Abstract

We consider the problem of recovering a low-rank tensor from a noisy observation. Previous work has shown \( O(n^{K/4}) \) recovery guarantee for recovering a \( K \)th order rank one tensor of size \( n \times \cdots \times n \) by an algorithm called recursive unfolding. In this paper, we first improve this to \( O(n^{K/4}) \) by a much simpler approach but with a more careful analysis. Then we propose a new norm based on the Kronecker products of factors obtained by the proposed simple estimator. The imposed Kronecker structure of the new norm allows us to show a nearly ideal \( O(\sqrt{n} + \sqrt{m}) \) bound for the proposed subspace norm, in which the parameter \( m \) controls the blend from the non-convex estimator to mode-wise nuclear norm minimization. Furthermore we empirically demonstrate that with \( m = O(1) \), the proposed norm achieves near ideal denoising performance.

1 Introduction

Tensor is a natural way to express higher order interactions for variety of data and tensor decomposion has been successfully applied to wide areas ranging from chemometrics (Smilde et al., 2005), signal processing (Cardoso, 1991) to neuroimaging (Mørup, 2011); see Kolda and Bader (2009) for a survey. Moreover, recently it has become an active area in the context of learning latent variable models (Anandkumar et al., 2014) and learning distributed representation of words (Kiros et al., 2014).

Although tensors can be considered as a natural generalization of matrices, their mathematical properties are widely different. For example, finding the rank of a tensor or finding a best rank-one approximation of it is known to be NP hard (Hastad, 1990; Hillar and Lim, 2013).

A related statistical problem is, assuming that we observe a randomly corrupted version of a low-rank tensor, how well we can recover the underlying true tensor with a polynomial time algorithm. Since we can convert a tensor into a matrix by an operation known as unfolding, some recent work (Tomioka et al., 2011b; Mu et al., 2014; Richard and Montanari, 2014; Jain and Oh, 2014) has shown that we do get nontrivial recovery guarantees by using some norms or singular value decompositions. More specifically, Richard and Montanari (2014) has shown that when a rank-one \( K \)th order tensor of size \( n \times \cdots \times n \) is corrupted by standard Gaussian noise, a non-trivial bound can be shown if the signal to noise ratio \( \beta/\sigma \geq O_p(n^{K/4}) \) for even order tensors and \( \beta/\sigma \geq O_p(n^{\lceil K/2 \rceil/2}) \) for odd order tensors by recursive unfolding. Note that \( \beta/\sigma \geq O_p(\sqrt{n}) \) is sufficient for matrices (\( K = 2 \)) and also for tensors if we use the best rank-one approximation (which is known to be NP hard) as an estimator. Jain and Oh (2014) analyzed the tensor completion problem and proposed an algorithm that requires \( O(n^{3/2} \cdot \log(n)) \) samples for \( K = 3 \); here information theoretically we need at least \( \Omega(n) \) samples and the intractable maximum likelihood estimator would require \( O(n \cdot \log(n)) \) samples. Therefore, there is a wide gap between the ideal estimator and what we can achieve using a polynomial time algorithm. A subtle question that we will address in this paper is whether we need to unfold the tensor so that the resulting matrix become as square as possible, which was the reasoning underlying both Mu et al. (2014); Richard and Montanari (2014).

Nevertheless, non-convex estimators based on alternating minimization or nonlinear optimization (Acar et al., 2011; Sorber et al., 2013) have been widely applied and have performed well when appropriately set up. Therefore it would be of fundamental importance to connect the wisdom of non-convex estimators with the more theoretically motivated estimators that recently emerged.
Table 1: Comparison of required signal-to-noise ratio $\beta/\sigma$ of different algorithms for recovering a $K$th order rank one tensor of size $n \times \cdots \times n$ contaminated by Gaussian noise with Standard deviation $\sigma$. See model (2). The power method and recursive unfolding is from [Richard and Montanari (2014)]. The square norm is from [Mu et al. (2014)]. The overlapped trace norm is from [Tomioka et al. (2011b)]. The latent trace norm is from [Tomioka and Suzuki (2013)]. The bound for the ordinary unfolding is shown in Corollary 1. The bound for the subspace norm is shown in Theorem 2. The ideal estimator is proven in Appendix A.

| Power method                  | Recursive unfolding / square norm | Overlapped norm / Latent trace norm | Ordinary unfolding | Subspace norm (proposed) | Ideal                  |
|-------------------------------|----------------------------------|------------------------------------|--------------------|--------------------------|------------------------|
| $O_p(n^{K/2})$                | $O_p(n^{K/2}/2)$                 | $O_p(n^{(K-1)/2})$                 | $O_p(n^{K/4})$     | $O_p(\sqrt{n + \sqrt{m}})$ | $O_p(\sqrt{nK \log(K)})$ |

In this paper, we explore such a connection by defining a new norm based on Kronecker products of factors that can be obtained by simple mode-wise singular value decomposition of unfolding\(^1\) (or higher-order singular value decomposition, HOSVD; [De Lathauwer et al., 2000a]). Our contributions are two folds. We first study the non-asymptotic behavior of the leading singular vector from the ordinary rectangular unfolding $\tilde{X}_k$ and show a nontrivial bound for signal to noise ratio $\beta/\sigma = O_p(n^{K/4})$ without the square unfolding. Thus the result also applies to odd order tensors confirming a conjecture in [Richard and Montanari (2014)]. Furthermore, this motivates us to use the solution of mode-wise truncated SVDs to construct a new norm. We propose the subspace norm, which predicts an unknown low-rank tensor as a mixture of $K$ low-rank tensors, in which each term takes the form

$$\text{fold}_k(M^{(k)}(\tilde{A}^{(k-1)} \otimes \cdots \otimes \tilde{A}^{(k+1)})^\top),$$

where $\text{fold}_k$ is the inverse of unfolding $(\cdot)_{(k)}$ and $\tilde{A}^{(k)} \in \mathbb{R}^{n \times \tilde{m}}$ is orthonormal matrix estimated from the mode-$k$ unfolding of the observed tensor, for $k = 1, \ldots, K$; $\tilde{m}$ is a parameter, and $M^{(k)} \in \mathbb{R}^{n \times m}$, where $m = \tilde{m} K^{-1}$. Our theory tells us that with sufficiently high signal-to-noise ratio the estimated $\tilde{A}^{(k)}$ spans the true factors. Moreover, we penalize the nuclear norm of each $M^{(k)}$ to be robust to the error in estimating $\tilde{A}^{(k)}$'s.

We highlight our contributions below:

1. We prove that the required signal-to-noise ratio for recovering a $K$th order rank one tensor from the rectangular unfolding is $O_p(n^{K/4})$. Our analysis shows a curious two phase behavior: when $O_p(n^{K/4}) \leq \beta/\sigma \leq O_p(n^{K/2})$, the error shows a fast decay as $1/\beta^4$. For $\beta/\sigma \geq O_p(n^{K/2})$, the error decays slowly as $1/\beta^2$. We confirm this in a numerical simulation.

