (1) the resulting estimates for both the unknown tensor factors and tensor entries are nearly unbiased, and (2) the associated uncertainty of the estimates is nearly zero-mean Gaussian, whose (co)-variance can be accurately determined in a data-driven manner.

Entrywise confidence intervals. Our distributional theory leads to construction of *entrywise* confidence intervals for both the unknown tensor and the associated tensor factors. Our inferential procedure is fully data-driven: it does not require prior knowledge about the noise distributions, and it automatically adapts to local variability of noise.

Optimality. We develop lower bounds under i.i.d. Gaussian noise, showing that the proposed entrywise confidence intervals are the shortest possible. Our results also reveal that nonconvex optimization achieves un-improvable ℓ_2 estimation accuracy (including both the rates and pre-constants).

All in all, our results shed light on the *unreasonable effectiveness of nonconvex optimization* in noisy low-rank tensor completion, which enables optimal estimation and uncertainty quantification all at once.

1.5. Notation

For any matrix $\boldsymbol{M}$, let $\|\boldsymbol{M}\|$ (resp. $\|\boldsymbol{M}\|_{\mathrm{F}}$) denote the spectral (resp. Frobenius) norm of $\boldsymbol{M}$, denote by $\|\boldsymbol{M}\|_{2,\infty} := \max_l \|\boldsymbol{M}_{l,:}\|_2$ (resp. $\|\boldsymbol{M}\|_{\infty} := \max_{i,j} |M_{i,j}|$) the $\ell_{2,\infty}$ norm (resp. entrywise ℓ_∞ norm) of $\boldsymbol{M}$, and let $\boldsymbol{M}_{i,:}$ (resp. $\boldsymbol{M}_{:,i}$) be the i-th row (resp. column). For any tensors $\boldsymbol{T} \in \mathbb{R}^{d\times d\times d}$, the Frobenius (resp. entrywise ℓ_∞) norm of $\boldsymbol{T}$ is $\|\boldsymbol{T}\|_{\mathrm{F}} := \sqrt{\sum_{i,j,k} T_{i,j,k}^2}$ (resp. $\|\boldsymbol{T}\|_{\infty} := \max_{i,j,k} |T_{i,j,k}|$). We denote by $[a \pm b]$ the interval $[a-b, a+b]$. We use $u_{l,i}$ (resp. $u^\star_{l,i}$) to denote the i-th entry of $\boldsymbol{u}_l \in \mathbb{R}^d$ (resp. $\boldsymbol{u}^\star_l \in \mathbb{R}^d$). We shall often let (i,j) denote $(i-1)d+j$ whenever it is clear from the context.

2. Models and assumptions

In this paper, we shall consider a setting with random sampling and independent random noise as follows.

Assumption 1 (Random sampling)**.** *Suppose that Ω is a symmetric index set.*[2] *Assume that each (i,j,k) with $i \le j \le k$ is included in Ω independently with probability p.*

Assumption 2 (Random noise)**.** *Suppose that $\boldsymbol{E} = [E_{i,j,k}]_{1\le i,j,k\le d}$ is a symmetric tensor.*[3] *Assume that the $E_{i,j,k}$'s are independent sub-Gaussian random variables satisfying $\mathbb{E}[E_{i,j,k}] = 0$ and $\mathsf{Var}(E_{i,j,k}) = \sigma^2_{i,j,k}$. Denoting $\sigma_{\min} := \min_{i,j,k} \sigma_{i,j,k}$ and $\sigma_{\max} := \max_{i,j,k} \sigma_{i,j,k}$, we assume throughout that $\sigma_{\max}/\sigma_{\min} = O(1)$.*

Next, we introduce additional parameters about the unknown tensor of interest. Recall that

$$\boldsymbol{T}^\star = \sum\nolimits_{l=1}^{r} \boldsymbol{u}_l^\star \otimes \boldsymbol{u}_l^\star \otimes \boldsymbol{u}_l^\star = \sum\nolimits_{l=1}^{r} \boldsymbol{u}_l^{\star\otimes 3} \in \mathbb{R}^{d\times d\times d}.$$

To begin with, we define

$$\lambda^\star_{\min} := \min_{1\le l\le r} \|\boldsymbol{u}_l^\star\|_2^3 \ \text{ and } \ \lambda^\star_{\max} := \max_{1\le l\le r} \|\boldsymbol{u}_l^\star\|_2^3, \quad (6)$$

allowing us to define the condition number by

$$\kappa := \lambda^\star_{\max}/\lambda^\star_{\min}. \quad (7)$$

To enable reliable tensor completion, we impose further assumptions on the tensor factors $\{\boldsymbol{u}_l^\star\}$ as follows.

Assumption 3 (Incoherence and well-conditionedness)**.** *Suppose that $\boldsymbol{T}^\star$ satisfies*

$$\|\boldsymbol{T}^\star\|_\infty \le \sqrt{\frac{\mu_0}{d^3}}\, \|\boldsymbol{T}^\star\|_{\mathrm{F}}; \quad (8a)$$

$$\|\boldsymbol{u}_l^\star\|_\infty \le \sqrt{\frac{\mu_1}{d}}\, \|\boldsymbol{u}_l^\star\|_2, \quad 1\le l\le r; \quad (8b)$$

$$\left|\langle \boldsymbol{u}_l^\star, \boldsymbol{u}_j^\star\rangle\right| \le \sqrt{\frac{\mu_2}{d}}\, \|\boldsymbol{u}_l^\star\|_2 \left\|\boldsymbol{u}_j^\star\right\|_2, \quad l\ne j. \quad (8c)$$

Further, assume that $\boldsymbol{T}^\star$ is well-conditioned in the sense that κ (cf. (7)) obeys $\kappa = O(1)$.

Informally, when both μ_0 and μ_1 are small, the ℓ_2 energy of both $\boldsymbol{T}^\star$ and $\boldsymbol{u}_l^\star$ $(1\le l\le r)$ is dispersed more or less evenly across their entries. In addition, a small μ_2 implies that every pair of the tensor factors of interest is nearly orthogonal to (and hence incoherent with) each other. Finally, the well-conditionedness assumption guarantees that no single tensor component has significantly higher energy compared to the rest of them. For the sake of notational simplicity, we shall combine them into a single incoherence parameter

$$\mu := \max\{\mu_0, \mu_1, \mu_2\}. \quad (9)$$

The focal point of this paper lies in distributional characterization of and uncertainty assessment for the nonconvex estimate (i.e. the solution $\boldsymbol{U}$ returned by Algorithm 1) in a strong entrywise sense. In particular, we set out the goal to (1) establish distributional representation of the estimate $\boldsymbol{U}$, and (2) construct short yet valid confidence intervals for each entry of the tensor factor $\{\boldsymbol{u}_l^\star\}_{1\le l\le r}$ as well as each entry of the unknown tensor $\boldsymbol{T}^\star$. To phrase the latter task more precisely: given any target coverage level $0 < 1-\alpha < 1$, any $1\le l\le r$ and any $1\le i,j,k\le d$, the aim is to compute intervals $[c_{1,T}, c_{2,T}]$ and $[c_{1,u}, c_{2,u}]$ such that

$$\mathbb{P}\left\{T^\star_{i,j,k} \in [c_{1,T}, c_{2,T}]\right\} = 1-\alpha+o(1)$$

up to global permutation and

$$\mathbb{P}\left\{u^\star_{l,i} \in [c_{1,u}, c_{2,u}]\right\} = 1-\alpha+o(1)$$

Ideally, this should be accomplished in a data-driven manner without requiring prior knowledge about noise distributions.

3. Main results

This section presents our distributional theory for nonconvex noisy tensor completion, and demonstrates how to conduct uncertainty quantification in a data-driven and optimal manner. In the sequel, we denote by $\boldsymbol{U} = [\boldsymbol{u}_1, \cdots, \boldsymbol{u}_r] \in \mathbb{R}^{d\times r}$ the estimate returned by Algorithm 1, and let $\boldsymbol{T} \in \mathbb{R}^{d\times d\times d}$ indicate the resulting tensor estimate as follows

$$\boldsymbol{T} := \sum\nolimits_{l=1}^{r} \boldsymbol{u}_l \otimes \boldsymbol{u}_l \otimes \boldsymbol{u}_l. \quad (10)$$

Given that one can only hope to recover $\boldsymbol{U}^\star$ up to global permutation[4], we introduce a permutation matrix as follows

$$\boldsymbol{\Pi} := \min_{\boldsymbol{Q}\in \mathsf{perm}_r} \|\boldsymbol{U}\boldsymbol{Q} - \boldsymbol{U}^\star\|_{\mathrm{F}}, \quad (11)$$

[2] This means that if $(i,j,k)\in\Omega$, then (j,i,k), (i,k,j), (j,k,i), (k,i,j), (k,j,i) are all in Ω.

