Convex Tensor Decomposition via Structured Schatten Norm Regularization

Ryota Tomioka
Toyota Technological Institute at Chicago
Chicago, IL 60637
tomioka@ttic.edu

Taiji Suzuki
Department of Mathematical
and Computing Sciences
Tokyo Institute of Technology
Tokyo 152-8552, Japan
s-taiji@is.titech.ac.jp

Abstract

We study a new class of structured Schatten norms for tensors that includes two recently proposed norms ("overlapped" and "latent") for convex-optimization-based tensor decomposition. We analyze the performance of "latent" approach for tensor decomposition, which was empirically found to perform better than the "overlapped" approach in some settings. We show theoretically that this is indeed the case. In particular, when the unknown true tensor is low-rank in a specific unknown mode, this approach performs as well as knowing the mode with the smallest rank. Along the way, we show a novel duality result for structured Schatten norms, which is also interesting in the general context of structured sparsity. We confirm through numerical simulations that our theory can precisely predict the scaling behaviour of the mean squared error.

1 Introduction

Decomposition of tensors [10, 14] (or multi-way arrays) into low-rank components arises naturally in many real world data analysis problems. For example, in neuroimaging, spatio-temporal patterns of neural activities that are related to certain experimental conditions or subjects can be found by computing the tensor decomposition of the data tensor, which can be of size channels $\times$ time-points $\times$ subjects $\times$ conditions [18]. More generally, any multivariate spatio-temporal data (e.g., environmental monitoring) can be regarded as a tensor. If some of the observations are missing, low-rank modeling enables the imputation of missing values. Tensor modelling may also be valuable for collaborative filtering with temporal or contextual dimension.

Conventionally, tensor decomposition has been tackled through non-convex optimization problems, using alternate least squares or higher-order orthogonal iteration [6]. Compared to its empirical success, little has been theoretically understood about the performance of tensor decomposition algorithms. De Lathauwer et al. [5] showed an approximation bound for a truncated higher-order SVD (also known as the Tucker decomposition). Nevertheless the generalization performance of these approaches has been widely open. Moreover, the model selection problem can be highly challenging, especially for the Tucker model [5, 27], because we need to specify the rank r_k for each mode (here a mode refers to one dimensionality of a tensor); that is, we have K hyper-parameters to choose for a K-way tensor, which is challenging even for $K = 3$.

Recently a convex-optimization-based approach for tensor decomposition has been proposed by several authors [9, 15, 23, 25], and its performance has been analyzed in [26].

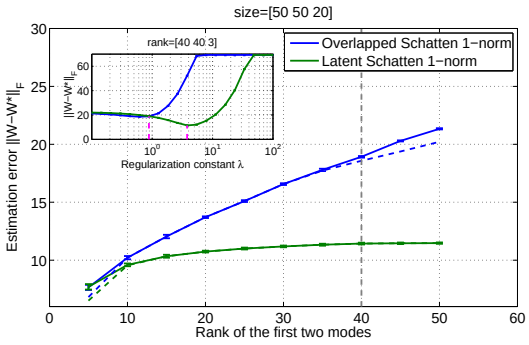

Figure 1: Estimation of a low-rank $50\times50\times20$ tensor of rank $r \times r \times 3$ from noisy measurements. The noise standard deviation is $\sigma = 0.1$. The estimation errors of two convex optimization based methods are plotted against the rank r of the first two modes. The solid lines show the error at the fixed regularization constant λ, which is 0.89 for the overlapped approach and 3.79 for the latent approach (see also Figure 2). The dashed lines show the minimum error over candidates of the regularization constant λ from 0.1 to 100. In the inset, the errors of the two approaches are plotted against the regularization constant λ for rank $r = 40$ (marked with gray dashed vertical line in the outset). The two values (0.89 and 3.79) are marked with vertical dashed lines. Note that both approaches need no knowledge of the true rank; the rank is automatically learned.

The basic idea behind their convex approach, which we call *overlapped approach*, is to unfold[1] a tensor into matrices along different modes and penalize the unfolded matrices to be *simultaneously low-rank* based on the Schatten 1-norm, which is also known as the trace norm and nuclear norm [7, 22, 24]. This approach does not require the rank of the decomposition to be specified beforehand, and due to the low-rank inducing property of the Schatten 1-norm, the rank of the decomposition is *automatically* determined.

However, it has been noticed that the above overlapped approach has a limitation that it performs poorly for a tensor that is only low-rank in a certain mode. The authors of [25] proposed an alternative approach, which we call *latent approach*, that decomposes a given tensor into a mixture of tensors that each are low-rank in a specific mode. Figure 1 demonstrates that the latent approach is preferable to the overlapped approach when the underlying tensor is almost full rank in all but one mode. However, so far no theoretical analysis has been presented to support such an empirical success.

In this paper, we rigorously study the performance of the latent approach and show that the mean squared error of the latent approach scales no greater than the minimum mode-k rank of the underlying true tensor, which clearly explains why the latent approach performs better than the overlapped approach in Figure 1.

Along the way, we show a novel *duality* between the two types of norms employed in the above two approaches, namely the overlapped Schatten norm and the latent Schatten norm. This result is closely related and generalize the results in structured sparsity literature [2, 13, 17, 21]. In fact, the *(plain) overlapped group lasso* constrains the weights to be simultaneously group sparse over overlapping groups. The *latent group lasso* predicts with a mixture of group sparse weights [see also 1, 3, 12]. These approaches clearly correspond to the two variations of tensor decomposition algorithms we discussed above.

Finally we empirically compare the overlapped approach and latent approach and show that even when the unknown tensor is simultaneously low-rank, which is a favorable situation for the overlapped approach, the latent approach performs better in many cases. Thus we provide both theoretical and empirical evidence that for noisy tensor decomposition, the latent approach is preferable to the overlapped approach. Our result is complementary to the previous study [25, 26], which mainly focused on the noise-less tensor completion setting.

[1]For a K-way tensor, there are K ways to unfold a tensor into a matrix. See Section 2.

This paper is structured as follows. In Section 2, we provide basic definitions of the two variations of structured Schatten norms, namely the overlapped/latent Schatten norms, and discuss their properties, especially the *duality* between them. Section 3 presents our main theoretical contributions; we establish the consistency of the latent approach, and we analyze the denoising performance of the latent approach. In Section 4, we empirically confirm the scaling predicted by our theory. Finally, Section 5 concludes the paper. Most of the proofs are presented in the supplementary material.

2 Structured Schatten norms for tensors

In this section, we define the overlapped Schatten norm and the latent Schatten norm and discuss their basic properties.

First we need some basic definitions.

Let $\mathcal{W} \in \mathbb{R}^{n_1 \times \cdots n_K}$ be a K-way tensor. We denote the total number of entries in $\mathcal{W}$ by $N = \prod_{k=1}^{K} n_k$. The dot product between two tensors $\mathcal{W}$ and $\mathcal{X}$ is defined as $\langle \mathcal{W}, \mathcal{X} \rangle = \mathrm{vec}(\mathcal{W})^\top \mathrm{vec}(\mathcal{X})$; i.e., the dot product as vectors in $\mathbb{R}^N$. The Frobenius norm of a tensor is defined as $\left|\!\left|\!\left| \mathcal{W} \right|\!\right|\!\right|_F = \sqrt{\langle \mathcal{W}, \mathcal{W} \rangle}$. Each dimensionality of a tensor is called a *mode*. The mode k *unfolding* $\boldsymbol{W}_{(k)} \in \mathbb{R}^{n_k \times N/n_k}$ is a matrix that is obtained by concatenating the mode-k fibers along columns; here a mode-k fiber is an n_k dimensional vector obtained by fixing all the indices but the kth index of $\mathcal{W}$. The mode-k rank $\underline{r}_k$ of $\mathcal{W}$ is the rank of the mode-k unfolding $\boldsymbol{W}_{(k)}$. We say that a tensor $\mathcal{W}$ has multilinear rank $(\underline{r}_1, \ldots, \underline{r}_K)$ if the mode-k rank is $\underline{r}_k$ for $k = 1, \ldots, K$ [14]. The mode k folding is the inverse of the unfolding operation.