2. The proposed subspace norm is an interpolation between the intractable estimators that directly control the rank (e.g., HOSVD, [De Lathauwer et al., 2000b]) and the tractable norm-based estimators. It becomes equivalent to the latent trace norm proposed by [Tomioka and Suzuki (2013)] when $\tilde{m} = n$ at the cost of increased signal-to-noise ratio threshold (see Table 1).

3. The proposed estimator is also more efficient than previously proposed norm based estimators because the size of the SVD required in the algorithm is reduced from $n \times n^{K-1}$ to $n \times m$.

4. We also empirically demonstrate that the proposed subspace norm performs nearly optimally for constant order $m$.

**Notation**

Here we summarize the notations we use in this paper. The numbers of dimension of the tensor is denoted by $n_1, \ldots, n_K$. In the simpler square case, we use $n$. We define $n_{\ell k} = \prod_{\ell \neq k} n_{\ell}$, which equals $n^{K-1}$ in the square case. The mode-$k$ unfolding $X_{(k)}$ of $X$ is the $n_k \times n_{\backslash k}$ matrix obtained by concatenating all mode-$k$ fibers. For

\(^1\)Mode-$k$ unfolding $X_{(k)}$ of a tensor $X$ is an $n_k \times \prod_{\ell \neq k} n_{\ell}$ matrix constructed by concatenating the mode-$k$ fibers along columns.
$K \geq 4$, a more general unfolding can be defined by partitioning the $K$ indices of $X$ into two parts. For example, for $K = 4$, we denote by $X_{(1,2;3,4)}$ the $n_1 n_2 \times n_3 n_4$ matrix obtained by considering the first two indices as linear index for the rows and the last two indices as the linear index for the columns. Note that the ordinary rectangular unfolding can be written as $X_{(k,k−1...,k+1)}$. The inner product between two vectors $u$ and $v$ is denoted by $\langle u, v \rangle$. The inner product between two matrices and tensors are defined as the inner product of them as vectors. We denote the spectral norm, nuclear norm, and Frobenius norm for matrices by $\| \cdot \|$, $\| \cdot \|_*$, and $\| \cdot \|_F$ respectively. For tensors we use $\| \cdot \|_\bullet$.

2  Theory

2.1 Perturbation bound for the left singular vector

We first establish a bound on recovering the left singular vector of a rank-one $n \times m$ matrix perturbed by random Gaussian noise.

Consider the following model known as the information plus noise model [Benaych-Georges and Nadakuditi 2011]:

$$\hat{X} = \beta uv^T + \sigma E,$$

where $u$ and $v$ are unit vectors, $\beta$ is the signal strength, $\sigma$ is the noise standard deviation, and the noise matrix $E$ is assumed to be random with entries sampled i.i.d. from the standard normal distribution. Our goal is to lower-bound the correlation between $u$ and the top left singular vector $\hat{u}$ of $\hat{X}$ for signal-to-noise ratio $\beta/\sigma \geq O_p((mn)^{1/4})$.

A direct application of the classic Wedin perturbation theorem [Welin 1972] to the rectangular matrix $\hat{X}$ does not provide us the desired result. This is because it requires the signal to noise ratio $\beta/\sigma \geq 2\|E\|$. Since the spectral norm of $E$ scales as $O_p(\sqrt{n} + \sqrt{m})$ [Vershynin 2010], this would mean that we require $\beta/\sigma \geq O_p(m^{1/2})$; i.e., the threshold is dominated by the number of columns $m$.

Alternatively, we can view $\hat{u}$ as the leading eigenvector of $\hat{X}\hat{X}^T$, a square matrix. Our key insight is that we can decompose $\hat{X}\hat{X}^T$ as follows:

$$\hat{X}\hat{X}^T = (\beta^2 uu^T + m\sigma^2 I) + (\sigma^2 EE^T - m\sigma^2 I) + \beta\sigma(uv^TE^T + Evu^T).$$

Note that $u$ is the leading eigenvector of the first term because adding an identity matrix does not change the eigenvectors. Moreover, we notice that there are two noise terms: the first term is a centered Wishart matrix and it is independent of the signal $\beta$; the second term is Gaussian distributed and depends on the signal $\beta$.

This implies a two-phase behavior corresponding to either the Wishart noise term or the Gaussian noise term being dominant depending on the value of $\beta$. Interestingly, we get a different speed of convergence for each of these phases as we show in the next theorem.

**Theorem 1.** There exists a constant $C$ such that with probability at least $1 - 4e^{-n}$, if $m/n \geq C$, then

$$|\langle \hat{u}, u \rangle| \geq \begin{cases} 1 - \frac{Cnm}{(\beta/\sigma)^2}, & \text{if } \sqrt{m} \geq \frac{\beta}{\sigma} \geq (Cnm)^{1/2}, \\ 1 - \frac{Cn}{(\beta/\sigma)^2}, & \text{if } \beta/\sigma \geq \sqrt{m}, \end{cases}$$

otherwise, $|\langle \hat{u}, u \rangle| \geq 1 - \frac{Cn}{(\beta/\sigma)^2}$ if $\beta/\sigma \geq \sqrt{Cn}$.

**Proof.** We prove the theorem in Appendix B

In other words, if $\hat{X}$ has sufficiently many more columns than rows, as the signal to noise ratio $\beta/\alpha$ increases, $\hat{u}$ first converges to $u$ as $1/\beta^2$, and then as $1/\beta^4$.

Figure 1 illustrates these results. We randomly generate a rank-one $100 \times 10000$ matrix perturbed by Gaussian noise, and measure the distance between $\hat{u}$ and $u$. It shows that the phase transition happens at $\beta/\sigma = (nm)^{1/4}$, and there are two regimes of different convergence rates as Theorem 1 predicts.
Figure 1: Synthetic experiment showing phase transition at $\beta/\sigma = (nm)^{1/4}$ and regimes with different rates of convergence. The observed matrix $\hat{X}$ is generated as in Theorem 1. As $\beta/\sigma$ grows, the distance between $\hat{u}$ and $u$ decreases as $1/\beta^4$ between $(nm)^{1/4}$ and $\sqrt{m}$, and as $1/\beta^2$ after $\sqrt{m}$.

2.2 Tensor Unfolding

Now let’s apply the above result to the tensor version of information plus noise model studied by [Richard and Montanari, 2014]. We consider a rank one $n \times n \times \cdots \times n$ tensor (signal) contaminated by Gaussian noise as follows:

$$ Y = \mathcal{X}^* + \sigma E = \beta u^{(1)} \otimes \cdots \otimes u^{(K)} + \sigma E, $$

(2)

where factors $u^{(k)} \in \mathbb{R}^n$, $k = 1, \ldots, K$, are unit vectors, which are not necessarily identical, and the entries of $E \in \mathbb{R}^{n \times \cdots \times n}$ are i.i.d samples from the normal distribution $\mathcal{N}(0, 1)$. Note that this is slightly more general (and easier to analyze) than the symmetric setting studied by [Richard and Montanari, 2014].