[3] This means that $E_{i,j,k} = E_{j,i,k} = E_{i,k,j} = E_{j,k,i} = E_{k,i,j} = E_{k,j,i}$ for any $1\le i,j,k\le d$.

[4] More precisely, we cannot distinguish $\boldsymbol{u}_1^\star, \cdots, \boldsymbol{u}_r^\star$ from an arbitrary permutation of them based on the observed data (2).

where perm_r is the set of permutation matrices in $\mathbb{R}^{r\times r}$. In order to guarantee reliable convergence of Algorithm 1, there are several algorithmic parameters (e.g. learning rates) that need to be properly chosen. We shall adopt the choices suggested by (Cai et al., 2019) throughout. Given that our theory can be presented regardless of the reader's understanding of these choices, we defer the specification of these algorithmic parameters to Appendix A.2 to avoid distraction. All proofs are deferred to (Cai et al., 2020b).

3.1. Distributional guarantees for nonconvex estimates

We now establish distributional guarantees for the nonconvex estimate. For notational convenience, we introduce an auxiliary matrix $\widetilde{\boldsymbol{U}}^\star \in \mathbb{R}^{d^2\times r}$ as well as a collection of diagonal matrices $\boldsymbol{D}_k^\star \in \mathbb{R}^{d^2\times d^2}$ $(1\le k\le d)$ such that

$$\widetilde{\boldsymbol{U}}^\star := \big[\boldsymbol{u}_1^\star\otimes\boldsymbol{u}_1^\star,\cdots,\boldsymbol{u}_r^\star\otimes\boldsymbol{u}_r^\star\big]\in\mathbb{R}^{d^2\times r}; \tag{12}$$

$$\big(\boldsymbol{D}_k^\star\big)_{(i,j),(i,j)} := \sigma_{i,j,k}^2, \qquad 1\le i,j\le d; \tag{13}$$

here, we abuse the notation (i,j) to denote $(i-1)d+j$. In words, $\widetilde{\boldsymbol{U}}^\star$ lifts the tensor factors to a higher order, and $\boldsymbol{D}_k^\star$ records the noise variance in the k-th slice of $\boldsymbol{E}$. To simplify presentation, we begin with the Gaussian noise case.

Theorem 1 (Distributional guarantees for tensor factor estimates (Gaussian noise))**.** *Suppose that the $E_{i,j,k}$'s are Gaussian, and that Assumptions 1-3 hold. Assume that $\mu,\kappa,r=O(1)$ and that $t_0=c_0\log d$,*

$$p\ge\frac{c_1\log^5 d}{d^{3/2}},\qquad \frac{c_2}{d^{100}}\le\frac{\sigma_{\max}}{\|\boldsymbol{T}^\star\|_\infty}\le c_3\sqrt{\frac{pd^{3/2}}{\log^4 d}} \tag{14}$$

for some sufficiently large (resp. small) constants $c_0,c_1,c_2>0$ (resp. $c_3>0$). Then with probability at least $1-o(1)$, one has the following decomposition:

$$\boldsymbol{U}\boldsymbol{\Pi}-\boldsymbol{U}^\star=\boldsymbol{Z}+\boldsymbol{W},$$

where $\boldsymbol{\Pi}$ is defined in (11), $\|\boldsymbol{W}\|_{2,\infty}=o\big(\frac{\sigma_{\min}}{\lambda_{\max}^{\star 2/3}\sqrt{p}}\big)$, and for any $1\le k\le d$, $\boldsymbol{Z}_{k,:}\sim\mathcal{N}(\boldsymbol{0},\boldsymbol{\Sigma}_k^\star)$ with

$$\boldsymbol{\Sigma}_k^\star := \tfrac{2}{p}\big(\widetilde{\boldsymbol{U}}^{\star\top}\widetilde{\boldsymbol{U}}^\star\big)^{-1}\widetilde{\boldsymbol{U}}^{\star\top}\boldsymbol{D}_k^\star\widetilde{\boldsymbol{U}}^\star\big(\widetilde{\boldsymbol{U}}^{\star\top}\widetilde{\boldsymbol{U}}^\star\big)^{-1}. \tag{15}$$

Remark 1. As an interpretation of Condition (14): (i) the sample complexity pd^3 is $O(d^{3/2}\mathrm{poly}\log(d))$, which is widely conjectured to be computationally optimal (up to some log factor) (Barak & Moitra, 2016); (ii) the typical size of each noise component (as captured by $\{\sigma_{i,j,k}\}$) is allowed to be substantially larger than the maximum magnitude of the entries of $\boldsymbol{T}^\star$ under our sample size assumption.

In words, Theorem 1 reveals that the estimation error of $\boldsymbol{U}$ can be decomposed into a Gaussian component $\boldsymbol{Z}$ and a residual term $\boldsymbol{W}$. Encouragingly, the residual term $\boldsymbol{W}$ is, in some sense, dominated by the Gaussian term and can be safely neglected. To see this, recall that $\sigma_{i,j,k}\ge\sigma_{\min}$, leading to a lower bound[5]

$$\begin{aligned}\boldsymbol{\Sigma}_k^\star &\succeq \tfrac{2\sigma_{\min}^2}{p}\big(\widetilde{\boldsymbol{U}}^{\star\top}\widetilde{\boldsymbol{U}}^\star\big)^{-1}\succeq\tfrac{(2-o(1))\sigma_{\min}^2}{p}\mathrm{diag}\big(\|\boldsymbol{u}_i^\star\|_2^{-4}\big)\\ &\succeq(1-o(1))\tfrac{2\sigma_{\min}^2}{p\lambda_{\max}^{\star 4/3}}\boldsymbol{I}.\end{aligned}$$

This tells us that the typical ℓ_2 norm of each row $\boldsymbol{Z}_{k,:}$ exceeds the order of $\frac{\sigma_{\min}\sqrt{r}}{\sqrt{p}\lambda_{\max}^{\star 2/3}}$, which is hence much larger than $\|\boldsymbol{W}\|_{2,\infty}$ (by virtue of Theorem 1). To conclude, the nonconvex estimate $\boldsymbol{U}$ is — up to global permutation — a nearly un-biased estimate of the true tensor factors $\boldsymbol{U}^\star$, with estimation errors being approximately Gaussian.

As it turns out, this distributional characterization can be extended to accommodate a much broader class of noise beyond Gaussian noise, as stated below.

Theorem 2 (Distributional guarantees for tensor factor estimates (general noise))**.** *Suppose that $\{E_{i,j,k}\}$ are not necessarily Gaussian but satisfy Assumption 2. Then the decomposition in Theorem 1 continues to hold, except that $\boldsymbol{Z}$ is not necessarily Gaussian but instead obeys*

$$\big|\mathbb{P}\big\{\boldsymbol{Z}_{k,:}\in\mathcal{A}\big\}-\mathbb{P}\big\{\boldsymbol{g}_k\in\mathcal{A}\big\}\big|=o(1),\quad 1\le k\le d$$

for any convex set $\mathcal{A}\subset\mathbb{R}^r$. Here, $\boldsymbol{g}_k\sim\mathcal{N}(\boldsymbol{0},\boldsymbol{\Sigma}_k^\star)$ with covariance matrix $\boldsymbol{\Sigma}_k^\star$ defined in (15).

Before continuing, there is another important observation that is worth pointing out (which is not stated in Theorems 1-2 but will be made precise in the analysis): for any three different rows i,j,k, the corresponding errors $\boldsymbol{Z}_{i,:}$, $\boldsymbol{Z}_{j,:}$ and $\boldsymbol{Z}_{k,:}$ are "nearly" statistically independent. This is a crucial observation that immediately leads to entrywise distributional characterizations for the resulting tensor estimate $\boldsymbol{T}$, as summarized below.