2.1 Overlapped Schatten norms

The low-rank inducing norm studied in [9, 15, 23, 25], which we call overlapped Schatten 1-norm, can be written as follows:

$$\left|\!\left|\!\left| \mathcal{W} \right|\!\right|\!\right|_{\underline{S_1/1}} = \sum\nolimits_{k=1}^{K} \|\boldsymbol{W}_{(k)}\|_{S_1}. \tag{1}$$

In this paper, we consider the following more general *overlapped S_p/q-norm*, which includes the Schatten 1-norm as the special case $(p, q) = (1, 1)$. The overlapped S_p/q-norm is written as follows:

$$\left|\!\left|\!\left| \mathcal{W} \right|\!\right|\!\right|_{\underline{S_p/q}} = \Big(\sum\nolimits_{k=1}^{K} \|\boldsymbol{W}_{(k)}\|_{S_p}^{q} \Big)^{1/q}, \tag{2}$$

where $1 \leq p, q \leq \infty$; here

$$\|\boldsymbol{W}\|_{S_p} = \Big(\sum\nolimits_{j=1}^{r} \sigma_j^p(\boldsymbol{W}) \Big)^{1/p}$$

is the Schatten p-norm for matrices, where $\sigma_j(\boldsymbol{W})$ is the jth largest singular value of $\boldsymbol{W}$.

When used as a regularizer, the overlapped Schatten 1-norm penalizes all modes of $\mathcal{W}$ to be jointly low-rank. It is related to the overlapped group regularization [see 13, 16] in a sense that the same object $\mathcal{W}$ appears repeatedly in the norm.

The following inequality relates the overlapped Schatten 1-norm with the Frobenius norm, which was a key step in the analysis of [26]:

$$\left|\!\left|\!\left| \mathcal{W} \right|\!\right|\!\right|_{\underline{S_1/1}} \leq \sum_{k=1}^{K} \sqrt{\underline{r}_k} \left|\!\left|\!\left| \mathcal{W} \right|\!\right|\!\right|_F, \tag{3}$$

where $\underline{r}_k$ is the mode-k rank of $\mathcal{W}$.

Now we are interested in the dual norm of the overlapped S_p/q-norm, because deriving the dual norm is a key step in solving the minimization problem that involves the norm (2) [see 16], as well as computing various complexity measures, such as, Rademacher complexity [8] and Gaussian width [4]. It turns out that the dual norm of the overlapped S_p/q-norm is the *latent S_{p^*}/q^*-norm* as shown in the following lemma (proof is presented in Appendix A).

Lemma 1. *The dual norm of the overlapped S_p/q-norm is the latent S_{p^*}/q^*-norm, where $1/p + 1/p^* = 1$ and $1/q + 1/q^* = 1$, which is defined as follows:*

$$\left|\!\left|\!\left| \mathcal{X} \right|\!\right|\!\right|_{\overline{S_{p^*}/q^*}} = \inf_{\left(\mathcal{X}^{(1)}+\cdots+\mathcal{X}^{(K)}\right)=\mathcal{X}} \left(\sum\nolimits_{k=1}^{K} \|\boldsymbol{X}_{(k)}^{(k)}\|_{S_{p^*}}^{q^*} \right)^{1/q^*}. \tag{4}$$

Here the infimum is taken over the K-tuple of tensors $\mathcal{X}^{(1)}, \ldots, \mathcal{X}^{(K)}$ that sums to $\mathcal{X}$.

In the supplementary material, we show a slightly more general version of the above lemma that naturally generalizes the duality between overlapped/latent group sparsity norms [1, 12, 17, 21]; see Section A. Note that when the groups have no overlap, the overlapped/latent group sparsity norms become identical, and the duality is the ordinary duality between the group S_p/q-norms and the group S_{p^*}/q^*-norms.

2.2 Latent Schatten norms

The latent approach for tensor decomposition [25] solves the following minimization problem

$$\underset{\mathcal{W}^{(1)},\ldots,\mathcal{W}^{(K)}}{\text{minimize}} \quad L(\mathcal{W}^{(1)} + \cdots + \mathcal{W}^{(K)}) + \lambda \sum_{k=1}^{K} \|\boldsymbol{W}_{(k)}^{(k)}\|_{S_1}, \tag{5}$$

where L is a loss function, λ is a regularization constant, and $\boldsymbol{W}_{(k)}^{(k)}$ is the mode-k unfolding of $\mathcal{W}^{(k)}$. Intuitively speaking, the latent approach for tensor decomposition predicts with a mixture of K tensors that each are regularized to be low-rank in a specific mode.

Now, since the loss term in the minimization problem (5) only depends on the sum of the tensors $\mathcal{W}^{(1)}, \ldots, \mathcal{W}^{(K)}$, minimization problem (5) is equivalent to the following minimization problem

$$\underset{\mathcal{W}}{\text{minimize}} \quad L(\mathcal{W}) + \lambda \left|\!\left|\!\left| \mathcal{W} \right|\!\right|\!\right|_{\overline{S_1/1}}.$$

In other words, we have identified the structured Schatten norm employed in the latent approach as the latent $S_1/1$-norm (or latent Schatten 1-norm for short), which can be written as follows:

$$\left|\!\left|\!\left| \mathcal{W} \right|\!\right|\!\right|_{\overline{S_1/1}} = \inf_{\left(\mathcal{W}^{(1)}+\cdots+\mathcal{W}^{(K)}\right)=\mathcal{W}} \sum_{k=1}^{K} \|\boldsymbol{W}_{(k)}^{(k)}\|_{S_1}. \tag{6}$$

According to Lemma 1, the dual norm of the latent $S_1/1$-norm is the overlapped S_∞/∞-norm

$$\left|\!\left|\!\left| \mathcal{X} \right|\!\right|\!\right|_{\underline{S_\infty/\infty}} = \max_k \|\boldsymbol{X}_{(k)}\|_{S_\infty}, \tag{7}$$

where $\|\cdot\|_{S_\infty}$ is the spectral norm.

The following lemma is similar to inequality (3) and is a key in our analysis (proof is presented in Appendix B).

Lemma 2.

$$\left|\!\left|\!\left| \mathcal{W} \right|\!\right|\!\right|_{\overline{S_1/1}} \leq \left(\min_k \sqrt{\underline{r}_k} \right) \left|\!\left|\!\left| \mathcal{W} \right|\!\right|\!\right|_F,$$

where $\underline{r}_k$ is the mode-k rank of $\mathcal{W}$.

Compared to inequality (3), the latent Schatten 1-norm is bounded by the *minimal* square root of the ranks instead of the sum. This is the fundamental reason why the latent approach performs betters than the overlapped approach as in Figure 1.

3 Main theoretical results

In this section, combining the duality we presented in the previous section with the techniques from Agarwal et al. [1], we study the generalization performance of the latent approach for tensor decomposition in the context of recovering an unknown tensor $\mathcal{W}^*$ from noisy measurements. This is the setting of the experiment in Figure 1. We first prove a generic consistency statement that does not take the low-rank-ness of the truth into account. Next we show that a tighter bound that takes the low-rank-ness into account can be obtained with some incoherence assumption. Finally, we discuss the difference between overlapped approach and latent approach and provide an explanation for the empirically observed superior performance of the latent approach in Figure 1.

3.1 Consistency

Let $\mathcal{W}^*$ be the underlying true tensor and the noisy version $\mathcal{Y}$ is obtained as follows:

$$\mathcal{Y} = \mathcal{W}^* + \mathcal{E},$$

where $\mathcal{E} \in \mathbb{R}^{n_1 \times \cdots \times n_K}$ is the noise tensor.

A consistency statement can be obtained as follows (proof is presented in Appendix C):

Theorem 1. *Assume that the regularization constant λ satisfies $\lambda \geq \left|\!\left|\!\left| \mathcal{E} \right|\!\right|\!\right|_{\underline{S_\infty/\infty}}$ (overlapped S_∞/∞ norm of the noise), then the estimator defined by $\hat{\mathcal{W}} = \operatorname{argmin}_{\mathcal{W}} \left(\frac{1}{2} \left|\!\left|\!\left| \mathcal{Y} - \mathcal{W} \right|\!\right|\!\right|_F^2 + \lambda \left|\!\left|\!\left| \mathcal{W} \right|\!\right|\!\right|_{\overline{S_1/1}} \right)$, satisfies the inequality*

$$\left|\!\left|\!\left| \hat{\mathcal{W}} - \mathcal{W}^* \right|\!\right|\!\right|_F \leq 2\lambda \sqrt{\min_k n_k}. \tag{8}$$

In particular when the noise goes to zero $\mathcal{E} \to 0$, the right hand side of inequality (8) shrinks to zero.