Several estimators for recovering the low-rank part $\mathcal{X}^*$ from its noisy version $Y$ have been proposed. The overlapped trace norm [Signoretto et al., 2010; Gandy et al., 2011; Liu et al., 2009; Tomioka et al., 2011b] is an extension of nuclear norm (Fazel et al., 2001; Srebro and Shraibman, 2005) and is defined as follows:

$$ \| \mathcal{X} \|_{\text{overlap}} = \sum_{k=1}^{K} \| \mathcal{X}^{(k)} \|_*, $$

where $\| \mathcal{X} \|_* = \sum_{j=1}^{r} \sigma_j(\mathcal{X})$ is the nuclear norm (also known as the trace norm), $r$ is the rank of $\mathcal{X}$, and $\mathcal{X}^{(k)}$ denotes the mode-$k$ unfolding of $\mathcal{X}$.

It was shown in [Tomioka et al., 2011b] that the estimator $\hat{\mathcal{X}}$ defined as

$$ \hat{\mathcal{X}} = \arg \min_{\mathcal{X}} \left( \frac{1}{2} \| Y - \mathcal{X} \|_F^2 + \lambda \| \mathcal{X} \|_{\text{overlap}} \right) $$
achieves the relative performance guarantee

\[
\frac{1}{\beta} \| \hat{X} - X^* \|_F \leq O_p \left( \frac{\sigma \sqrt{nK^{-1}}}{\beta} \right).
\]

The latent trace norm (Tomioka and Suzuki, 2013) and the scaled latent trace norm (Wimalawarne et al., 2014) posed subsequently achieve the same performance guarantee in this setting. The above bound implies that if we want to obtain relative error smaller than \( \varepsilon \), we need the signal to noise ratio \( \beta / \sigma \) to scale as \( \beta / \sigma \geq O_p(\sqrt{nK^{-1}}/\varepsilon) \).

Mu et al. (2014) proposed the square norm defined as follows:

\[
\| \mathcal{X} \|_{\text{square}} = \| X_{(1,\ldots,[K/2];[K/2],\ldots,K)} \|_*.
\]

This norm improves the right hand side of inequality (3) to \( O_p(\sigma \sqrt{nK/2}/\beta) \), which translates to requiring \( \beta / \sigma \geq O_p(\sqrt{nK/2}/\varepsilon) \) for obtaining relative error \( \varepsilon \). The intuition here is the more square the unfolding is the better the bound becomes. However, there is no improvement for \( K = 3 \).

Richard and Montanari (2014) studied model (2) and proved that a recursive unfolding algorithm and the tensor power method (De Lathauwer et al. 2000b; Kolda and Mayo 2011; Anandkumar et al. 2014) can achieve factor recovery error \( \text{dist}^2(\hat{u}^{(k)}, u^{(k)}) = \varepsilon \) with \( \beta / \sigma \geq O_p(n\sqrt{K/2}/\varepsilon) \) and \( \beta / \sigma \geq O_p(\max(\sqrt{n}/\varepsilon, nK/2)/\sqrt{K \log K}) \), respectively, where \( \text{dist}(u, u') := \min(\|u - u'\|, \|u + u'\|) \).

The reasoning underlying both Mu et al. (2014) and Richard and Montanari (2014) is that square unfolding is better. However, if we take the mode-\( k \) (rectangular) unfolding

\[
Y_{(k)} = \beta u^{(k)}(u^{(k-1)} \otimes \ldots \otimes u^{(1)} \otimes u^{(K)} \otimes \ldots \otimes u^{(k+1)})^T + \sigma E_{(k)},
\]

we can see recovering factors \( u^{(k)} \) in (2) as an information plus noise model (1) where \( m/n = nK^{-2} \). Thus the rectangular unfolding satisfies the condition of Theorem 1 for \( n \) or \( K \) large enough. Therefore we have the next corollary.

**Corollary 1.** Consider a \( K \)th order rank one tensor contaminated by Gaussian noise as in (2) for \( K \geq 3 \). There exists a constant \( C \) such that if \( nK^{-2} \geq C \), then with probability at least 1 - 4Ke\(^{-n}\), for \( k = 1, \ldots, K \) we have

\[
\text{dist}^2(\hat{u}^{(k)}, u^{(k)}) \leq \begin{cases} 
2Cn^K \left( \frac{\beta}{\sigma} \right)^{\frac{1}{2}}, & \text{if} \ n^{K-1} > \beta / \sigma \geq C_n^{1/2}n^{K-1}, \\
2Cn^{K} \left( \frac{\beta}{\sigma} \right)^{\frac{1}{2}}, & \text{if} \ \beta / \sigma \geq n^{K-1},
\end{cases}
\]

where \( \hat{u}^{(k)} \) is the leading left singular vector of the rectangular unfolding \( Y_{(k)} \), and \( \text{dist} \) is defined as above.

This proves that as conjectured by Richard and Montanari (2014), the threshold \( \beta / \sigma \geq O_p(nK^{1/4}) \) applies not only to the even order case but also to the odd order case.

The statement easily extends to more general \( n_1 \times \cdots \times n_K \) tensor by replacing the conditions by \( \sqrt{nK} > \beta / \sigma \geq (C \prod_{k=1}^Kn_k)^{1/4} \) and \( \beta / \sigma \geq \sqrt{nK} \).

We demonstrate this result in Figure 2. The model behind the experiment is a slightly more general case in which \( [n_1, n_2, n_3] = [20, 40, 60] \) or \( [40, 80, 120] \) and the signal \( X^* \) is rank two with \( \beta_1 = 20 \) and \( \beta_2 = 10 \). The plot shows the inner products \( \langle u_1^{(1)}, u_1^{(1)} \rangle \) and \( \langle u_2^{(1)}, u_2^{(1)} \rangle \) as a measure of the quality of estimating the two mode-1 factors. The horizontal axis is the normalized noise standard deviation \( \sigma / \left( \prod_{k=1}^Kn_k \right)^{1/4} \). We can clearly see that the inner product decays symmetrically around \( \beta_1 \) and \( \beta_2 \) as predicted by Corollary 1 for both tensors of dimensions \([20, 40, 60]\) and \([40, 80, 120]\).