Theorem 3 (Distributional guarantees for tensor entry estimates)**.** *Instate the assumptions of Theorem 2. Consider any $1\le i\le j\le k\le d$ obeying*

$$\frac{\big\|\widetilde{\boldsymbol{U}}^\star_{(j,k),:}\big\|_2+\big\|\widetilde{\boldsymbol{U}}^\star_{(i,j),:}\big\|_2+\big\|\widetilde{\boldsymbol{U}}^\star_{(i,k),:}\big\|_2}{\|\widetilde{\boldsymbol{U}}^\star\|_{2,\infty}}\ge c_7\sigma_{\max}\sqrt{\frac{\log^2 d}{d^2p}} \tag{16}$$

for some large constant $c_7>0$, with $\widetilde{\boldsymbol{U}}^\star$ defined in (12). Then the estimate $\boldsymbol{T}$ defined in (10) obeys

$$\sup_{\tau\in\mathbb{R}}\Big|\mathbb{P}\Big\{T_{i,j,k}\le T_{i,j,k}^\star+\tau\sqrt{v_{i,j,k}^\star}\Big\}-\Phi(\tau)\Big|=o(1), \tag{17}$$

[5] To see why the penultimate inequality holds, note that under our assumptions,

$$\begin{aligned}\widetilde{\boldsymbol{U}}^{\star\top}\widetilde{\boldsymbol{U}}^\star&=\big[(\boldsymbol{u}_i^{\star\top}\boldsymbol{u}_j^\star)^2\big]_{1\le i,j\le r}\preceq\mathsf{diag}\big(\|\boldsymbol{u}_i^\star\|_2^4\big)\\&\quad+\big(r\max_{i\ne j}(\boldsymbol{u}_i^{\star\top}\boldsymbol{u}_j^\star)^2\big)\boldsymbol{I}=(1+o(1))\mathsf{diag}\big(\|\boldsymbol{u}_i^\star\|_2^4\big).\end{aligned}$$

where $\Phi(\cdot)$ is the CDF of a standard Gaussian random variable. Here, the variance parameters $\{v^\star_{i,j,k}\}$ are defined such that for any three distinct numbers i, j, k,

$$v^\star_{i,j,k} := \widetilde{\boldsymbol{U}}^\star_{(j,k),:}\boldsymbol{\Sigma}^\star_i\big(\widetilde{\boldsymbol{U}}^\star_{(j,k),:}\big)^\top + \widetilde{\boldsymbol{U}}^\star_{(i,k),:}\boldsymbol{\Sigma}^\star_j\big(\widetilde{\boldsymbol{U}}^\star_{(i,k),:}\big)^\top + \widetilde{\boldsymbol{U}}^\star_{(i,j),:}\boldsymbol{\Sigma}^\star_k\big(\widetilde{\boldsymbol{U}}^\star_{(i,j),:}\big)^\top, \tag{18a}$$

$$v^\star_{i,i,k} := 4\,\widetilde{\boldsymbol{U}}^\star_{(i,k),:}\boldsymbol{\Sigma}^\star_i\big(\widetilde{\boldsymbol{U}}^\star_{(i,k),:}\big)^\top + \widetilde{\boldsymbol{U}}^\star_{(i,i),:}\boldsymbol{\Sigma}^\star_k\big(\widetilde{\boldsymbol{U}}^\star_{(i,i),:}\big)^\top, \tag{18b}$$

$$v^\star_{i,i,i} := 9\,\widetilde{\boldsymbol{U}}^\star_{(i,i),:}\boldsymbol{\Sigma}^\star_i\big(\widetilde{\boldsymbol{U}}^\star_{(i,i),:}\big)^\top, \tag{18c}$$

where $\boldsymbol{\Sigma}^\star_k$ is defined in (15).

In short, the above theorem indicates that: if the "strength" of a tensor entry $T^\star_{i,j,k}$ is not exceedingly small, then our nonconvex estimate of this entry is nearly unbiased, whose estimation error is approximately zero-mean Gaussian with variance $v^\star_{i,j,k}$ (which admits a closed-form expression). To see this, note that when (14) holds, the right-hand side of Condition (16) is at most $O\left(d^{-1/4}/\sqrt{\log d}\right)$, which is vanishingly small. In other words, the Gaussian approximation is nearly tight unless the energy $\big\|\widetilde{\boldsymbol{U}}^\star_{(j,k),:}\big\|_2 + \big\|\widetilde{\boldsymbol{U}}^\star_{(i,j),:}\big\|_2 + \big\|\widetilde{\boldsymbol{U}}^\star_{(i,j),:}\big\|_2$ is vanishingly small compared to the average size. This entrywise distributional theory allows one to accommodate a broad family of noise models.

3.2. Confidence intervals

The preceding distributional guarantees pave the way for uncertainty quantification. To achieve this, it remains to compute the unknown covariance matrices $\{\boldsymbol{\Sigma}^\star_k\}$ and the variance parameters $\{v^\star_{i,j,k}\}$, which are functions of both the ground truth $\{\boldsymbol{u}^\star_l\}$ and the noise variance $\{\sigma^2_{i,j,k}\}$ and are not known *a priori*. In particular, in the heteroscedastic case where $\{\sigma^2_{i,j,k}\}$ are location-varying, it might be infeasible to estimate all variance parameters reliably.

Variance and covariance estimation. Fortunately, despite the absence of prior knowledge about the truth and the noise parameters, we are still able to faithfully estimate these important parameters from the data at hand, using simple plug-in rules. Specifically:

1. Rather than estimating all $\{\sigma_{i,j,k}\}$ directly, we turn attention to estimating the noise components $\{E_{i,j,k}\}$ instead, with the assistance of our tensor estimate $\boldsymbol{T}$ as follows

$$\widehat{E}_{i,j,k} := T^{\mathsf{obs}}_{i,j,k} - T_{i,j,k}, \qquad (i,j,k)\in\Omega. \tag{19}$$

We then construct a diagonal matrix $\boldsymbol{D}_k \in \mathbb{R}^{d^2\times d^2}$ obeying

$$\big(\boldsymbol{D}_k\big)_{(i,j),(i,j)} = p^{-1}\widehat{E}^2_{i,j,k}\,\mathbb{1}_{\{(i,j,k)\in\Omega\}}. \tag{20}$$

Note that $\boldsymbol{D}_k$ $(1\le k\le d)$ is not really a faithful estimate of the $\boldsymbol{D}^\star_k$ defined in (13), but it suffices for our purpose.

2. Estimate $\widetilde{\boldsymbol{U}}^\star$ via the plug-in estimator

$$\widetilde{\boldsymbol{U}} := [\boldsymbol{u}_1\otimes\boldsymbol{u}_1, \cdots, \boldsymbol{u}_r\otimes\boldsymbol{u}_r] \in \mathbb{R}^{d^2\times r}.$$

3. Substitute the above estimators into the expressions of the (co)-variance parameters to yield our estimates. Specifically, for any $1\le k\le d$, we compute

$$\boldsymbol{\Sigma}_k = \tfrac{2}{p}(\widetilde{\boldsymbol{U}}^\top\widetilde{\boldsymbol{U}})^{-1}\widetilde{\boldsymbol{U}}^\top\boldsymbol{D}_k\widetilde{\boldsymbol{U}}(\widetilde{\boldsymbol{U}}^\top\widetilde{\boldsymbol{U}})^{-1} \tag{21}$$

as an estimate of $\boldsymbol{\Sigma}^\star_k$. We also produce estimates for $\{v^\star_{i,j,k}\}$ as follows: for any three distinct numbers $1\le i,j,k\le d$,

$$v_{i,j,k} := \widetilde{\boldsymbol{U}}_{(j,k),:}\boldsymbol{\Sigma}_i\big(\widetilde{\boldsymbol{U}}_{(j,k),:}\big)^\top + \widetilde{\boldsymbol{U}}_{(i,k),:}\boldsymbol{\Sigma}_j\big(\widetilde{\boldsymbol{U}}_{(i,k),:}\big)^\top + \widetilde{\boldsymbol{U}}_{(i,j),:}\boldsymbol{\Sigma}_k\big(\widetilde{\boldsymbol{U}}_{(i,j),:}\big)^\top; \tag{22a}$$

$$v_{i,i,k} := 4\,\widetilde{\boldsymbol{U}}_{(i,k),:}\boldsymbol{\Sigma}_i\big(\widetilde{\boldsymbol{U}}_{(i,k),:}\big)^\top + \widetilde{\boldsymbol{U}}_{(i,i),:}\boldsymbol{\Sigma}_k\big(\widetilde{\boldsymbol{U}}_{(i,i),:}\big)^\top; \tag{22b}$$

$$v_{i,i,i} := 9\,\widetilde{\boldsymbol{U}}_{(i,i),:}\boldsymbol{\Sigma}_i\big(\widetilde{\boldsymbol{U}}_{(i,i),:}\big)^\top. \tag{22c}$$

Confidence intervals. With the above variance and covariance estimates in place, we are positioned to introduce our uncertainty quantification procedure. This is accomplished by constructing *entrywise* confidence intervals for both the tensor factors and the unknown tensor as follows.