3.2 Deterministic bound

The consistency statement in the previous section only deals with the sum $\hat{\mathcal{W}} = \sum_{k=1}^K \hat{\mathcal{W}}^{(k)}$ and the statement does not take into account the low-rank-ness of the truth. In this section, we establish a tighter statement that bounds the errors of individual terms $\hat{\mathcal{W}}^{(k)}$.

To this end, we need some additional assumptions. First, we assume that the unknown tensor $\mathcal{W}^*$ is a mixture of K tensors that each are low-rank in a certain mode and we have a noisy observation $\mathcal{Y}$ as follows:

$$\mathcal{Y} = \mathcal{W}^* + \mathcal{E} = \sum\nolimits_{k=1}^K \mathcal{W}^{*(k)} + \mathcal{E}, \tag{9}$$

where $\bar{r}_k = \operatorname{rank}(\boldsymbol{W}^{(k)}_{(k)})$ is the mode-k rank of the kth component $\mathcal{W}^{*(k)}$; note that this does not equal the mode-k rank $\underline{r}_k$ of $\mathcal{W}^*$ in general.

Second, we assume that the spectral norm of the mode-k unfolding of the lth component is bounded by a constant α for all $k \neq l$ as follows:

$$\|\boldsymbol{W}^{*(l)}_{(k)}\|_{S_\infty} \leq \alpha \quad (\forall l \neq k, k, l = 1, \ldots, K). \tag{10}$$

Note that such an additional incoherence assumption has also been used in [1, 3, 11].

We employ the following optimization problem to recover the unknown tensor $\mathcal{W}^*$:

$$\hat{\mathcal{W}} = \operatorname*{argmin}_{\mathcal{W}} \left(\frac{1}{2} \left|\!\left|\!\left| \mathcal{Y} - \mathcal{W} \right|\!\right|\!\right|_F^2 + \lambda \left|\!\left|\!\left| \mathcal{W} \right|\!\right|\!\right|_{\overline{S_1/1}} \quad \text{s.t. } \mathcal{W} = \sum_{k=1}^K \mathcal{W}^{(k)}, \ \|\boldsymbol{W}^{(l)}_{(k)}\|_{S_\infty} \leq \alpha, \quad \forall l \neq k \right), \tag{11}$$

where $\lambda > 0$ is a regularization constant. Notice that we have introduced additional spectral norm constraints to control the correlation between the components; see also [1].

Our deterministic performance bound can be stated as follows (proof is presented in Appendix D):

Theorem 2. *Let $\hat{\mathcal{W}}^{(k)}$ be an optimal decomposition of $\hat{\mathcal{W}}$ induced by the latent Schatten 1-norm* (6)*. Assume that the regularization constant λ satisfies $\lambda \geq 2 \left|\!\left|\!\left| \mathcal{E} \right|\!\right|\!\right|_{\underline{S_\infty/\infty}} + \alpha(K-1)$. Then there is a universal constant c such that, any solution $\hat{\mathcal{W}}$ of the minimization problem* (11) *satisfies the following deterministic bound:*

$$\sum\nolimits_{k=1}^K \left|\!\left|\!\left| \hat{\mathcal{W}}^{(k)} - \mathcal{W}^{*(k)} \right|\!\right|\!\right|_F^2 \leq c\lambda^2 \sum\nolimits_{k=1}^K \bar{r}_k. \tag{12}$$

Moreover, the overall error can be bounded in terms of the multilinear rank of $\mathcal{W}^*$ as follows:

$$\left|\!\left|\!\left| \hat{\mathcal{W}} - \mathcal{W}^* \right|\!\right|\!\right|_F^2 \leq c\lambda^2 \min_k \underline{r}_k. \tag{13}$$

Note that in order to get inequality (13), we exploit the arbitrariness of the decomposition $\mathcal{W}^* = \sum_{k=1}^K \mathcal{W}^{*(k)}$ to replace the sum over the ranks with the minimal *mode-k rank*. This is possible because a *singleton decomposition*, i.e., $\mathcal{W}^{*(k)} = \mathcal{W}^*$ and $\mathcal{W}^{*(k')} = 0$ for $k' \neq k$, is allowed for any k.

Comparing two inequalities (8) and (13), we see that there are two regimes. When the noise is small, (8) is tighter. On the other hand, when the noise is larger and/or $\min_k \underline{r}_k \ll \min_k n_k$, (13) is tighter.

3.3 Gaussian noise

When the elements of the noise tensor $\mathcal{E}$ are Gaussian, we obtain the following theorem.

Theorem 3. *Assume that the elements of the noise tensor $\mathcal{E}$ are independent zero-mean Gaussian random variables with variance σ^2. In addition, assume without loss of generality that the dimensionalities of $\mathcal{W}^*$ are sorted in the descending order, i.e., $n_1 \geq \cdots \geq n_K$. Then there is a universal constant c such that, with probability at least $1-\delta$, any solution of the minimization problem* (11) *with regularization constant $\lambda = 2\sigma(\sqrt{N/n_K} + \sqrt{n_1} + \sqrt{2\log(K/\delta)}) + \alpha(K-1)$ satisfies*

$$\frac{1}{N}\sum_{k=1}^K \left|\!\left|\!\left| \hat{\mathcal{W}}^{(k)} - \mathcal{W}^{*(k)} \right|\!\right|\!\right|_F^2 \leq cF\sigma^2 \frac{\sum_{k=1}^K \bar{r}_k}{n_K}, \tag{14}$$

where $F = \left(\left(1+\sqrt{\frac{n_1 n_K}{N}}\right) + \left(\sqrt{2\log(K/\delta)} + \frac{\alpha(K-1)}{2\sigma}\right)\sqrt{\frac{n_K}{N}}\right)^2$ is a factor that mildly depends on the dimensionalities and the constant α in (10)*.*

Note that the theoretically optimal choice of regularization constant λ is independent of the ranks of the truth $\mathcal{W}^*$ or its factors in (9), which are unknown in practice.

Again we can obtain a bound corresponding to the minimum rank singleton decomposition as in inequality (13) as follows:

$$\frac{1}{N}\left|\!\left|\!\left| \hat{\mathcal{W}} - \mathcal{W}^* \right|\!\right|\!\right|_F^2 \leq cF\sigma^2 \frac{\min_k \underline{r}_k}{n_K}, \tag{15}$$

where F is the same factor as in Theorem 3.

3.4 Comparison with the overlapped approach

Inequality (15) explains the superior performance of the latent approach for tensor decomposition in Figure 1. The inequality obtained in [26] for the overlapped approach that uses overlapped Schatten 1-norm (1) can be stated as follows:

$$\frac{1}{N}\left|\!\left|\!\left| \hat{\mathcal{W}} - \mathcal{W}^* \right|\!\right|\!\right|_F^2 \leq c'\sigma^2 \left(\frac{1}{K}\sum_{k=1}^K \sqrt{\tfrac{1}{n_k}}\right)^2 \left(\frac{1}{K}\sum_{k=1}^K \sqrt{\underline{r}_k}\right)^2. \tag{16}$$

Comparing inequalities (15) and (16), we notice that the complexity of the overlapped approach depends on the average (square root) of the mode-k ranks $\underline{r}_1, \ldots, \underline{r}_K$, whereas that of the latent approach only grows linearly against the *minimum* mode-k rank. Interestingly, the latent approach performs *as if it knows the mode with the minimum rank*, although such information is not given.

Recently, Mu et al. [19] proved a lower bound of the number of measurements for solving linear inverse problem via the overlapped approach. Although the setting is different, the lower bound depends on the minimum mode-k rank, which agrees with the complexity of the latent approach.

4 Numerical results

In this section, we numerically confirm the theoretically obtained scaling behavior.