### 3 Subspace norm for tensors

In this section, we propose the *subspace norm* for low rank tensor decomposition.
σ(\prod_{i} n_{i})^{1/4}

0
5
10
15
20
25
30

0
0.2
0.4
0.6
0.8
1

|\langle u_{1}, \hat{u}_{1} \rangle| - [20, 40, 60]
|\langle u_{2}, \hat{u}_{2} \rangle| - [20, 40, 60]
|\langle u_{1}, \hat{u}_{1} \rangle| - [40, 80, 120]
|\langle u_{2}, \hat{u}_{2} \rangle| - [40, 80, 120]

\beta_{1}
\beta_{2}

Figure 2: Synthetic experiment showing phase transition at \( \beta = \sigma(\prod_{k} n_{k})^{1/4} \) for odd order tensors. The observed tensor \( \mathcal{Y} \) is generated as \( \mathcal{Y} = \mathcal{X}^{*} + \sigma \mathcal{E} \) with the signal tensor \( \mathcal{X}^{*} \) defined in (4). The inner products go down from one to zero symmetrically around \( \sigma = \beta/(\prod_{k} n_{k})^{1/4} \) as predicted by Corollary 1 for both \( [n_{1}, n_{2}, n_{3}] = [20, 40, 60] \) and \( [40, 80, 120] \).

We observe that if a \( K \)th order tensor \( \mathcal{X} \) has low CP rank or low multilinear rank, the right factor of \( \mathcal{X}_{(k)} \) is spanned by

\[ A^{(k-1)} \otimes \cdots \otimes A^{(k+1)}, \]

where the columns of \( A^{(k)} \) are left singular vectors of \( \mathcal{X}_{(k)} \). Our results in Corollary 1 show that this specific Kronecker structure can be exactly recovered under mild conditions. Inspired by this, we model \( \mathcal{X} \) as a mixture of tensors \( \mathcal{Z}^{(1)}, \ldots, \mathcal{Z}^{(K)} \), such that \( \mathcal{Z}_{(k)}^{(k)} \) has a low rank factorization

\[ \mathcal{Z}_{(k)}^{(k)} = \mathcal{M}_{(k)}^{(k)} \mathcal{S}_{(k)}^{(k)\top}, \]

where the right factor \( \mathcal{S}_{(k)}^{(k)} \) is the Kronecker structure. Since it has higher dimensionality than the truth, we penalize the trace norm of the left factor \( \mathcal{M}_{(k)}^{(k)} \).

In the following, we define the subspace norm, suggest an approach to construct the right factor, and prove the denoising bound in the end.

### 3.1 The subspace norm

**Definition 1.** Let \( \mathcal{S}_{(1)}, \ldots, \mathcal{S}_{(K)}^{(K)} \) be matrices s.t. \( \mathcal{S}_{(k)}^{(k)} \in \mathbb{R}^{n_{k} \times m_{k}}, m_{k} \leq n_{k} \). The subspace norm for a \( K \)th order tensor \( \mathcal{X} \) associated with \( \mathcal{S}_{(k)}^{(k)} \) for \( k = 1, \ldots, K \) is defined as

\[
\| \mathcal{X} \|_{s} = \begin{cases} 
\inf_{\mathcal{M}_{(k)}^{(k)}} \sum_{k=1}^{K} \| \mathcal{M}_{(k)}^{(k)} \|_{s}, & \text{if } \mathcal{X} \in \text{Span}(\mathcal{S}_{(k)}^{(k)}), \\
+\infty, & \text{otherwise},
\end{cases}
\]

where \( \| \cdot \|_{s} \) is the Schatten norm.
where

\[
\text{Span}(\{S^{(k)}\}_{k=1}^K) := \{ X : \exists \mathbf{M}^{(1)}, \ldots, \mathbf{M}^{(K)} \text{ s.t. } X = \sum_{k=1}^K \text{fold}_k(\mathbf{M}^{(k)}\mathbf{S}^{(k)^T}) \}.
\]

**Lemma 1.** The dual norm of \(\|\cdot\|_s\) is a semi-norm \(\|X\|_{s^*} = \max_k \|X(k)S^{(k)}\|\).

**Proof.** By definition,

\[
\|Y\|_{s^*} = \sup_{\{\mathbf{M}^{(k)}\}_{k=1}^K} \left\{ \langle Y, \sum_{k=1}^K \text{fold}_k(\mathbf{M}^{(k)}\mathbf{S}^{(k)^T}) \rangle \right\}_s
\]

\[
s.t. \sum_{k=1}^K \|\mathbf{M}^{(k)}\|_s \leq 1
\]

\[
= \sup_{\{\mathbf{M}^{(k)}\}_{k=1}^K} \sum_{k=1}^K \langle Y(k)S^{(k)}, \mathbf{M}^{(k)} \rangle
\]

\[
s.t. \sum_{k=1}^K \|\mathbf{M}^{(k)}\|_s \leq 1
\]

\[
= \max_k \|Y(k)S^{(k)}\|
\]

where we used the H"older inequality in the last line.

3.2 Choosing the subspace

A natural question that arises is how to choose the matrices \(S^{(1)}, \ldots, S^{(k)}\). Suppose the true tensor has CP rank \(R\):

\[
X^{*} = \sum_{r=1}^R \beta_r u_r^{(1)} \otimes \cdots \otimes u_r^{(K)}.
\]

The mode-\(k\) unfolding of a noisy observation \(Y\) reads

\[
Y(k) = X^{*}(k) + \sigma E(k)
\]

\[
= U^{(k)} C \left( U^{(k-1)} \otimes \cdots \otimes U^{(k+1)} \right)^T + \sigma E(k),
\]

where \(C\) is a \(R \times R\) diagonal matrix such that \(C_{rr} = \beta_r, r = 1, \ldots, R\); \(U^{(k)}\) is a \(n_k \times R\) matrix concatenating the mode-\(k\) factors \(u_1^{(k)}, \ldots, u_R^{(k)}\); \(\otimes\) is the column-wise Kronecker product a.k.a Khatri-Rao product.

**Lemma 2.** Let the \(X^{*}(k) = A^{(k)} \Lambda^{(k)} B^{(k)}\) be the SVD of \(X^{*}(k)\) for each \(k\). Assume that for all \(k\), \(R \leq n_k\) and \(U^{(k)}\) has full column rank. It holds that for all \(k\),

i) \(U^{(k)} \in \text{Span}(A^{(k)})\),

ii) \(U^{(k-1)} \otimes \cdots \otimes U^{(k+1)} \in \text{Span}(A^{(k-1)} \otimes \cdots \otimes A^{(k+1)})\).

**Proof.** We prove the lemma in Appendix D.

In Corollary 1 we have shown when the signal to noise ratio is high enough, with high probability we could detect \(A^{(k)}\). Hence we suggest the following three-step approach for tensor denoising:

(i) For each \(k\), unfold the observation tensor in mode \(k\) and compute the top \(\bar{m}_k\) left singular vectors. Concatenate these vectors to obtain a \(n_k \times \bar{m}_k\) matrix \(\tilde{A}^{(k)}\).