- For each $1\le k\le d$ and $1\le l\le r$, we construct a $(1-\alpha)$-confidence interval for the k-th entry of the l-th tensor factor (up to global permutation) as follows

$$\mathsf{CI}^{1-\alpha}_{u_{l,k}} := \big[u_{l,k} \pm \sqrt{(\boldsymbol{\Sigma}_k)_{l,l}}\cdot\Phi^{-1}(1-\alpha/2)\big], \tag{23}$$

  where $\Phi^{-1}(\cdot)$ is the inverse CDF of $\mathcal{N}(0,1)$, $[a\pm b] := [a-b, a+b]$, and $\boldsymbol{\Sigma}_k$ is constructed in (21).

- For each $1\le i,j,k\le d$, we construct a $(1-\alpha)$-confidence interval for the (i,j,k)-th entry of $\boldsymbol{T}^\star$:

$$\mathsf{CI}^{1-\alpha}_{T_{i,j,k}} := \big[T_{i,j,k} \pm \sqrt{v_{i,j,k}}\cdot\Phi^{-1}(1-\alpha/2)\big], \tag{24}$$

  where $v_{i,j,k}$ is constructed in (22).

As it turns out, the proposed (entrywise) confidence intervals are nearly accurate, as revealed by the following theorem.

Theorem 4 (Validity of confidence intervals)**.** *Instate the assumptions of Theorem 2. There is a permutation $\pi(\cdot): [d]\mapsto[d]$ such that for any $0<\alpha<1$, the confidence interval constructed in (23) obeys*

$$\mathbb{P}\left\{u^\star_{\pi(l),k}\in\mathsf{CI}^{1-\alpha}_{u_{l,k}}\right\} = 1-\alpha+o(1),\ 1\le l\le r, 1\le k\le d.$$

In addition, for any $1\le i,j,k\le d$ obeying (16) and any $0<\alpha<1$, the confidence interval (24) obeys

$$\mathbb{P}\left\{T^\star_{i,j,k}\in\mathsf{CI}^{1-\alpha}_{T_{i,j,k}}\right\} = 1-\alpha+o(1).$$

This theorem justifies the validity of the inferential procedure we propose. Several important features are worth emphasizing:

- *"Fine-grained" entrywise uncertainty quantification*. Our results enable trustworthy uncertainty quantification down to quite fine scales, namely, we are capable of assessing the uncertainty reliably at the entrywise level for both the tensor factors and the tensor of interest. To the best of our knowledge, accurate entrywise uncertainty characterization for noisy tensor completion is previously unavailable.
- *Adaptivity to heterogeneous and unknown noise distributions*. The proposed confidence intervals do not require prior knowledge about the noise distributions, and automatically adapt to noise heteroscedasticity (i.e. the case when the noise variance varies across entries). Such model-free and adaptive features are of eminently practical value.
- *No need of sample splitting*. The whole procedure studied here — including both estimation and uncertainty quantification — does not rely on any sort of data splitting, thus effectively preventing unnecessary information loss due to sample splitting.

Lower bounds. One might naturally wonder whether the proposed confidence intervals can be further improved; concretely, is it possible to identify a shorter confidence interval that remains valid? As it turns out, our procedures are, in some sense, statistically optimal under Gaussian noise, as confirmed by the following fundamental lower bound.

Theorem 5 (Fundamental lower bounds)**.** *Consider any unbiased estimator $\widehat{\boldsymbol{u}}_l$ for $\boldsymbol{u}_l^\star$ $(1 \le l \le r)$ and any unbiased estimator $\widehat{\boldsymbol{T}}$ for $\boldsymbol{T}^\star$. Suppose $\{E_{i,j,k}\}$ are i.i.d. Gaussian. Under the assumptions of Theorem 2, one necessarily has*

$$\mathsf{Var}\big[\widehat{u}_{l,k}\big] \ge (1-o(1))\big(\boldsymbol{\Sigma}_k^\star\big)_{l,l}, \quad 1 \le k \le d; \tag{25a}$$

$$\mathsf{Var}\big[\widehat{T}_{i,j,k}\big] \ge (1-o(1))v_{i,j,k}^\star, \quad 1 \le i,j,k \le d; \tag{25b}$$

$$\mathbb{E}\left[\big\|\widehat{\boldsymbol{u}}_l - \boldsymbol{u}_l^\star\big\|_2^2\right] \ge \frac{(2-o(1))\sigma_{\min}^2 d}{p\big\|\boldsymbol{u}_l^\star\big\|_2^4}, \quad 1 \le l \le r; \tag{25c}$$

$$\mathbb{E}\left[\big\|\widehat{\boldsymbol{T}} - \boldsymbol{T}^\star\big\|_{\mathrm{F}}^2\right] \ge \frac{(6-o(1))\sigma_{\min}^2 dr}{p}. \tag{25d}$$

Taken collectively with Theorems 2 and 3, the above result reveals that our nonconvex estimators $\{\boldsymbol{u}_l\}$ and $\boldsymbol{T}$ achieve minimal mean square estimation errors in a very sharp manner at the entrywise level. Recognizing that the proposed confidence intervals allow for accurate assessment of the uncertainty (by virtue of Theorem 4), we conclude that the proposed entrywise inferential procedures are, in some sense, un-improvable under i.i.d. Gaussian noise (including both the rates and the pre-constants).

3.3. Back to estimation: ℓ_2 optimality of nonconvex estimates

Thus far, we have established the distributions of the estimators $u_{l,k}$ $(1 \le l \le r, 1 \le k \le d)$ and $T_{i,j,k}$ (for those i,j,k obeying (16)). These results taken together allow one to pin down the ℓ_2 risk of the nonconvex optimization approach in a sharp manner. Our result is this:

Theorem 6 (Sharp ℓ_2 estimation risk)**.** *Instate the assumptions of Theorem 2. With probability $1-o(1)$, the estimates returned by Algorithm 1 obey*

$$\big\|\boldsymbol{u}_{\pi(l)} - \boldsymbol{u}_l^\star\big\|_2^2 = \frac{(2+o(1))\sigma_{\max}^2 d}{p\big\|\boldsymbol{u}_l^\star\big\|_2^4}, \quad 1 \le l \le r; \tag{26a}$$

$$\|\boldsymbol{T} - \boldsymbol{T}^\star\|_{\mathrm{F}}^2 = \frac{(6+o(1))\sigma_{\max}^2 dr}{p} \tag{26b}$$

for some permutation $\pi(\cdot) : [d] \mapsto [d]$.

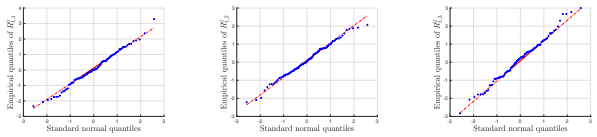
Here, the characterization of the ℓ_2 risk (26a) for $\boldsymbol{u}_l$ is a straightforward consequence of Theorems 1-2. In comparison, establishing the ℓ_2 risk (26b) for $\boldsymbol{T}$ requires more work, as Theorem 3 is valid only for a subset of the entries obeying (16). Fortunately, a majority of the entries of $\boldsymbol{T}^\star$ satisfy (16), thus allowing for a nearly accurate approximation of the Euclidean risk of $\boldsymbol{T}$. Theorem 6 taken together with Theorem 5 delivers an encouraging news: when the noise is i.i.d. Gaussian, nonconvex optimization is information-theoretically optimal in a sharp manner when estimating both the unknown tensor and its underlying tensor factors.