The goal of this experiment is to recover the true low rank tensor $\mathcal{W}^*$ from a noisy observation $\mathcal{Y}$. We randomly generated the true low rank tensors $\mathcal{W}^*$ of size $50 \times 50 \times 20$ or $80 \times 80 \times 40$ with various mode-k ranks $(\underline{r}_1, \underline{r}_2, \underline{r}_3)$. A low-rank tensor is generated by first randomly drawing the

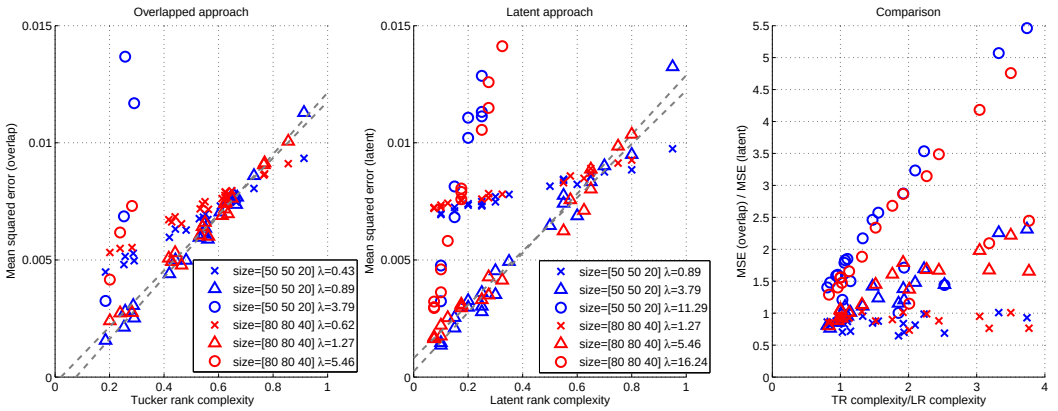

Figure 2: Performance of the overlapped approach and latent approach for tensor decomposition are shown against their theoretically predicted complexity measures (see Eqs. (17) and (18)). The right panel shows the improvement of the latent approach from the overlapped approach against the ratio of their complexity measures.

$\underline{r}_1 \times \underline{r}_2 \times \underline{r}_3$ core tensor from the standard normal distribution and multiplying an orthogonal factor matrix drawn uniformly to its each mode. The observation tensor $\mathcal{Y}$ is obtained by adding Gaussian noise with standard deviation $\sigma = 0.1$. There is no missing entries in this experiment.

For each observation $\mathcal{Y}$, we computed tensor decompositions using the overlapped approach and the latent approach (11). For the optimization, we used the algorithms[2] based on alternating direction method of multipliers described in Tomioka et al. [25]. We computed the solutions for 20 candidate regularization constants ranging from 0.1 to 100 and report the results for three representative values for each method.

We measured the quality of the solutions obtained by the two approaches by the mean squared error (MSE) $\left\|\left\|\hat{\mathcal{W}} - \mathcal{W}^*\right\|\right\|_F^2 / N$. In order to make our theoretical predictions more concrete, we define the quantities in the right hand side of the bounds (16) and (14) as *Tucker rank (TR) complexity* and *Latent rank (LR) complexity*, respectively, as follows:

$$\text{TR complexity} = \left(\frac{1}{K}\sum_{k=1}^K \sqrt{\frac{1}{n_k}}\right)^2 \left(\frac{1}{K}\sum_{k=1}^K \sqrt{\underline{r}_k}\right)^2, \tag{17}$$

$$\text{LR complexity} = \frac{\sum_{k=1}^K \bar{r}_k}{n_K}, \tag{18}$$

where without loss of generality we assume $n_1 \geq \cdots \geq n_K$. We have ignored terms like $\sqrt{n_k/N}$ because they are negligible for $n_k \approx 50$ and $N \approx 50,000$. The TR complexity is equivalent to the *normalized rank* in [26]. Note that the TR complexity (17) is defined in terms of the multilinear rank $(\underline{r}_1, \ldots, \underline{r}_K)$ of the truth $\mathcal{W}^*$, whereas the LR complexity (18) is defined in terms of the ranks of the latent factors $(\bar{r}_1, \ldots, \bar{r}_K)$ in (9). In order to find a decomposition that minimizes the right hand side of (18), we ran the latent approach to the true tensor $\mathcal{W}^*$ without noise, and took the minimum of the sum of ranks found by the run and $\min_k \underline{r}_k$, i.e., the minimal mode-k rank (because a singleton solution is also allowed). The whole procedure is repeated 10 times and averaged.

Figure 2 shows the results of the experiment. The left panel shows the MSE of the overlapped approach against the TR complexity (17). The middle panel shows the MSE of the latent approach against the LR complexity (18). The right panel shows the improvement (i.e., MSE of the overlap approach over that of the latent approach) against the ratio of the respective complexity measures.

First, from the left panel, we can confirm that as predicted by [26], the MSE of the overlapped approach scales linearly against the TR complexity (17) for each value of the regularization constant.

From the central panel, we can clearly see that the MSE of the latent approach scales linearly against the LR complexity (18) as predicted by Theorem 3. The series with $\triangle$ ($\lambda = 3.79$ for $50 \times 50 \times 20$,

[2]The solver is available online: `https://github.com/ryotat/tensor`.

$\lambda = 5.46$ for $80 \times 80 \times 40$) is mostly below other series, which means that the optimal choice of the regularization constant is independent of the rank of the true tensor and only depends on the size; this agrees with the condition on λ in Theorem 3. Since the blue series and red series with the same markers lie on top of each other (especially the series with $\triangle$ for which the optimal regularization constant is chosen), we can see that our theory predicts not only the scaling against the latent ranks but also that against the size of the tensor correctly. Note that the regularization constants are scaled by roughly 1.6 to account for the difference in the dimensionality.

The right panel reveals that in many cases the latent approach performs better than the overlapped approach, i.e., MSE (overlap)/ MSE (latent) greater than one. Moreover, we can see that the success of the latent approach relative to the overlapped approach is correlated with high TR complexity to LR complexity ratio. Indeed, we found that an optimal decomposition of the true tensor $\mathcal{W}^*$ was typically a singleton decomposition corresponding to the smallest tucker rank (see Section 3.2). Note that the two approaches perform almost identically when they are under-regularized (crosses).

The improvements here are milder than that in Figure 1. This is because most of the randomly generated low-rank tensors were simultaneously low-rank to some degree. It is encouraging that the latent approach perform at least as well as the overlapped approach in such situations as well.

5 Conclusion

In this paper, we have presented a framework for structured Schatten norms. The current framework includes both the overlapped Schatten 1-norm and latent Schatten 1-norm recently proposed in the context of convex-optimization-based tensor decomposition [9, 15, 23, 25], and connects these studies to the broader studies on structured sparsity [2, 13, 17, 21]. Moreover, we have shown a *duality* that holds between the two types of norms.

Furthermore, we have rigorously studied the performance of the latent approach for tensor decomposition. We have shown the consistency of the latent Schatten 1-norm minimization. Next, we have analyzed the denoising performance of the latent approach and shown that the error of the latent approach is upper bounded by the *minimal* mode-k rank, which contrasts sharply against the average (square root) dependency of the overlapped approach analyzed in [26]. This explains the empirically observed superior performance of the latent approach compared to the overlapped approach. The most difficult case for the overlapped approach is when the unknown tensor is only low-rank in one mode as in Figure 1.

We have also confirmed through numerical simulations that our analysis precisely predicts the scaling of the mean squared error as a function of the dimensionalities and the sum of ranks of the factors of the unknown tensor, which is dominated by the minimal mode-k rank. Unlike mode-k ranks, the ranks of the factors are not easy to compute. However, note that the theoretically optimal choice of the regularization constant does not depend on these quantities.

Thus, we have theoretically and empirically shown that for noisy tensor decomposition, the latent approach is more likely to perform better than the overlapped approach. Analyzing the performance of the latent approach for tensor completion would be an important future work.

The structured Schatten norms proposed in this paper include norms for tensors that are not employed in practice yet. Therefore, it would be interesting to explore various extensions, such as, using the overlapped S_1/∞-norm instead of the $S_1/1$-norm or a *non-sparse* tensor decomposition.

Acknowledgment: This work was carried out while both authors were at The University of Tokyo. This work was partially supported by JSPS KAKENHI 25870192 and 25730013, and the Aihara Project, the FIRST program from JSPS, initiated by CSTP.

References

[1] A. Agarwal, S. Negahban, and M. J. Wainwright. Noisy matrix decomposition via convex relaxation: Optimal rates in high dimensions. *The Annals of Statistics*, 40(2):1171–1197, 2012.

[2] F. Bach, R. Jenatton, J. Mairal, and G. Obozinski. Convex optimization with sparsity-inducing norms. In *Optimization for Machine Learning*. MIT Press, 2011.

[3] E. J. Candes, X. Li, Y. Ma, and J. Wright. Robust principal component analysis? Technical report, arXiv:0912.3599, 2009.