7
Algorithm 1 Tensor denoising via the subspace norm

**Input:** noisy tensor \( \hat{Y} \), input ranks \( \hat{m}_1, \ldots, \hat{m}_K \), regularization constant \( \lambda \)

for \( k = 1 \) to \( K \) do

\( \hat{A}^{(k)} \leftarrow \) top \( \hat{m}_k \) left singular vectors of \( Y^{(k)} \)

end for

for \( k = 1 \) to \( K \) do

\( S^{(k)} \leftarrow \hat{A}^{(k-1)} \otimes \cdots \otimes \hat{A}^{(k+1)} \)

end for

**Output:** \( \hat{X} = \arg\min_X \frac{1}{2} \| Y - X \|_F^2 + \lambda \| X \|_s \).

(i) Compute the sums of products of \( M \) computed in closed forms. We further combine the updates of \( D \in \mathbb{R}^{n \times d} \) where \( D \) is the SVD of \( Y \) (Algorithm 2).

(ii) Construct \( S^{(k)} \) as \( \hat{A}^{(k-1)} \otimes \cdots \otimes \hat{A}^{(k+1)} \). The size of \( S^{(k)} \) is \( n_k \times m_k \), where \( m_k = \prod_{\ell \neq k} \hat{m}_\ell \).

(iii) Solve the subspace norm regularized minimization problem

\[
\min_{X} \frac{1}{2} \| Y - X \|_F^2 + \lambda \| X \|_s,
\]

where the subspace norm is associated with \( \{ S^{(k)} \}_{k=1}^K \).

See Algorithm 1 for details.

### 3.3 Optimization

For solving problem (5), we follow the alternating direction method of multipliers described in [Tomioka et al., 2011a]. We scale the objective function in (5) by \( 1/\lambda \), and consider the dual problem

\[
\min_{D, \{ W^{(k)} \}_{k=1}^K} \frac{\lambda}{2} \| D \|_F^2 - \langle D, Y \rangle
\]

subject to \( \max_k \| W^{(k)} \| \leq 1 \),

\[
W^{(k)} = D^{(k)} S^{(k)}, \quad k = 1, \ldots, K,
\]

where \( D \in \mathbb{R}^{n_1 \times n_2 \times \cdots \times n_K} \) is the dual tensor that corresponds to the residual in the primal problem (5), and \( W^{(k)} \)'s are auxiliary variables introduced to make the problem equality constrained.

The augmented Lagrangian function of problem (6) could be written as follows:

\[
L_{\eta}(D, \{ W^{(k)} \}_{k=1}^K, \{ M^{(k)} \}_{k=1}^K) = \frac{\lambda}{2} \| D \|_F^2 - \langle D, Y \rangle + \sum_{k=1}^K \left( \langle M^{(k)}, D^{(k)} S^{(k)} - W^{(k)} \rangle + \frac{\eta}{2} \| D^{(k)} S^{(k)} - W^{(k)} \|_F^2 + 1_{\| W^{(k)} \| \leq 1} \right),
\]

where \( M^{(k)} \)'s are the multipliers, \( \eta \) is the augmenting parameter, and \( 1_{\| W^{(k)} \| \leq 1} \) is the indicator function of the unit spectral norm ball.

We follow the derivation in [Tomioka et al., 2011a] and conclude that the updates of \( D, M^{(k)} \) and \( W^{(k)} \) can be computed in closed forms. We further combine the updates of \( W^{(k)} \) and other steps so that it needs not to be explicitly computed. The sum of the products of \( M^{(k)} \) and \( S^{(k)} \) finally converges to the solution of the primal problem (5), see Algorithm 2.

The update for the Lagrangian multipliers \( M^{(k)} \) \( (k = 1, \ldots, K) \) is written as singular value soft-thresholding operator defined as

\[
\text{prox}_{tr}^r(Z) = P \text{ max}(\Sigma - r, 0) Q^T,
\]

where \( Z = P \Sigma Q^T \) is the SVD of \( Z \).

A notable property of the subspace norm is the computational efficiency. The update of \( M^{(k)} \) requires singular value decomposition, which usually dominates the costs of computation. For problem (6), the size of \( M^{(k)} \) is only \( n_k \times m_k \). Comparing with previous approaches, e.g. the latent approach whose multipliers are \( n_k \times n_\backslash k \) matrices, the size of our variables is much smaller, so the per-iteration cost is reduced.
We define a slightly modified estimator
\[ \hat{Y} \in \mathbb{R}^{n_1 \times \cdots \times n_K} \]
Let \( \eta \) be any tensor that can be expressed as
\[ \eta = \sum_{k=1}^{K} \text{fold}_k \left( M^{(k)} S^{(k)\top} \right) \]
where \( M^{(k)} \in \mathbb{R}^{n_k \times m_k} \) defined as follows:
\[ \mathcal{M}(\rho) := \left\{ \{M^{(k)}\}_{k=1}^{K} : \|\text{fold}_k(M^{(k)})\|_F \leq \frac{\rho}{K} \sqrt{n_k + m_k}, \forall k \neq \ell \right\} \]
This restriction makes sure that \( M^{(k)} \), \( k = 1, \ldots, K \), are incoherent, i.e., each \( M^{(k)} \) has a spectral norm that is as low as a random matrix when unfolded at a different mode \( \ell \). Similar assumptions were used in low-rank plus sparse matrix decomposition (Agarwal et al. 2012; Hsu et al. 2011) and for the denoising bound for the latent trace norm (Tomioka and Suzuki 2013).

Then we have the following statement.

**Theorem 2.** Let \( \mathcal{X}_p \) be any tensor that can be expressed as
\[ \mathcal{X}_p = \sum_{k=1}^{K} \text{fold}_k \left( M^{(k)} S^{(k)\top} \right) \]
which satisfies the above incoherence condition \( \{M^{(k)}\}_{k=1}^{K} \in \mathcal{M}(\rho) \) and let \( r_k \) be the rank of \( M^{(k)} \) for \( k = 1, \ldots, K \). In addition, we assume that each \( S^{(k)} \) is constructed as \( S^{(k)} = \tilde{A}^{(k-1)} \otimes \cdots \otimes \tilde{A}^{(k+1)} \) with \( \tilde{A}^{(k)} \top = \tilde{A}^{(k)} = I_{m_k} \).

Then there are universal constants \( c_0 \) and \( c_1 \) such that any solution \( \hat{X} \) of the minimization problem (8) with \( \lambda = \|\mathcal{X}_p - \mathcal{X}^*\|_F^2 + c_0 \sigma \left( \max_k (\sqrt{n_k} + \sqrt{m_k}) + \sqrt{2 \log(K/\delta)} \right) \) satisfies the following bound
\[ \|\hat{X} - \mathcal{X}^*\|_F \leq \|\mathcal{X}_p - \mathcal{X}^*\|_F + c_1 \lambda \sqrt{\sum_{k=1}^{K} r_k}, \]
with probability at least \( 1 - \delta \).
Note that the right-hand side of the bound consists of two terms. The first term is the approximation error. This term will be zero if $X^*$ lies in $\text{Span}\{\{S^{(k)}\}_{k=1}^K\}$. This is the case, if we choose $S^{(k)} = I_{m_k}$ as in the latent trace norm, or if the condition of Corollary 1 is satisfied for all components $\beta_1, \ldots, \beta_R$ when we use the Kronecker product construction we proposed. Note that the regularization constant $\lambda$ should also scale with the dual subspace norm of the misspecification $X_p - X^*$.