3.4. Numerical experiments

To validate our theory and demonstrate the practical applicability of our inferential procedures, we perform a series of numerical experiments for a variety of settings. Specifically, we set $d = 100$, $p = 0.2$, and generate the ground-truth tensor $\boldsymbol{T}^\star = \sum_{l=1}^r (\boldsymbol{u}_l^\star)^{\otimes 3}$ in a random fashion such that $\boldsymbol{u}_l^\star \overset{\text{i.i.d.}}{\sim} \mathcal{N}(\boldsymbol{0}, \boldsymbol{I}_d)$. Regarding the algorithmic parameters for nonconvex optimization (i.e. Algorithm 1 and Algorithm 2 in Supplementary materials A.1), we choose $L = r^2$, $\epsilon_{\mathsf{th}} = 0.4$, $\eta_t \equiv 3 \times 10^{-5}/p$, and $t_0 = 100$. The noise components are independently drawn from Gaussian distributions, obeying $E_{i,j,k} \sim \mathcal{N}(0, \sigma_{i,j,k}^2), 1 \le i \le j \le k \le d$ with variance $\sigma_{i,j,k}^2$ constructed as follows. We generate $w_{i,j,k} \overset{\text{i.i.d.}}{\sim} \mathsf{Unif}[0,1], 1 \le i,j,k \le d$ and let

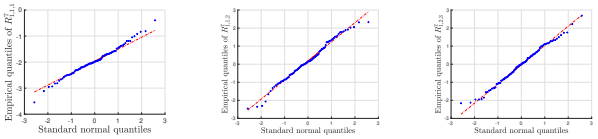
$$\sigma_{i,j,k}^2 = \frac{\sigma^2 w_{i,j,k}^\beta}{\sum_{1 \le i \le j \le k \le d} w_{i,j,k}^\beta} \frac{d^3}{6},$$

where β dictates the degree of heteroscedasticity. The noise becomes more heteroscedastic as β increases, and setting $\beta = 0$ reduces to the homoscedastic case where the noise variances are identical across all entries. In what follows, we shall set $\beta = 5$.



(a) (b) (c)

Figure 1. Q-Q (quantile-quantile) plots of $R^{\mathsf{U}}_{1,1}$, $R^{\mathsf{U}}_{1,2}$ and $R^{\mathsf{U}}_{1,3}$ vs. a standard Gaussian distribution (where $r = 4$, $p = 0.2$, $\sigma = 0.1$ and $\beta = 5$).



(a) (b) (c)

Figure 2. Q-Q (quantile-quantile) plots of $R^{\mathsf{T}}_{1,1,1}$, $R^{\mathsf{T}}_{1,1,2}$ and $R^{\mathsf{T}}_{1,2,3}$ vs. a standard Gaussian distribution (where $r = 4$, $p = 0.2$, $\sigma = 0.1$ and $\beta = 5$).

Table 1. Empirical coverage rates of tensor factor entries for varying r and σ.

(r, σ)	Mean(CR)	Std(CR)
$(2, 10^{-2})$	0.9481	0.0201
$(2, 10^{-1})$	0.9477	0.0228
$(2, 1)$	0.9478	0.0215
$(4, 10^{-2})$	0.9450	0.0218
$(4, 10^{-1})$	0.9472	0.0231
$(4, 1)$	0.9462	0.0234

Table 2. Empirical coverage rates of tensor entries for different r and σ.

(r, σ)	Mean(CR)	Std(CR)
$(2, 10^{-2})$	0.9494	0.0218
$(2, 10^{-1})$	0.9513	0.0218
$(2, 1)$	0.9475	0.0222
$(4, 10^{-2})$	0.9434	0.0225
$(4, 10^{-1})$	0.9494	0.0220
$(4, 1)$	0.9494	0.0219

Tensor factor entries. We begin with inference for the entries of the tensor factors of interest. Consider the construction of 95% confidence intervals (i.e. $\alpha = 0.05$). Define the normalized estimation error as follows

$$R^{\mathsf{U}}_{l,k} := \frac{1}{\sqrt{(\boldsymbol{\Sigma}_k)_{l,l}}} \left(u_{l,k} - u^{\star}_{l,k}\right), \quad 1 \le l \le r,\, 1 \le k \le d.$$

For each $1 \le l \le r$ and $1 \le k \le d$, we denote by $\mathsf{CR}_{l,k}$ the empirical coverage rate for the tensor factor entry $u^{\star}_{l,k}$ over 100 independent trials. Let $\mathsf{Mean(CR)}$ (resp. $\mathsf{Std\,(CR)}$) denote the average (resp. the standard deviation) of $\{\mathsf{CR}_{l,k}\}$ over all tensor factor entries. Figure 1 displays the Q-Q (quantile-quantile) plots of $R^{\mathsf{U}}_{1,1}$, $R^{\mathsf{U}}_{1,2}$ and $R^{\mathsf{U}}_{1,3}$ vs. a standard Gaussian random variable, and Table 1 summarizes the numerical results for varying r and σ. Encouragingly, the empirical coverage rates are all very close to 95%, and the empirical distributions of the normalized estimation errors are all well approximated by a standard Gaussian distribution.

Tensor entries. Next, we turn to inference for tensor entries. Similar to the above case, we intend to construct 95% confidence intervals. Define

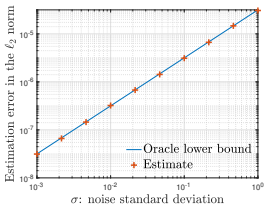
$$R^{\mathsf{T}}_{i,j,k} := \frac{1}{\sqrt{v_{i,j,k}}} \left(T_{i,j,k} - T^{\star}_{i,j,k}\right), \quad 1 \le i \le j \le k \le d.$$

For each $1 \le i \le j \le k \le d$, we record the empirical coverage rate $\mathsf{CR}_{i,j,k}$ for the tensor entry $T^{\star}_{i,j,k}$ over 100 Monte Carlo trials. Denote by $\mathsf{Mean(CR)}$ (resp. $\mathsf{Std(CR)}$) the average (resp. the standard deviation) of $\{\mathsf{CR}_{i,j,k}\}$ over entries $1 \le i \le j \le k \le d$. Figure 2 depicts the Q-Q (quantile-quantile) plots of $R^{\mathsf{T}}_{1,1,1}$, $R^{\mathsf{T}}_{1,1,2}$ and $R^{\mathsf{T}}_{1,2,3}$ vs. a standard Gaussian random variable. Table 2 collects the numerical results $\mathsf{Mean(CR)}$ and $\mathsf{Std(CR)}$ for a variety of settings. Similar to previous experiments, the confidence intervals and the Q-Q plots match our theoretical prediction in a reasonably well manner.

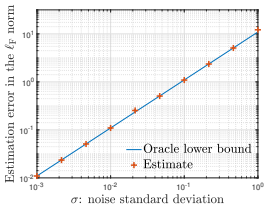
ℓ_2 estimation accuracy. Finally, we verify the Euclidean estimation guarantees we develop for Algorithm 1 in Theorem 6. Figure 3 plots the Euclidean estimation errors of the tensor factor estimate $\boldsymbol{u}_1$ (resp. the tensor estimate $\boldsymbol{T}$). In this series of experiments, we focus on the homoskedastic case, i.e. $\beta = 0$. As one can see, the empirical ℓ_2 risks are exceedingly close to the Cramér-Rao lower bounds supplied in Theorem 5.