[4] V. Chandrasekaran, B. Recht, P. Parrilo, and A. Willsky. The convex geometry of linear inverse problems, prepint. Technical report, arXiv:1012.0621v2, 2010.

[5] L. De Lathauwer, B. De Moor, and J. Vandewalle. A multilinear singular value decomposition. *SIAM J. Matrix Anal. Appl.*, 21(4):1253–1278, 2000.

[6] L. De Lathauwer, B. De Moor, and J. Vandewalle. On the best rank-1 and rank-($R_1, R_2, \ldots, R_N$) approximation of higher-order tensors. *SIAM J. Matrix Anal. Appl.*, 21(4):1324–1342, 2000.

[7] M. Fazel, H. Hindi, and S. P. Boyd. A Rank Minimization Heuristic with Application to Minimum Order System Approximation. In *Proc. of the American Control Conference*, 2001.

[8] R. Foygel and N. Srebro. Concentration-based guarantees for low-rank matrix reconstruction. Technical report, arXiv:1102.3923, 2011.

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

[10] F. L. Hitchcock. The expression of a tensor or a polyadic as a sum of products. *J. Math. Phys.*, 6(1): 164–189, 1927.

[11] D. Hsu, S. M. Kakade, and T. Zhang. Robust matrix decomposition with sparse corruptions. *Information Theory, IEEE Transactions on*, 57(11):7221–7234, 2011.

[12] A. Jalali, P. Ravikumar, S. Sanghavi, and C. Ruan. A dirty model for multi-task learning. In *Advances in NIPS 23*, pages 964–972. 2010.

[13] R. Jenatton, J. Audibert, and F. Bach. Structured variable selection with sparsity-inducing norms. *J. Mach. Learn. Res.*, 12:2777–2824, 2011.

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

[15] J. Liu, P. Musialski, P. Wonka, and J. Ye. Tensor completion for estimating missing values in visual data. In *Prof. ICCV*, 2009.

[16] J. Mairal, R. Jenatton, G. Obozinski, and F. Bach. Convex and network flow optimization for structured sparsity. *J. Mach. Learn. Res.*, 12:2681–2720, 2011.

[17] A. Maurer and M. Pontil. Structured sparsity and generalization. Technical report, arXiv:1108.3476, 2011.

[18] M. Mørup. Applications of tensor (multiway array) factorizations and decompositions in data mining. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 1(1):24–40, 2011.

[19] C. Mu, B. Huang, J. Wright, and D. Goldfarb. Square deal: Lower bounds and improved relaxations for tensor recovery. *arXiv preprint arXiv:1307.5870*, 2013.

[20] S. Negahban, P. Ravikumar, M. Wainwright, and B. Yu. A unified framework for high-dimensional analysis of m-estimators with decomposable regularizers. In *Advances in NIPS 22*, pages 1348–1356. 2009.

[21] G. Obozinski, L. Jacob, and J.-P. Vert. Group lasso with overlaps: the latent group lasso approach. Technical report, arXiv:1110.0413, 2011.

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

[23] M. Signoretto, L. De Lathauwer, and J. Suykens. Nuclear norms for tensors and their use for convex multilinear estimation. Technical Report 10-186, ESAT-SISTA, K.U.Leuven, 2010.

[24] N. Srebro and A. Shraibman. Rank, trace-norm and max-norm. In *Proc. of the 18th Annual Conference on Learning Theory (COLT)*, pages 545–560. Springer, 2005.

[25] R. Tomioka, K. Hayashi, and H. Kashima. Estimation of low-rank tensors via convex optimization. Technical report, arXiv:1010.0789, 2011.

[26] R. Tomioka, T. Suzuki, K. Hayashi, and H. Kashima. Statistical performance of convex tensor decomposition. In *Advances in NIPS 24*, pages 972–980. 2011.

[27] L. R. Tucker. Some mathematical notes on three-mode factor analysis. *Psychometrika*, 31(3):279–311, 1966.

[28] R. Vershynin. Introduction to the non-asymptotic analysis of random matrices. Technical report, arXiv:1011.3027, 2010.

Supplementary material for "Convex Tensor Decomposition via Structured Schatten Norms"

A Proof of Lemma 1

Proof. This follows from the following lemma.

Lemma 3. *Let $\|\cdot\|_\star$ be a norm and $\mathbf{\Phi}$ be a linear operator from $\mathbb{R}^N$ to $\mathbb{R}^M$. Assume that the right kernel of $\mathbf{\Phi}$ is empty. Then*

1. $\left|\!\left|\!\left|\mathcal{W}\right|\!\right|\!\right|_{\underline{\star}(\mathbf{\Phi})} := \|\mathbf{\Phi}(\mathcal{W})\|_\star$ *is a norm.*
2. $\left|\!\left|\!\left|\mathcal{W}\right|\!\right|\!\right|_{\overline{\star}(\mathbf{\Phi})} := \inf_{\boldsymbol{z}\in\mathbb{R}^M} \|\boldsymbol{z}\|_\star \quad \text{s.t.} \quad \mathbf{\Phi}^\top(\boldsymbol{z}) = \mathcal{W}$ *is also a norm.*
3. $\left|\!\left|\!\left|\cdot\right|\!\right|\!\right|_{\underline{\star}(\mathbf{\Phi})}$ *and* $\left|\!\left|\!\left|\cdot\right|\!\right|\!\right|_{\overline{\star^*}(\mathbf{\Phi})}$ *are dual to each other, where $\|\cdot\|_{\star^*}$ is the dual norm of $\|\cdot\|_\star$.*

In fact, let $M = KN$ and let $\mathbf{\Phi}(\mathcal{W}) = [\mathrm{vec}(\boldsymbol{W}_{(1)}); ...; \mathrm{vec}(\boldsymbol{W}_{(K)})]$, i.e., the column-wise concatenation of unfoldings of $\mathcal{W}$ vectorized, and define the $\star$-norm $\|\cdot\|_\star$ as follows:

$$\|\boldsymbol{z}\|_\star = \left(\sum\nolimits_{k=1}^K \|\boldsymbol{Z}_{(k)}^{(k)}\|_{S_p}^q\right)^{1/q}, \tag{19}$$

where $\boldsymbol{Z}_{(k)}^{(k)}$ denotes the inverse vectorization of an N dimensional sub-vector $\boldsymbol{z}_{(k-1)N+1:kN}$ of $\boldsymbol{z}$ into an $n_k \times N/n_k$ matrix. Now the dual norm of $\star$-norm (19) can be written as follows:

$$\|\boldsymbol{z}\|_{\star^*} = \left(\sum\nolimits_{k=1}^K \|\boldsymbol{Z}_{(k)}^{(k)}\|_{S_{p^*}}^{q^*}\right)^{1/q^*}.$$

Furthermore, the constraint $\mathbf{\Phi}^\top(\boldsymbol{z}) = \mathcal{X}$ can be easily rewritten as follows:

$$\sum\nolimits_{k=1}^K \mathcal{Z}^{(k)} = \mathcal{X},$$

where $\mathcal{Z}^{(k)}$ is the mode-k folding (inverse unfolding) of $\boldsymbol{Z}_{(k)}^{(k)}$. By replacing $\mathcal{Z}^{(k)}$ by $\mathcal{X}^{(k)}$ for $k = 1, \ldots, K$, we have Lemma 1.

Now we prove Lemma 3. The first and the second parts are straightforward to show. In order to prove the third part, let's start from the definition of a dual norm

$$\|\mathcal{X}\|_{\underline{\star}(\mathbf{\Phi})^*} = \sup \langle \mathcal{W}, \mathcal{X}\rangle \quad \text{s.t.} \quad \left|\!\left|\!\left|\mathcal{W}\right|\!\right|\!\right|_{\underline{\star}(\mathbf{\Phi})} \le 1. \tag{20}$$

This is a constrained maximization problem. Since the above maximization problem satisfies Slater's condition, the strong duality holds. Thus, we only need to show that its dual problem agrees with the definition of $\left|\!\left|\!\left|\cdot\right|\!\right|\!\right|_{\overline{\star^*}(\mathbf{\Phi})}$. Due to Fenchel's duality theorem, we have

$$\sup_{\mathcal{W}} \left(\langle \mathcal{W}, \mathcal{X}\rangle - \delta\left(\|\mathbf{\Phi}(\mathcal{W})\|_\star \le 1\right)\right) = \inf_{\boldsymbol{z}} \left(\delta(-\mathbf{\Phi}^\top(\boldsymbol{z}) + \mathcal{X} = 0) + \|\boldsymbol{z}\|_{\star^*}\right),$$

where $\delta(C)$ is the indicator function of condition C (0 if C is true, and ∞ otherwise). Now we can identify the left-hand side as the definition of dual norm (20) and the right-hand side as $\left|\!\left|\!\left|\mathcal{X}\right|\!\right|\!\right|_{\overline{\star^*}(\mathbf{\Phi})}$.