The second term is the estimation error with respect to $X_p$. If we take $X_p$ to be the orthogonal projection of $X^*$ to the $\text{Span}\{\{S^{(k)}\}_{k=1}^K\}$, we can ignore the misspecification term because $(X_p - X^*)_{(k)}S^{(k)}$ is zero. Then the estimation error scales mildly with the dimensions $n_k, m_k$ and with the sum of the ranks. Note that if we take $S^{(k)} = I_{m_k}, m_k = n \setminus k$ and we recover the guarantee in the square case $n_1 = \cdots = n_K = n$.

4 Experiments

In this section, we conduct tensor denoising experiments on synthetic and real datasets, to numerically justify our analysis in previous sections.

4.1 Synthetic data

We randomly generated a rank two tensor $X^*$ of size $20 \times 30 \times 40$ from the spiked model, with singular value $\beta_1 = 20$ and $\beta_2 = 10$. For each mode, the true factors are obtained by taking left singular vectors of a square matrix randomly drawn from the standard normal distribution. The observation tensor $Y$ is then generated by adding random Gaussian noise with standard deviation $\sigma$ to $X^*$. For $\sigma$, we chose 20 values linearly spaced between $0.1\beta_2/(\prod_i n_i)^{1/4}$ and $1.2\beta_1/(\prod_i n_i)^{1/4}$.

Our approach is compared to the CP decomposition, the overlapped approach, and the latent approach. The CP decomposition is computed by the tensorlab [Sorber et al., 2014] with 20 random initializations. We assume CP knows the true rank is 2. For the subspace norm, we use Algorithm 2 described in Section 3. We also select the top $2$ singular vectors when constructing $\hat{A}^{(k)}$’s. We computed the solutions for 20 values of regularization parameter $\lambda$ logarithmically spaced between $1$ and $100$. For the overlapped and the latent norm, we use ADMM described in [Tomioka et al., 2011a]; we also computed 20 solutions with the same $\lambda$’s used for the subspace norm.

We measure the performance in the relative error defined as $\|\hat{X} - X^*\|_F/\|X^*\|_F$. We report the minimum error obtained by choosing the optimal regularization parameter or the optimal initialization. Although the regularization parameter could be selected by leaving out some entries and measuring the error on these entries, we will not go into tensor completion here for the sake of simplicity.

Figure 3 shows the result of this experiment. The left panel shows the relative error for 3 representative values of $\lambda$ for the subspace norm. The black dash-dotted line shows the minimum error across all the $\lambda$’s. The magenta dashed line shows the error corresponding to the theoretically motivated choice $\lambda = \sigma (\max_k (\sqrt{m_k} + \sqrt{n_k}) + \sqrt{2 \log(K)})$ for each $\sigma$. The two vertical lines are thresholds of $\sigma$ from Corollary 1 corresponding to $\beta_1$ and $\beta_2$, namely, $\beta_1/(\prod_i n_i)^{1/4}$ and $\beta_2/(\prod_i n_i)^{1/4}$. It confirms that there is a rather sharp increase in the error around the theoretically predicted places (see also Figure 2). We can also see that the optimal $\lambda$ should grow linearly with $\sigma$. For large $\sigma$ (small SNR), the best relative error is $\sqrt{R}$ since the optimal choice of the regularization parameter $\lambda$ leads to predicting with $\hat{X} = 0$.

The right panel compares the performance of the subspace norm to other approaches. For each method the smallest error corresponding to the optimal choice of the regularization parameter $\lambda$ is shown. In addition, to place the numbers in context, we plot the line corresponding to

$$\text{Relative error} = \frac{\sqrt{R} \sum_n n_k \log(K)}{\|X^*\|_F} \cdot \sigma,$$  \hspace{1cm} (9)

which we call “optimistic”. This can be motivated from considering the (non-tractable) maximum likelihood estimator for CP decomposition [4]. The scaling for the maximum likelihood estimator is presented in Appendix A.

Clearly, the error of CP, the subspace norm, and “optimistic” grows at the same rate, much slower than overlap and latent. The error of CP increases beyond 1, as no regularization is imposed. The point when the optimistic error reaches 1 also gives us a critical standard derivation $\sigma_c = \frac{\|X^*\|_F}{\sqrt{R} \sum_n n_k \log(K)}$. One can see the points when CP or the
subspace norm reaches relative error 1 are close to $\sigma_c$. Thus we can see that both CP and the subspace norm are behaving near optimally in this setting, although such behavior is guaranteed for the subspace norm whereas it is hard to give any such guarantee for the CP decomposition based on nonlinear optimization.

### 4.2 Amino acids data

The amino acid dataset \cite{bro1997} is a semi-realistic dataset commonly used as a benchmark for low rank tensor modeling. It consists of five laboratory-made samples, each one contains different amounts of tyrosine, tryptophan and phenylalanine. The spectrum of their excitation wavelength (250-300 nm) and emission (250-450 nm) are measured by fluorescence, which gives a $5 \times 201 \times 61$ tensor. As the true factors are known to be these three acids, this data perfectly suits the CP model.

As for the synthetic dataset, we add random Gaussian noise with standard deviation $\sigma$ to the ground truth. The values of $\sigma$ are linearly spaced between 130 and 2020. We also fed the true rank into CP and the subspace approach. We computed the solutions of CP for 20 different random initializations, and the solutions of other approaches with 20 different values of $\lambda$. For the subspace and the overlapped approach, $\lambda$’s are logarithmically spaced between $10^3$ and $10^5$. For the latent approach, $\lambda$’s are logarithmically spaced between $10^4$ and $10^6$. Again, we include the optimistic scaling \cite{acar2011} to put the numbers in context.

Figure 4 shows the smallest relative error achieved by all methods we compare. Similar to the synthetic data, both CP and the subspace norm behaves near ideally, though the relative error of CP can be larger than 1 due to the lack of regularization. The performance of the overlap norm is slightly worse and the latent norm is the worst.

### References

E. Acar, D. M. Dunlavy, T. G. Kolda, and M. Mørup. Scalable tensor factorizations for incomplete data. Chemometrics and Intelligent Laboratory Systems, 106(1):41–56, 2011.
Figure 4: Comparison of tensor denoising on amino acids data.

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.

A. Anandkumar, R. Ge, D. Hsu, S. M. Kakade, and M. Telgarsky. Tensor decompositions for learning latent variable models. *J. Mach. Learn. Res.*, 15(1):2773–2832, 2014.