4. Prior art

Much progress has been made towards solving low-rank tensor completion. Inspired by the success of convex relaxation for matrix completion (Candès & Recht, 2009; Candès & Plan, 2010; Gross, 2011; Li, 2013; Chen et al., 2019c), an estimate based on tensor nuclear norm minimization was proposed by (Yuan & Zhang, 2016; 2017), which enables information-theoretically optimal sample complexity. Unfortunately, the tensor nuclear norm is itself NP-hard to compute and hence computationally infeasible in practice. To allow for more economical algorithms, a widely adopted strategy is to unfold the tensor data into a matrix (Tomioka et al., 2010; Gandy et al., 2011; Liu et al., 2013; Mu et al., 2014), thus transforming it into a low-rank matrix completion problem (Candès & Recht, 2009; Keshavan et al., 2010a; Chi et al., 2019). However, unfolding a third-order tensor often leads to an extremely unbalanced matrix,



(a)



(b)

Figure 3. (a) ℓ_2 estimation error of $\boldsymbol{u}_1$ vs. the Cramér–Rao lower bound; (b) Euclidean estimation errors of $\boldsymbol{T}$ vs. the Cramer-Rao lower bound (where $r = 4$, $p = 0.2$ and $\beta = 0$).

thereby resulting in sub-optimal sample complexity when directly invoking matrix completion theory. To address this issue, a recent line of work (Barak & Moitra, 2016; Potechin & Steurer, 2017) suggested the use of sum-of-squares (SOS) hierarchy, which performs convex relaxation after lifting the data into higher dimension. The SOS-based algorithms achieve a sample complexity on the order of $rd^{3/2}$ for third-order tensors, which is widely conjectured to be optimal among all polynomial-time algorithms. However, despite their polynomial-time complexity, the SOS-based methods remain too expensive for solving large-scale practical problems, primarily due to the lifting operation.

Motivated by the above computational concerns, several nonconvex approaches have been developed, which often consist of two stages: (1) finding a rough initialization; (2) local refinement. Existing initialization schemes include unfolding-based spectral methods (Xia & Yuan, 2019a; Xia et al., 2020; Montanari & Sun, 2018; Cai et al., 2020a; 2019; Liu & Moitra, 2020). tensor power methods (Jain & Oh, 2014), tensor SVD (Zhang & Aeron, 2017), and so on. To improve the estimation accuracy, the local refinement stage invokes nonconvex optimization algorithms like alternating minimization (Jain & Oh, 2014; Liu & Moitra, 2020), gradient descent (Cai et al., 2019; Han et al., 2020), manifold-based optimization (Xia & Yuan, 2019a), block coordinate decent (Ji et al., 2016), etc. These were motivated in part by the effectiveness of nonconvex optimization in solving nonconvex low-complexity problems (Burer & Monteiro, 2003; Srebro, 2004; Keshavan et al., 2010a;b; Jain et al., 2013; Chen & Candès, 2017; Chen & Wainwright, 2015; Ma et al., 2019; Chen & Candès, 2018; Chen et al., 2019b; Netrapalli et al., 2014; Hao et al., 2020; Zhang et al., 2016; Cai et al., 2017; Chen et al., 2019a; Wang & Giannakis, 2016; Chen et al., 2020; Sun et al., 2018; Qu et al., 2019; Tong et al., 2020; Zhang et al., 2017); see an overview of recent development in (Chi et al., 2019). Various statistical and computational guarantees have been provided for these algorithms, all of which have been shown to run in polynomial time. In particular, (unfolding-based) spectral initialization followed by gradient descent converges linearly to an accuracy that is within a logarithmic factor from optimal (Cai et al., 2019).

None of the above results, however, suggested how to evaluate the uncertainty of the resulting estimates in a meaningful way. Despite a large body of work on statistical estimation for noisy tensor completion, it remains completely unclear how to exploit existing results to construct valid yet short confidence intervals for the unknown tensor. Perhaps the work closest to the current paper is inference and uncertainty quantification for noisy matrix completion and matrix denoising (Chen et al., 2019d; Xia & Yuan, 2019b; Cheng et al., 2020), which enables optimal construction of confidence intervals on the basis of nonconvex matrix completion algorithms. Inference for singular subspaces has also been investigated under both low-rank matrix regression and denoising settings (Xia, 2018; 2019). While these results might potentially be applicable here by first matricizing the data, the resulting sample complexity, as discussed above, could be pessimistic. Finally, construction of confidence intervals has been extensively studied in a variety of high-dimensional sparse estimation settings (Zhang & Zhang, 2014; van de Geer et al., 2014; Javanmard & Montanari, 2014; Ren et al., 2015; Cai et al., 2016; Ning & Liu, 2017; Cai & Guo, 2017; Sur et al., 2019; Janková & van de Geer, 2018; Miolane & Montanari, 2018). Both the inferential approaches and analysis techniques therein, however, are drastically different from the ones employed for inference for either tensor completion or matrix completion.

5. Discussions

This paper has explored the problem of uncertainty quantification for nonconvex tensor completion. The main contributions lie in establishing (nearly) precise distributional guarantees for the nonconvex estimates down to an entrywise level. Our distributional representation enables data-driven construction of confidence intervals for both the unknown tensor and its underlying tensor factors. Our inferential procedure and the accompanying theory are model agnostic, which do not require prior knowledge about the noise distributions and are fully adaptive to location-varying noise levels. Our results uncover the unreasonable effectiveness of nonconvex optimization, which are statistically optimal for both estimation and confidence interval construction.

The findings of the current paper further suggest numerous possible extensions that are worth pursuing. To begin with, our current results are only optimal when both the rank r and the condition number κ are constants independent of the ambient dimension d. Can we further refine the analysis to enable optimal inference for more general settings? It would also be interesting to go beyond uniform random sampling by considering the type of sampling patterns with a heterogeneous missingness mechanism.

A. More details about Algorithm 1

A.1. The initialization scheme

For self-completeness, we record in this section the detailed initialization procedure employed in the two-stage nonconvex algorithm proposed in (Cai et al., 2019) (namely, Algorithm 1). This is summarized in Algorithm 2, with auxiliary procedures detailed in Algorithm 3. As a high-level interpretation, Algorithm 2 estimates the subspace spanned by the tensor factor $\{\boldsymbol{u}_l^\star\}_{1\le l\le r}$ via a spectral method (similar to PCA-type methods (Montanari & Sun, 2018; Zhang et al., 2018; Cai et al., 2020a)), whereas Algorithm 3 attempts to retrieve estimates for individual tensor factors from this subspace estimate $\boldsymbol{U}_{\mathsf{space}}$. Here and throughout, we denote $\boldsymbol{T}^{\mathsf{obs}} := [T^{\mathsf{obs}}_{i,j,k}]_{1\le i,j,k\le d}$, where we set $T^{\mathsf{obs}}_{i,j,k} = 0$ for any $(i,j,k)\notin\Omega$. In addition, for any vector $\boldsymbol{u}\in\mathbb{R}^d$, we define the vector product of a tensor $\boldsymbol{T}\in\mathbb{R}^{d\times d\times d}$ — denoted by $\boldsymbol{T}\times_3\boldsymbol{u}\in\mathbb{R}^{d\times d}$ — such that $\big[\boldsymbol{T}\times_3\boldsymbol{u}\big]_{ij} := \sum_{1\le k\le d} T_{i,j,k}u_k, \forall i,j\in[d]$.

Algorithm 2 Spectral initialization for nonconvex tensor completion

1: **Input:** tensor $\boldsymbol{T}^{\mathsf{obs}}$.
2: Let $\boldsymbol{U}_{\mathsf{space}}\boldsymbol{\Lambda}\boldsymbol{U}_{\mathsf{space}}^\top$ be the rank-r eigen-decomposition of $\mathcal{P}_{\mathsf{off\text{-}diag}}(\boldsymbol{A}\boldsymbol{A}^\top)$, where $\boldsymbol{A} = \mathsf{unfold}\big(\boldsymbol{T}^{\mathsf{obs}}\big)$ is the mode-1 matricization of $\boldsymbol{T}^{\mathsf{obs}}$, and $\mathcal{P}_{\mathsf{off\text{-}diag}}(\boldsymbol{Z})$ extracts out the off-diagonal entries of $\boldsymbol{Z}$.
3: **Output:** an initial estimate $\boldsymbol{U}^0\in\mathbb{R}^{d\times r}$ on the basis of $\boldsymbol{U}_{\mathsf{space}}\in\mathbb{R}^{d\times r}$ using Algorithm 3.