□

B Proof of Lemma 2

Proof. Since we are allowed to take a singleton decomposition $\mathcal{W}^{(k)} = \mathcal{W}$ and $\mathcal{W}^{(k')} = 0$ $(k' \neq k)$, we have

$$\begin{aligned}
\left|\!\left|\!\left|\mathcal{W}\right|\!\right|\!\right|_{\overline{S_1/1}} &= \inf_{(\mathcal{W}^{(1)}+\cdots+\mathcal{W}^{(K)})=\mathcal{W}} \sum_{k=1}^K \|\boldsymbol{W}_{(k)}^{(k)}\|_{S_1} \\
&\le \|\boldsymbol{W}_{(k)}\|_{S_1} \\
&\le \sqrt{r}_k \|\boldsymbol{W}_{(k)}\|_F \quad (\forall k = 1, \ldots, K)
\end{aligned}$$

Choosing k that minimizes the right hand side, we obtain our claim.

□

C Proof of Theorem 1

Proof. Due to the triangular inequality

$$\left\|\hat{\mathcal{W}} - \mathcal{W}^*\right\|_F \leq \left\|\hat{\mathcal{W}} - \mathcal{Y}\right\|_F + \left\|\mathcal{E}\right\|_F.$$

First, due the optimality of $\hat{\mathcal{W}}$, $\mathcal{Y} - \hat{\mathcal{W}} \in \lambda\partial\left\|\hat{\mathcal{W}}\right\|_{\overline{S_1/1}}$, where $\partial\left\|\hat{\mathcal{W}}\right\|_{\overline{S_1/1}}$ is the subdifferential of the latent $S_1/1$ norm at $\hat{\mathcal{W}}$. Thus, the first term of the right hand side satisfies

$$\left\|\hat{\mathcal{W}} - \mathcal{Y}\right\|_F \leq \sqrt{\min_k n_k}\left\|\hat{\mathcal{W}} - \mathcal{Y}\right\|_{\underline{S_\infty/\infty}} \leq \sqrt{\min_k n_k}\lambda,$$

The first inequality is true because $\left\|\mathcal{W}\right\|_F \leq \sqrt{n_k}\|\boldsymbol{W}_{(k)}\|_{S_\infty}$ for any tensor $\mathcal{W}$ and mode k. The last inequality is true because for any $\mathcal{G} \in \partial\left\|\hat{\mathcal{W}}\right\|_{\overline{S_1/1}}$, we have $\left\|\mathcal{G}\right\|_{\underline{S_\infty/\infty}} \leq 1$.

Similarly,

$$\left\|\mathcal{E}\right\|_F \leq \sqrt{\min_k n_k}\left\|\mathcal{E}\right\|_{\underline{S_\infty/\infty}} \leq \sqrt{\min_k n_k}\lambda.$$

The last inequality follows from the assumption. This completes the proof of Theorem 1. □

D Proof of Theorem 2

Let $\hat{\mathcal{W}} = \sum_{k=1}^K \hat{\mathcal{W}}^{(k)}$ be the solution and its optimal decomposition of the minimization problem (11); in addition let $\Delta^{(k)} := \hat{\mathcal{W}}^{(k)} - \mathcal{W}^{*(k)}$.

The proof is based on Lemmas 4 and 5, which we present below.

In order to present the first lemma, we need the following definitions. Let $\boldsymbol{U}_k\boldsymbol{S}_k\boldsymbol{V}_k = \boldsymbol{W}^{*(k)}_{(k)}$ be the singular value decomposition of the mode-k unfolding of the kth component of the unknown tensor $\mathcal{W}^*$. We define the orthogonal projection of $\Delta^{(k)}$ as follows:

$$\boldsymbol{\Delta}^{(k)}_{(k)} = \boldsymbol{\Delta}'_k + \boldsymbol{\Delta}''_k,$$

where

$$\boldsymbol{\Delta}''_k = (\boldsymbol{I}_{n_k} - \boldsymbol{U}_k{\boldsymbol{U}_k}^\top)\boldsymbol{\Delta}^{(k)}_{(k)}(\boldsymbol{I}_{N/n_k} - \boldsymbol{V}_k{\boldsymbol{V}_k}^\top).$$

Intuitively speaking, $\boldsymbol{\Delta}''_k$ lies in a subspace completely orthogonal to the unfolding of the kth component $\boldsymbol{W}^{*(k)}_{(k)}$, whereas $\boldsymbol{\Delta}'_k$ lies in a partially correlated subspace.

The following lemma is similar to Negahban et al. [20, Lemma 1] and Tomioka et al. [26, Lemma 2], and it bounds the Schatten 1-norm of the orthogonal part $\boldsymbol{\Delta}''_k$ with that of the partially correlated part $\boldsymbol{\Delta}'_k$ and also bounds the rank of $\boldsymbol{\Delta}'_k$.

Lemma 4. *Let $\hat{\mathcal{W}}$ be the solution of the minimization problem* (11) *with the regularization constant $\lambda \geq 2\left\|\mathcal{E}\right\|_{\underline{S_\infty/\infty}}$. Let $\Delta^{(k)}$ and its decomposition be as defined above. Then we have*

1. $\mathrm{rank}(\boldsymbol{\Delta}'_k) \leq 2\bar{r}_k$.
2. $\sum_{k=1}^K \|\boldsymbol{\Delta}''_k\|_{S_1} \leq 3\sum_{k=1}^K \|\boldsymbol{\Delta}'_k\|_{S_1}$.

Note that although the proof of the above statement closely follows that of Tomioka et al. [26, Lemma 2], the notion of rank is different. In their result, the rank is the mode-k rank $\underline{r}_k$, whereas the rank here is the mode-k rank of the kth component $\mathcal{W}^{*(k)}$ of the truth.

The following lemma relates the squared Frobenius norm of the difference of the sums $\left\|\sum_{k=1}^K \Delta^{(k)}\right\|_F^2$ with the sum of squared differences $\sum_{k=1}^K \left\|\Delta^{(k)}\right\|_F^2$

Lemma 5. *Let $\hat{\mathcal{W}}$ be the solution of the minimization problem* (11)*. Then we have,*

$$\frac{1}{2}\sum_{k=1}^{K}\left\|\Delta^{(k)}\right\|_F^2 \leq \frac{1}{2}\left\|\Delta\right\|_F^2 + \alpha(K-1)\sum_{k=1}^{K}\|\boldsymbol{\Delta}_{(k)}^{(k)}\|_{S_1},$$

where $\Delta = \sum_{k=1}^{K}\Delta^{(k)}$.

Proof of Theorem 2. First from the optimality of $\hat{\mathcal{W}}$, we have

$$\frac{1}{2}\left\|\mathcal{Y}-\hat{\mathcal{W}}\right\|_F^2 + \lambda\sum\nolimits_{k=1}^{K}\|\hat{\boldsymbol{W}}_{(k)}^{(k)}\|_{S_1} \leq \frac{1}{2}\left\|\mathcal{Y}-\mathcal{W}^*\right\|_F^2 + \lambda\sum\nolimits_{k=1}^{K}\|\boldsymbol{W}_{(k)}^{*(k)}\|_{S_1},$$

which implies

$$\begin{aligned}
\frac{1}{2}\left\|\Delta\right\|_F^2 &\leq \langle\Delta,\mathcal{E}\rangle + \lambda\sum\nolimits_{k=1}^{K}\left(\|\boldsymbol{W}_{(k)}^{*(k)}\|_{S_1} - \|\hat{\boldsymbol{W}}_{(k)}^{(k)}\|_{S_1}\right) && (21)\\
&\leq \langle\Delta,\mathcal{E}\rangle + \lambda\sum\nolimits_{k=1}^{K}\|\boldsymbol{\Delta}_{(k)}^{(k)}\|_{S_1}\\
&\leq (\left\|\mathcal{E}\right\|_{\underline{S_\infty/\infty}} + \lambda)\sum\nolimits_{k=1}^{K}\|\boldsymbol{\Delta}_{(k)}^{(k)}\|_{S_1}, && (22)
\end{aligned}$$

where we used the fact that $\mathcal{Y} = \mathcal{W}^* + \mathcal{E}$ in the first line, the triangular inequality in the second line, and Hölder's inequality in the third line. Note that there is an additional looseness in the third line due to the fact that $\Delta = \sum_{k=1}^{K}\Delta^{(k)}$ is not an optimal decomposition of Δ induced by the latent Schatten 1-norm.