F. Benaych-Georges and R. R. Nadakuditi. The eigenvalues and eigenvectors of finite, low rank perturbations of large random matrices. *Advances in Mathematics*, 227(1):494–521, 2011.

R. Bro. Parafac. tutorial and applications. *Chemometrics and intelligent laboratory systems*, 38(2):149–171, 1997.

J. Cardoso. Super-symmetric decomposition of the fourth-order cumulant tensor. blind identification of more sources than sensors. In *Acoustics, Speech, and Signal Processing, 1991. ICASSP-91., 1991 International Conference on*, pages 3109–3112, 1991.

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

L. De Lathauwer, B. De Moor, and J. Vandewalle. On the best rank-1 and rank-(R1, R2, . . . , RN) approximation of higher-order tensors. *SIAM J. Matrix Anal. Appl.*, 21(4):1324–1342, 2000b.

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.

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

C. J. Hillar and L.-H. Lim. Most tensor problems are np-hard. *Journal of the ACM*, 60(6):45, 2013.

J. Hästad. Tensor rank is NP-complete. *Journal of Algorithms*, 11(4):644–654, 1990.
D. Hsu, S. M. Kakade, and T. Zhang. Robust matrix decomposition with sparse corruptions. *Information Theory, IEEE Transactions on*, 57(11):7221–7234, 2011.

P. Jain and S. Oh. Provable tensor factorization with missing data. In *Adv. Neural. Inf. Process. Syst.* 27, pages 1431–1439, 2014.

R. Kiros, R. Zemel, and R. R. Salakhutdinov. A multiplicative model for learning distributed text-based attribute representations. In *Advances in Neural Information Processing Systems*, pages 2348–2356, 2014.

T. G. Kolda. Orthogonal tensor decompositions. *SIAM Journal on Matrix Analysis and Applications*, 23(1):243–255, 2001.

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

T. G. Kolda and J. R. Mayo. Shifted power method for computing tensor eigenpairs. *SIAM Journal on Matrix Analysis and Applications*, 32(4):1095–1124, 2011.

B. Laurent and P. Massart. Adaptive estimation of a quadratic functional by model selection. *Annals of Statistics*, pages 1302–1338, 2000.

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

M. Mørup. Applications of tensor (multiway array) factorizations and decompositions in data mining. *Wiley Interdisciplinary Rev.: Data Min. Knowl. Discov.*, 1(1):24–40, 2011.

C. Mu, B. Huang, J. Wright, and D. Goldfarb. Square deal: Lower bounds and improved relaxations for tensor recovery. In *Proc. ICML ’14*, 2014.

E. Richard and A. Montanari. A statistical model for tensor pca. In *Advances in Neural Information Processing Systems*, pages 2897–2905, 2014.

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.

A. Smilde, R. Bro, and P. Geladi. *Multi-way analysis: applications in the chemical sciences*. John Wiley & Sons, 2005.

L. Sorber, M. Van Barel, and L. De Lathauwer. Optimization-based algorithms for tensor decompositions: Canonical polyadic decomposition, decomposition in rank-(l1,r1,l2,r2,l3,r3) terms, and a new generalization. *SIAM Journal on Optimization*, 23(2):695–720, 2013.

L. Sorber, M. Van Barel, and L. T. De Lathauwer. v2. 0. *Available online, URL: http://www.tensorlab.net*, 2014.

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.

R. Tomioka and T. Suzuki. Convex tensor decomposition via structured Schatten norm regularization. In C. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K. Weinberger, editors, *Adv. Neural. Inf. Process. Syst.* 26, pages 1331–1339. 2013.

R. Tomioka and T. Suzuki. Spectral norm of random tensors. Technical report, arXiv:1407.1870, 2014.

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

R. Tomioka, T. Suzuki, K. Hayashi, and H. Kashima. Statistical performance of convex tensor decomposition. In *Adv. Neural. Inf. Process. Syst.* 24, pages 972–980. 2011b.

R. Vershynin. Introduction to the non-asymptotic analysis of random matrices. Technical report, arXiv:1011.3027, 2010.
P.-Å. Wedin. Perturbation bounds in connection with singular value decomposition. *BIT Numerical Mathematics*, 12(1):99–111, 1972.

K. Wimalawarne, M. Sugiyama, and R. Tomioka. Multitask learning meets tensor factorization: task imputation via convex optimization. In Z. Ghahramani, M. Welling, C. Cortes, N. Lawrence, and K. Weinberger, editors, *Adv. Neural. Inf. Process. Syst.* 27, pages 2825–2833. Curran Associates, Inc., 2014.
A Maximum likelihood estimator

Let $Y \in \mathbb{R}^{n_1 \times \cdots \times n_K}$ be a noisy observed tensor generated as follows:

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

where $\mathcal{X}^*$ can be factorized into $R$ rank one terms as in (4) and $\mathcal{E}$ is a noisy tensor whose entries are i.i.d. normal $\mathcal{N}(0,1)$.

Let $\hat{\mathcal{X}}_{\text{MLE}}$ be the (intractable) estimator defined as

$$\hat{\mathcal{X}}_{\text{MLE}} = \arg \min_{\mathcal{X}} \left( \|Y - \mathcal{X}\|_F^2 \text{ s.t. } \mathcal{X} \text{ is rank at most } R \right).$$

We have the following performance guarantee for $\hat{\mathcal{X}}_{\text{MLE}}$:

**Theorem 3.** Let $R \leq \min_k n_k/2$. Then there is a constant $c$ such that

$$\|\hat{\mathcal{X}}_{\text{MLE}} - \mathcal{X}^*\|_F \leq c \sigma \sqrt{R^K \sum_{k=1}^K n_k \log(2K/K_0) + \log(2/\delta)},$$

with probability at least $1 - \delta$, where $K_0 = \log(3/2)$.

Note that the factor $R^K$ in the square root is rather conservative. In the best case, this factor reduces to linear in $R$ and this is what we present in Section 4 as “optimistic” ignoring constants and $\delta$; see Eq. (9).

**Proof of Theorem 3.** Since $\hat{\mathcal{X}}_{\text{MLE}}$ is a minimizer and $\mathcal{X}^*$ is also feasible, we have

$$\|Y - \hat{\mathcal{X}}_{\text{MLE}}\|_F^2 \leq \|Y - \mathcal{X}^*\|_F^2,$$

which implies

$$\|\mathcal{X}^* - \hat{\mathcal{X}}_{\text{MLE}}\|_F^2 \leq \sigma \langle \mathcal{E}, \hat{\mathcal{X}}_{\text{MLE}} - \mathcal{X}^* \rangle \leq \sigma \|\mathcal{E}\|_\text{op} \|\hat{\mathcal{X}}_{\text{MLE}} - \mathcal{X}^*\|_{\text{nuc}},$$

where

$$\|\mathcal{X}\|_\text{op} := \sup_{u(1), \ldots, u(K)} \left\{ \sum_{i_1, i_2, \ldots, i_K} \mathcal{X}_{i_1, i_2, \ldots, i_K} u_{i_1}^{(1)} u_{i_2}^{(2)} \cdots u_{i_K}^{(K)} : \|u^{(1)}\| = \|u^{(2)}\| = \cdots = \|u^{(K)}\| = 1 \right\}$$

is the tensor spectral norm and the nuclear norm

$$\|\mathcal{X}\|_{\text{nuc}} := \inf_{u(1), \ldots, u(K)} \left\{ \sum_r \|u_r^{(1)}\| \cdot \|u_r^{(2)}\| \cdots \|u_r^{(K)}\| : \mathcal{X} = \sum_{r=1}^R u_r^{(1)} \circ \cdots \circ u_r^{(K)} \right\}$$

is the dual of the spectral norm.