A.2. Choices of algorithmic parameters

To guarantee fast convergence of Algorithm 1, there are a couple of algorithmic parameters — namely, the number of restart attempts L, the pruning threshold ϵ_{th} in Algorithm 3, as well as the learning rates η_t — that need to be properly chosen. Unless otherwise noted, this paper adopts the following choices suggested by (Cai et al., 2019):

$$L = c_4 r^{2\kappa^2}\log^{3/2} d, \quad \eta_t \equiv c_5\frac{\lambda_{\min}^{\star 4/3}}{p\lambda_{\max}^{\star 8/3}},$$
$$\epsilon_{\mathsf{th}} = c_6\bigg(\frac{\mu r\log d}{d\sqrt{p}} + \frac{\sigma_{\min}}{\lambda_{\min}^\star}\sqrt{\frac{rd\log^2 d}{p}} + \sqrt{\frac{\mu r\log d}{d}}\bigg),$$

where $c_4 > 0$ is some sufficiently large constant, and $c_5, c_6 > 0$ are some sufficiently small constants. The interested reader is referred to (Cai et al., 2019) for justification.

Algorithm 3 Retrieval of low-rank tensor factors from a given subspace estimate.

1: **Input:** number of restarts L, pruning threshold ϵ_{th}, subspace estimate $\boldsymbol{U}_{\mathsf{space}}\in\mathbb{R}^{d\times r}$ given by Algorithm 2.
2: **for** $\tau = 1,\dots,L$ **do**
3: Generate an independent vector $\boldsymbol{g}^\tau\sim\mathcal{N}(0,\boldsymbol{I}_d)$.
4: Compute
$$\big(\boldsymbol{\nu}^\tau,\lambda_\tau,\mathsf{gap}_\tau\big) \leftarrow \textbf{Algorithm 4}\,\big(\boldsymbol{T}^{\mathsf{obs}}, p, \boldsymbol{U}_{\mathsf{space}}, \boldsymbol{g}^\tau\big).$$
5: **end for**
6: Generate tensor factor estimates
$$\big\{(\boldsymbol{w}^1,\lambda_1),\dots,(\boldsymbol{w}^r,\lambda_r)\big\} \leftarrow$$
$$\textbf{Algorithm 5}\,\Big(\big\{\big(\boldsymbol{\nu}^\tau,\lambda_\tau,\mathsf{gap}_\tau\big)\big\}_{\tau=1}^L, \epsilon_{\mathsf{th}}\Big).$$
7: **Output:** initial estimate
$$\boldsymbol{U}^0 = \big[\lambda_1^{1/3}\boldsymbol{w}^1,\dots,\lambda_r^{1/3}\boldsymbol{w}^r\big].$$

Algorithm 4 Retrieve-one-tensor-factor

1: **Input:** tensor $\boldsymbol{T}$, sampling rate p, subsapce estimate $\boldsymbol{U}_{\mathsf{space}}$, Gaussian random vector $\boldsymbol{g}$.
2: Compute
$$\boldsymbol{\theta} = \boldsymbol{U}_{\mathsf{space}}\boldsymbol{U}_{\mathsf{space}}^\top\boldsymbol{g}, \tag{27a}$$
$$\boldsymbol{M} = p^{-1}\boldsymbol{T}^{\mathsf{obs}}\times_3\boldsymbol{\theta}, \tag{27b}$$
3: Let $\boldsymbol{\nu}$ be the leading singular vector of $\boldsymbol{M}$ obeying $\langle\boldsymbol{T}^{\mathsf{obs}},\boldsymbol{\nu}^{\otimes 3}\rangle\ge 0$, and set $\lambda = \langle p^{-1}\boldsymbol{T}^{\mathsf{obs}},\boldsymbol{\nu}^{\otimes 3}\rangle$.
4: **Output:** $\big(\boldsymbol{\nu},\lambda,\sigma_1(\boldsymbol{M})-\sigma_2(\boldsymbol{M})\big)$.

Algorithm 5 Prune

1: **Input:** tensor estimate tuples $\big\{\big(\boldsymbol{\nu}^\tau,\lambda_\tau,\mathsf{gap}_\tau\big)\big\}_{\tau=1}^L$, pruning threshold ϵ_{th}.
2: Set $\Theta = \big\{\big(\boldsymbol{\nu}^\tau,\lambda_\tau,\mathsf{gap}_\tau\big)\big\}_{\tau=1}^L$.
3: **for** $i = 1,\dots,r$ **do**
4: Choose $(\boldsymbol{\nu}^\tau,\lambda_\tau,\mathsf{gap}_\tau)$ from Θ with the largest gap_τ; set $\boldsymbol{w}^i = \boldsymbol{\nu}^\tau$ and $\lambda_i = \lambda_\tau$.
5: Update $\Theta \leftarrow \Theta\setminus\big\{\big(\boldsymbol{\nu}^\tau,\lambda_\tau,\mathsf{gap}_\tau\big)\in\Theta : |\langle\boldsymbol{\nu}^\tau,\boldsymbol{w}^i\rangle| > 1-\epsilon_{\mathsf{th}}\big\}$.
6: **end for**
7: **Output:** $\big\{(\boldsymbol{w}^1,\lambda_1),\dots,(\boldsymbol{w}^r,\lambda_r)\big\}$.

Acknowledgements

Y. Chen is supported in part by AFOSR YIP award FA9550-19-1-0030, ONR N00014-19-1-2120, ARO YIP W911NF-20-1-0097 and W911NF-18-1-0303, NSF CCF-1907661, IIS-1900140 and DMS-2014279, and Princeton SEAS Innovation Award. H. V. Poor is supported in part by the National Science Foundation under Grant CCF-1908308,

and a Princeton Schmidt Data-X Research Award. C. Cai is supported in part by Gordon Y. S. Wu Fellowships in Engineering.

References

Anandkumar, A., Ge, R., Hsu, D., Kakade, S. M., and Telgarsky, M. Tensor decompositions for learning latent variable models. *The Journal of Machine Learning Research*, 15:2773–2832, 2014.

Barak, B. and Moitra, A. Noisy tensor completion via the sum-of-squares hierarchy. In *Proceedings of the Conference on Learning Theory*, pp. 417–445, 2016.

Burer, S. and Monteiro, R. D. A nonlinear programming algorithm for solving semidefinite programs via low-rank factorization. *Mathematical Programming*, 95(2):329–357, 2003.

Cai, C., Li, G., Poor, H. V., and Chen, Y. Nonconvex low-rank tensor completion from noisy data. In *Advances in Neural Information Processing Systems*, pp. 1861–1872, 2019.

Cai, C., Li, G., Chi, Y., Poor, H. V., and Chen, Y. Subspace estimation from unbalanced and incomplete data matrices: $\ell_{2,\infty}$ statistical guarantees. *The Annals of Statistics,* to appear, 2020a.

Cai, C., Poor, H. V., and Chen, Y. Uncertainty quantification for nonconvex tensor completion: Confidence intervals, heteroscedasticity and optimality. *arXiv preprint arXiv:2006.08580*, 2020b.

Cai, J.-F., Liu, H., and Wang, Y. Fast rank one alternating minimization algorithm for phase retrieval. *arXiv preprint arXiv:1708.08751*, 2017.

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

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

Candès, E. and Plan, Y. Matrix completion with noise. *Proceedings of the IEEE*, 98(6):925 –936, June 2010.

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

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

Chen, Y. and Candès, E. The projected power method: An efficient algorithm for joint alignment from pairwise differences. *Communications on Pure and Applied Mathematics*, 71(8):1648–1714, 2018.

Chen, Y. and Candès, E. J. Solving random quadratic systems of equations is nearly as easy as solving linear systems. *Communications on Pure and Applied Mathematics*, 70(5):822–883, 2017.

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

Chen, Y., Chi, Y., Fan, J., and Ma, C. Gradient descent with random initialization: Fast global convergence for nonconvex phase retrieval. *Mathematical Programming*, 176(1-2):5–37, July 2019b.

Chen, Y., Chi, Y., Fan, J., Ma, C., and Yan, Y. Noisy matrix completion: Understanding statistical guarantees for convex relaxation via nonconvex optimization. *arXiv:1902.07698*, 2019c.

Chen, Y., Fan, J., Ma, C., and Yan, Y. Inference and uncertainty quantification for noisy matrix completion. *Proceedings of the National Academy of Sciences of the U.S.A.*, 116(46):22931–22937, 2019d.

Chen, Y., Fan, J., Ma, C., and Yan, Y. Bridging convex and nonconvex optimization in robust PCA: Noise, outliers, and missing data. *arXiv preprint arXiv:2001.05484*, 2020.