Next, combining inequality (22) with Lemma 5, we have

$$\frac{1}{2}\sum\nolimits_{k=1}^{K}\left\|\Delta^{(k)}\right\|_F^2 \leq 2\lambda\sum\nolimits_{k=1}^{K}\|\boldsymbol{\Delta}_{(k)}^{(k)}\|_{S_1},$$

where we used the fact that $\lambda \geq \left\|\mathcal{E}\right\|_{\underline{S_\infty/\infty}} + \alpha(K-1)$.

Finally, combining the above inequality with Lemma 4, we have

$$\begin{aligned}
\frac{1}{2}\sum\nolimits_{k=1}^{K}\left\|\Delta^{(k)}\right\|_F^2 &\leq 2\lambda\sum\nolimits_{k=1}^{K}(\|\boldsymbol{\Delta}_k'\|_{S_1} + \|\boldsymbol{\Delta}_k''\|_{S_1})\\
&\leq 8\lambda\sum\nolimits_{k=1}^{K}\|\boldsymbol{\Delta}_k'\|_{S_1}\\
&\leq 8\lambda\sum\nolimits_{k=1}^{K}\sqrt{2\bar{r}_k}\|\boldsymbol{\Delta}_k'\|_F\\
&\leq 8\lambda\sum\nolimits_{k=1}^{K}\sqrt{2\bar{r}_k}\left\|\Delta^{(k)}\right\|_F\\
&\leq 8\sqrt{2}\lambda\sqrt{\sum\nolimits_{k=1}^{K}\bar{r}_k}\sqrt{\sum\nolimits_{k=1}^{K}\left\|\Delta^{(k)}\right\|_F^2}, && (23)
\end{aligned}$$

where we used the triangular inequality in the first line, Lemma 4 in the second line, Hölder's inequality in the third line (combined with Lemma 4), the fact that $\boldsymbol{\Delta}_{(k)}^{(k)} = \boldsymbol{\Delta}_k' + \boldsymbol{\Delta}_k''$ is an orthogonal decomposition in the fourth line, and Cauchy-Schwarz inequality in the last line. Dividing both sides of the last inequality by $\sqrt{\sum_{k=1}^{K}\left\|\mathcal{W}^{(k)}\right\|_F^2}$ completes the first part of our claim.

In order to obtain the bound on the overall error $\left\|\Delta\right\|_F$, we consider two cases: if

$$\left\|\Delta\right\|_F \leq \sqrt{\sum\nolimits_{k=1}^{K}\left\|\Delta^{(k)}\right\|_F^2}, \tag{24}$$

we can lower-bound the left hand side of inequality (23) by (24) and obtain

$$\frac{1}{2}\left\|\Delta\right\|_F \leq 8\sqrt{2}\lambda\sqrt{\sum\nolimits_{k=1}^{K}\bar{r}_k}.$$

We obtain our claim by choosing $\mathcal{W}^{*(k)} = \mathcal{W}^*$ for $k = \operatorname{argmin}_k \underline{r}_k$ and $\mathcal{W}^{*(k)} = 0$, otherwise.

On the other hand, if

$$\sum_{k=1}^{K} \left\| \Delta^{(k)} \right\|_F^2 \leq \left\| \Delta \right\|_F^2, \tag{25}$$

we can apply the derivation up to (23) to the right hand side of inequality (22) to obtain

$$\begin{aligned} \frac{1}{2} \left\| \Delta \right\|_F^2 &\leq 8\sqrt{2}\lambda \sqrt{\sum_{k=1}^{K} \bar{r}_k} \sqrt{\sum_{k=1}^{K} \left\| \Delta^{(k)} \right\|_F^2} \\ &\leq 8\sqrt{2}\lambda \sqrt{\sum_{k=1}^{K} \bar{r}_k} \left\| \Delta \right\|_F, \end{aligned}$$

where we used assumption (25) in the second line. We obtain our claim by dividing both sides by $\left\| \Delta \right\|_F$ and choosing $\mathcal{W}^{*(k)} = \mathcal{W}^*$ for $k = \operatorname{argmin}_k \underline{r}_k$ and $\mathcal{W}^{*(k)} = 0$, otherwise.

□

E Proof of Theorem 3

Proof. Since each entry of $\mathcal{E}$ is an independent zero men Gaussian random variable with variance σ^2, for each mode k we have the following tail bound (Corollary 5.35 in [28])

$$P\left(\|\boldsymbol{E}_{(k)}\|_{S_\infty} > \sigma \left(\sqrt{N/n_k} + \sqrt{n_k} \right) + t \right) \leq \exp\left(-t^2/(2\sigma^2) \right).$$

Next, taking a union bound

$$P\left(\max_k \|\boldsymbol{E}_{(k)}\|_{S_\infty} > \sigma \max_k \left(\sqrt{N/n_k} + \sqrt{n_k} \right) + t \right) \leq K \exp\left(-t^2/(2\sigma^2) \right).$$

Thus defining $\delta = K \exp(-t^2/(2\sigma^2))$, we have

$$\left\| \mathcal{E} \right\|_{\underline{S_\infty/\infty}} \leq \sigma \max_k \left(\sqrt{N/n_k} + \sqrt{n_k} \right) + \sigma \sqrt{2 \log(K/\delta)} \quad \text{with probability at least } 1 - \delta.$$

Therefore if we take

$$\lambda = 2\sigma \left(\sqrt{N/n_K} + \sqrt{n_1} + \sqrt{2 \log(K/\delta)} \right) + \alpha(K-1),$$

the condition of Theorem 2 will be satisfied with probability at least $1 - \delta$. Substituting the above λ into the right hand side of the error bound (12) in Theorem 2, we have the statement of Theorem 3.

□

F Discussion on the identifiability

Let $\bar{r}_k = \operatorname{rank}(\boldsymbol{W}_{(k)}^{(k)})$ be the mode-k rank of the kth component $\mathcal{W}^{(k)}$ in the decomposition

$$\mathcal{W} = \mathcal{W}^{(1)} + \mathcal{W}^{(2)} + \cdots + \mathcal{W}^{(K)}. \tag{26}$$

We say that a decomposition (26) is *locally identifiable* when there is no other decomposition $\sum_{k=1}^{K} \tilde{\mathcal{W}}^{(k)}$ having the same rank $(\bar{r}_1, \ldots, \bar{r}_K)$. The following theorem fully characterizes the local identifiability of the decomposition (26).

Theorem 4. *The decomposition* (26) *is* locally identifiable *if and only if* $\mathcal{W}^{(k^*)} = \mathcal{W}$ *for* $k = k^*$ *and* $\mathcal{W}^{(k)} = 0$ *otherwise, for some* k^*.

Proof. We first prove the "if" direction. suppose that there is another decomposition

$$\sum_{k=1}^{K} \mathcal{W}^{(k)} = \sum_{k=1}^{K} \tilde{\mathcal{W}}^{(k)},$$

such that $\mathrm{rank}(\boldsymbol{W}^{(k)}_{(k)}) = \mathrm{rank}(\tilde{\boldsymbol{W}}^{(k)}_{(k)})$. Note that $\mathcal{W}^{(k)} \neq \tilde{\mathcal{W}}^{(k)}$ can happen only when $\mathcal{W}^{(k)} \neq 0$ (otherwise the rank would increase). Now if $\mathcal{W}^{(k)} \neq 0$ and $\mathcal{W}^{(k)} \neq \tilde{\mathcal{W}}^{(k)}$, then there must be $\ell \neq k$ such that $\mathcal{W}^{(\ell)} \neq \tilde{\mathcal{W}}^{(\ell)}$. This, however, would mean $\mathrm{rank}(\tilde{\boldsymbol{W}}^{(\ell)}_{(\ell)}) > \mathrm{rank}(\boldsymbol{W}^{(\ell)}_{(\ell)}) = 0$, which is a contradiction.