Cheng, C., Wei, Y., and Chen, Y. Tackling small eigen-gaps: Fine-grained eigenvector estimation and inference under heteroscedastic noise. *arXiv preprint arXiv:2001.04620*, 2020.

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

Gandy, S., Recht, B., and Yamada, I. Tensor completion and low-n-rank tensor recovery via convex optimization. *Inverse Problems*, 27(2):025010, 2011.

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

Han, R., Willett, R., and Zhang, A. An optimal statistical and computational framework for generalized tensor estimation. *arXiv preprint arXiv:2002.11255*, 2020.

Hao, B., Zhang, A., and Cheng, G. Sparse and low-rank tensor estimation via cubic sketchings. *IEEE Transactions on Information Theory*, 2020.

Hillar, C. J. and Lim, L.-H. Most tensor problems are np-hard. *Journal of the ACM (JACM)*, 60(6):1–39, 2013.

Huang, B., Mu, C., Goldfarb, D., and Wright, J. Provable models for robust low-rank tensor completion. *Pacific Journal of Optimization*, 11(2):339–364, 2015.

Jain, P. and Oh, S. Provable tensor factorization with missing data. In *Advances in Neural Information Processing Systems*, pp. 1431–1439, 2014.

Jain, P., Netrapalli, P., and Sanghavi, S. Low-rank matrix completion using alternating minimization. In *Proceedings of the Forty-fifth Annual ACM Symposium on Theory of Computing*, pp. 665–674, 2013.

Janková, J. and van de Geer, S. De-biased sparse pca: Inference and testing for eigenstructure of large covariance matrices. *arXiv preprint arXiv:1801.10567*, 2018.

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

Ji, T.-Y., Huang, T.-Z., Zhao, X.-L., Ma, T.-H., and Liu, G. Tensor completion using total variation and low-rank matrix factorization. *Information Sciences*, 326:243–257, 2016.

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

Keshavan, R. H., Montanari, A., and Oh, S. Matrix completion from noisy entries. *The Journal of Machine Learning Research*, 11:2057–2078, 2010b. ISSN 1532-4435.

Kolda, T. G. and Bader, B. W. Tensor decompositions and applications. *SIAM Review*, 51(3):455–500, 2009.

Kreimer, N., Stanton, A., and Sacchi, M. D. Tensor completion based on nuclear norm minimization for 5D seismic data reconstruction. *Geophysics*, 78(6):V273–V284, 2013.

Li, X. Compressed sensing and matrix completion with constant proportion of corruptions. *Constructive Approximation*, 37:73–99, 2013.

Liu, A. and Moitra, A. Tensor completion made practical. *arXiv preprint arXiv:2006.03134*, 2020.

Liu, J., Musialski, P., Wonka, P., and Ye, J. Tensor completion for estimating missing values in visual data. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(1):208–220, 2013.

Ma, C., Wang, K., Chi, Y., and Chen, Y. Implicit regularization in nonconvex statistical estimation: Gradient descent converges linearly for phase retrieval, matrix completion, and blind deconvolution. *Foundations of Computational Mathematics*, pp. 1–182, 2019.

Miolane, L. and Montanari, A. The distribution of the lasso: Uniform control over sparse balls and adaptive parameter tuning. *arXiv preprint arXiv:1811.01212*, 2018.

Montanari, A. and Sun, N. Spectral algorithms for tensor completion. *Communications on Pure and Applied Mathematics*, 71(11):2381–2425, 2018.

Mu, C., Huang, B., Wright, J., and Goldfarb, D. Square deal: Lower bounds and improved relaxations for tensor recovery. In *Proceedings of the International Conference on Machine Learning*, pp. 73–81, 2014.

Netrapalli, P., Niranjan, U., Sanghavi, S., Anandkumar, A., and Jain, P. Non-convex robust PCA. In *Advances in Neural Information Processing Systems*, pp. 1107–1115, 2014.

Ning, Y. and Liu, H. A general theory of hypothesis tests and confidence regions for sparse high dimensional models. *The Annals of Statistics*, 45(1):158–195, 2017.

Potechin, A. and Steurer, D. Exact tensor completion with sum-of-squares. In *Proceedings of the Conference on Learning Theory*, pp. 1619–1673, 2017.

Qu, Q., Zhang, Y., Eldar, Y. C., and Wright, J. Convolutional phase retrieval via gradient descent. *IEEE Transactions on Information Theory*, 2019.

Ren, Z., Sun, T., Zhang, C.-H., and Zhou, H. H. Asymptotic normality and optimalities in estimation of large Gaussian graphical models. *The Annals of Statistics*, 43(3):991–1026, 2015.

Romera-Paredes, B. and Pontil, M. A new convex relaxation for tensor completion. In *Advances in Neural Information Processing Systems*, pp. 2967–2975, 2013.

Semerci, O., Hao, N., Kilmer, M. E., and Miller, E. L. Tensor-based formulation and nuclear norm regularization for multienergy computed tomography. *IEEE Transactions on Image Processing*, 23(4):1678–1693, 2014.

Sidiropoulos, N. D., De Lathauwer, L., Fu, X., Huang, K., Papalexakis, E. E., and Faloutsos, C. Tensor decomposition for signal processing and machine learning. *IEEE Transactions on Signal Processing*, 65(13):3551–3582, 2017.

Srebro, N. *Learning with matrix factorizations*. PhD thesis, Massachusetts Institute of Technology, 2004.

Sun, J., Qu, Q., and Wright, J. A geometric analysis of phase retrieval. *Foundations of Computational Mathematics*, 18 (5):1131–1198, 2018.

Sur, P., Chen, Y., and Candès, E. J. The likelihood ratio test in high-dimensional logistic regression is asymptotically a rescaled chi-square. *Probability Theory and Related Fields*, 175:487–558, 2019.

Tomioka, R., Hayashi, K., and Kashima, H. Estimation of low-rank tensors via convex optimization. *arXiv preprint arXiv:1010.0789*, 2010.

Tong, T., Ma, C., and Chi, Y. Accelerating ill-conditioned low-rank matrix estimation via scaled gradient descent. *arXiv preprint arXiv:2005.08898*, 2020.

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

Wang, G. and Giannakis, G. Solving random systems of quadratic equations via truncated generalized gradient flow. In *Advances in Neural Information Processing Systems*, pp. 568–576, 2016.

Xia, D. Confidence interval of singular vectors for high-dimensional and low-rank matrix regression. *arXiv preprint arXiv:1805.09871*, 2018.

Xia, D. Data-dependent confidence regions of singular subspaces. *arXiv preprint arXiv:1901.00304*, 2019.

Xia, D. and Yuan, M. On polynomial time methods for exact low-rank tensor completion. *Foundations of Computational Mathematics*, pp. 1265–1313, 2019a.

Xia, D. and Yuan, M. Statistical inferences of linear forms for noisy matrix completion. *arXiv preprint arXiv:1909.00116*, 2019b.

Xia, D., Yuan, M., and Zhang, C.-H. Statistically optimal and computationally efficient low rank tensor completion from noisy entries. *The Annals of Statistics,* to appear, 2020.

Yuan, M. and Zhang, C.-H. On tensor completion via nuclear norm minimization. *Foundations of Computational Mathematics*, 16(4):1031–1068, 2016.

Yuan, M. and Zhang, C.-H. Incoherent tensor norms and their applications in higher order tensor completion. *IEEE Transactions on Information Theory*, 63(10):6753–6766, 2017.

Zhang, A. Cross: Efficient low-rank tensor completion. *The Annals of Statistics*, 47(2):936–964, 2019.

Zhang, A., Cai, T. T., and Wu, Y. Heteroskedastic PCA: Algorithm, optimality, and applications. *arXiv preprint arXiv:1810.08316*, 2018.

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

Zhang, H., Chi, Y., and Liang, Y. Provable non-convex phase retrieval with outliers: Median truncated Wirtinger flow. In *Proceedings of the International Conference on Machine Learning*, pp. 1022–1031, 2016.

Zhang, H., Zhou, Y., Liang, Y., and Chi, Y. A nonconvex approach for phase retrieval: Reshaped wirtinger flow and incremental algorithms. *The Journal of Machine Learning Research*, 18(1):5164–5198, 2017.

Zhang, Z. and Aeron, S. Exact tensor completion using t-svd. *IEEE Transactions on Signal Processing*, 65(6): 1511–1526, 2017.