Conversely, suppose that there are $k \neq \ell$ such that $\mathcal{W}^{(k)} \neq 0$ and $\mathcal{W}^{(\ell)} \neq 0$, we can write[3]

$$\begin{aligned}\mathcal{W}^{(k)} &= \mathcal{C}^{(k)} \times_k \boldsymbol{U}_k,\\ \mathcal{W}^{(\ell)} &= \mathcal{C}^{(\ell)} \times_\ell \boldsymbol{U}_\ell,\end{aligned}$$

where $\boldsymbol{U}_k \in \mathbb{R}^{n_k \times \bar{r}_k}$, $\mathcal{C}^{(k)} \in \mathbb{R}^{n_1 \times \cdots \times n_{k-1} \times \bar{r}_k \times \cdots \times n_K}$, and $\boldsymbol{U}_\ell$ and $\mathcal{C}^{(\ell)}$ are defined similarly. Since $\mathcal{C}^{(k)}$ and $\mathcal{C}^{(\ell)}$ are allowed to be full rank, we can define

$$\begin{aligned}\tilde{\mathcal{C}}^{(k)} &= \mathcal{C}^{(k)} + \mathcal{D}^{(k,\ell)} \times_\ell \boldsymbol{U}_\ell,\\ \tilde{\mathcal{C}}^{(\ell)} &= \mathcal{C}^{(\ell)} - \mathcal{D}^{(k,\ell)} \times_k \boldsymbol{U}_k,\end{aligned}$$

for any $\mathcal{D} \in \mathbb{R}^{n_1 \times \cdots \times \bar{r}_k \times \cdots \times \bar{r}_\ell \times \cdots \times n_K}$. Then we have

$$\begin{aligned}\mathcal{W}^{(k)} + \mathcal{W}^{(\ell)} &= \mathcal{C}^{(k)} \times_k \boldsymbol{U}_k + \mathcal{C}^{(\ell)} \times_\ell \boldsymbol{U}_\ell\\ &= \left(\mathcal{C}^{(k)} + \mathcal{D}^{(k,\ell)} \times_\ell \boldsymbol{U}_\ell\right) \times_k \boldsymbol{U}_k\\ &\quad + \left(\mathcal{C}^{(\ell)} - \mathcal{D}^{(k,\ell)} \times_k \boldsymbol{U}_k\right) \times_\ell \boldsymbol{U}_\ell\\ &= \tilde{\mathcal{C}}^{(k)} \times_k \boldsymbol{U}_k + \tilde{\mathcal{C}}^{(\ell)} \times_\ell \boldsymbol{U}_\ell\\ &= \tilde{\mathcal{W}}^{(k)} + \tilde{\mathcal{W}}^{(\ell)}.\end{aligned}$$

Note that $\mathrm{rank}(\tilde{\boldsymbol{W}}^{(k')}_{(k')}) = \bar{r}_{k'}$ for $k' = k, \ell$. Therefore, there are infinitely many decompositions that have the same rank $(\bar{r}_1, \ldots, \bar{r}_K)$.

□

The above theorem partly explains the difficulty of estimating individual components $\mathcal{W}^{*(k)}$ *without additional incoherence assumption* as in (10). In fact, most decompositions of the form (9) are not identifiable.

G Proof of Lemma 4

Proof. The first inequality is true, because

$$\begin{aligned}\boldsymbol{\Delta}'_k &= \boldsymbol{\Delta}^{(k)}_{(k)} - (\boldsymbol{I}_{n_k} - \boldsymbol{U}_k \boldsymbol{U}_k{}^\top)\boldsymbol{\Delta}^{(k)}_{(k)}(\boldsymbol{I}_{N/n_k} - \boldsymbol{V}_k \boldsymbol{V}_k{}^\top)\\ &= \boldsymbol{U}_k \boldsymbol{U}_k{}^\top \boldsymbol{\Delta}^{(k)}_{(k)}(\boldsymbol{I}_{N/n_k} - \boldsymbol{V}_k \boldsymbol{V}_k{}^\top) + \boldsymbol{\Delta}^{(k)}_{(k)} \boldsymbol{V}_k \boldsymbol{V}_k{}^\top.\end{aligned}$$

Next, to show the second inequality, notice that

$$\begin{aligned}\|\hat{\boldsymbol{W}}^{(k)}_{(k)}\|_{S_1} &= \|\boldsymbol{W}^{*(k)}_{(k)} + \boldsymbol{\Delta}''_k + \boldsymbol{\Delta}'_k\|_{S_1}\\ &\geq \|\boldsymbol{W}^{*(k)}_{(k)} + \boldsymbol{\Delta}''_k\|_{S_1} - \|\boldsymbol{\Delta}'_k\|_{S_1}\\ &= \|\boldsymbol{W}^{*(k)}_{(k)}\|_{S_1} + \|\boldsymbol{\Delta}''_k\|_{S_1} - \|\boldsymbol{\Delta}'_k\|_{S_1}\end{aligned}$$

where we used the decomposability [20] of the Schatten 1-norm in the third line.

[3]Here the tensor mode-k product $\mathcal{A} = \mathcal{B} \times_k \boldsymbol{C}$ is defined as $a_{i_1 \ldots i_K} = \sum_{\ell=1}^{d_k} b_{i_1 i_2 \ldots \ell \ldots i_K} c_{\ell i_k}$ where $\mathcal{A} = (a_{i_1 \ldots i_K}) \in \mathbb{R}^{n_1 \times \cdots \times n_K}$, $\mathcal{B} = (b_{i_1 \ldots \ell \ldots i_K}) \in \mathbb{R}^{n_1 \times \cdots \times d_k \times \cdots \times n_K}$, and $\boldsymbol{C} = (c_{\ell i_k}) \in \mathbb{R}^{d_k \times n_k}$

Substituting the above inequality into inequality (21), we obtain

$$0 \leq \frac{1}{2}\left|\!\left|\!\left|\Delta\right|\!\right|\!\right|_F^2 \leq \frac{\lambda}{2}\sum_{k=1}^{K}\|\boldsymbol{\Delta}_{(k)}^{(k)}\| + \lambda\sum_{k=1}^{K}(\|\boldsymbol{\Delta}_k'\|_{S_1} - \|\boldsymbol{\Delta}_k''\|_{S_1})$$
$$\leq \lambda\sum_{k=1}^{K}\left(\frac{3}{2}\|\boldsymbol{\Delta}_k'\|_{S_1} - \frac{1}{2}\|\boldsymbol{\Delta}_k''\|_{S_1}\right),$$

from which the statement follows. Here we used $\left|\!\left|\!\left|\mathcal{E}\right|\!\right|\!\right|_{\underline{S_\infty/\infty}} \leq \lambda/2$ in the second inequality and the triangular inequality in the last line. □

H Proof of Lemma 5

Proof.

$$\left|\!\left|\!\left|\Delta\right|\!\right|\!\right|_F^2 = \sum_{k=1}^{K}\left|\!\left|\!\left|\Delta^{(k)}\right|\!\right|\!\right|_F^2 + \sum_{k=1}^{K}\sum_{k'\neq k}\left\langle \boldsymbol{\Delta}_{(k)}^{(k)}, \boldsymbol{\Delta}_{(k)}^{(k')}\right\rangle$$
$$\geq \sum_{k=1}^{K}\left|\!\left|\!\left|\Delta^{(k)}\right|\!\right|\!\right|_F^2 - \sum_{k=1}^{K}\|\boldsymbol{\Delta}_{(k)}^{(k)}\|_{S_1}\sum_{k'\neq k}\|\boldsymbol{\Delta}_{(k)}^{(k')}\|_{S_\infty}$$
$$\geq \sum_{k=1}^{K}\left|\!\left|\!\left|\Delta^{(k)}\right|\!\right|\!\right|_F^2 - 2\alpha(K-1)\sum_{k=1}^{K}\|\boldsymbol{\Delta}_{(k)}^{(k)}\|_{S_1},$$

where the last inequality follows from $\|\boldsymbol{\Delta}_{(k)}^{(k')}\|_{S_\infty} \leq \|\hat{\mathcal{W}}_{(k)}^{(k')}\|_{S_\infty} + \|\mathcal{W}_{(k)}^{*(k')}\|_{S_\infty} \leq 2\alpha$ for $k' \neq k$. Lemma follows by dividing both sides by two. □