Since both $\hat{\mathcal{X}}_{\text{MLE}}$ and $\mathcal{X}^*$ are rank at most $R$, the difference $\hat{\mathcal{X}}_{\text{MLE}} - \mathcal{X}^*$ is rank at most $2R$. Moreover, any rank-$R$ CP decomposition with $R \leq \min_k n_k$ can be reduced to an orthogonal CP decomposition with rank at most $R^K$ via the Tucker decomposition (Kolda, 2001). Thus, denoting this orthogonal decomposition by $\hat{\mathcal{X}}_{\text{MLE}} - \mathcal{X}^* = \sum_{r=1}^{R^K} \tilde{u}_r^{(1)} \circ \cdots \circ \tilde{u}_r^{(K)}$ and using $\beta_r := \|\tilde{u}_r^{(1)}\| \cdots \|\tilde{u}_r^{(K)}\|$, we have

$$\|\hat{\mathcal{X}}_{\text{MLE}} - \mathcal{X}^*\|_{\text{nuc}} \leq \sum_{r=1}^{R^K} \beta_r \leq \sqrt{R^K} \sqrt{\sum_{r=1}^{R^K} \beta_r^2} = \sqrt{R^K} \|\hat{\mathcal{X}}_{\text{MLE}} - \mathcal{X}^*\|_F,$$

where the last equality follows because the decomposition is orthogonal.

Finally applying the tail bound for the spectral norm of random Gaussian tensor (Tomioka and Suzuki, 2014), we obtain what we wanted.
B Proof of Theorem

We consider the second moment of $\tilde{X}$:

$$\tilde{X} \tilde{X}^\top = \beta^2 uu^\top + \sigma^2 E E^\top + \beta \sigma (uv^\top E^\top + Evu^\top)$$

$$= \beta^2 uu^\top + \sigma^2 I + G \sigma^2 EE^\top - \sigma^2 I + \beta \sigma (uv^\top E^\top + Evu^\top).$$

The eigenvalue decomposition of $B$ can be written as

$$B = [u \ U_2] \begin{bmatrix} \beta^2 + m \sigma^2 & \sigma^2 I \\ m \sigma^2 I & \sigma^2 I \end{bmatrix} [u \ U_2]^\top.$$

We first show a deterministic lower bound for $|\langle \hat{u}, u \rangle|$ assuming $\beta^2 \geq 2\|G\|$, where $\hat{u}$ is the leading eigenvector of $\tilde{X} \tilde{X}^\top$. Then we bound the spectral norm $\|G\|$ of the noise term (Lemma 3) and derive the sufficient condition for $\beta$.

Let $\hat{u}$ be the leading eigenvector of $\tilde{X} \tilde{X}^\top$ with eigenvalue $\hat{\lambda}$, $r = B \hat{u} - \hat{\lambda} \hat{u} = -G \hat{u}$. We have $U_2^\top r = (m \sigma^2 - \hat{\lambda})U_2^\top \hat{u}$. Hence, for all $\beta^2 > 2\|G\|$, it holds that

$$|\sin(\hat{u}, u)| = ||U_2^\top \hat{u}||_2 = \frac{||U_2^\top r||_2}{\sqrt{\lambda - m \sigma^2}} \leq \frac{\|G\|}{\beta^2 - \|G\|} \leq \frac{2\|G\|}{\beta^2},$$

where we used $||U_2^\top r||_2 = ||U_2^\top G \hat{u}||_2 \leq \|G\|$, and $\hat{\lambda} \geq u^\top \tilde{X} \tilde{X}^\top u \geq \beta^2 + m \sigma^2 - \|G\|$. Therefore,

$$|\langle \hat{u}, u \rangle| = |\cos(\hat{u}, u)| \geq \sqrt{1 - \frac{4\|G\|^2}{\beta^4}} \geq 1 - \frac{4\|G\|^2}{\beta^4},$$

if $\beta^2 \geq 2\|G\|$.

It follows from Lemma 3 (shown below) that

$$\|G\| \leq \begin{cases} 2\bar{C}\sigma^2 \sqrt{mn}, & \text{if } \beta/\sigma < \sqrt{m}, \\ 2\bar{C}\beta \sigma \sqrt{n}, & \text{otherwise}, \end{cases}$$

where $\bar{C}$ is a universal constant with probability at least $1 - 4e^{-n}$.

Now consider the first case ($\beta/\sigma < \sqrt{m}$) and assume $\beta^2 \geq 4\bar{C}\sigma^2 \sqrt{mn} \geq 2\|G\|$. Note that this case only arises when $\sqrt{m} \geq 4\bar{C}\sqrt{n}$. Denoting $\bar{C} = 16\bar{C}^2$, we obtain the first case in the theorem. Next, consider the second case ($\beta/\sigma \geq \sqrt{m}$). If $\sqrt{m} \geq 4\bar{C}\sqrt{n}$ as above, we have $\beta/\sigma \geq 4\bar{C}\sqrt{n}$, which implies $\beta^2 \geq 2\|G\|$ and we obtain the second case in the theorem. On the other hand, if $\sqrt{m} < 4\bar{C}\sqrt{n}$, we require $\beta/\sigma \geq 4\bar{C}\sqrt{n}$ to obtain the last case in the theorem.

**Lemma 3.** Let $G$ be constructed as in Theorem 4 If $m \geq n$, there exists a universal constant $\bar{C}$ such that

$$\|G\| \leq \bar{C}\sigma^2 \left(\sqrt{mn} + \sqrt{n(\beta/\sigma)^2}\right),$$

with probability at least $1 - 4e^{-n}$.

**Proof.** The proof is an $\varepsilon$-net argument. Let

$$\lambda = 2\sigma^2 \left(\sqrt{4mn} + 4n + \sqrt{8n(\beta/\sigma)^2}\right).$$

The goal is to control $|x^\top Gx|$ for all the vectors $x$ on the unit Euclidean sphere $S^{n-1}$. In order to do this, we first bound the probability of the tail event $|x^\top Gx| > \lambda$, for any fixed $x \in S^{n-1}$. Then we bound the probability that
$|x^\top Gx| > \lambda$ for all the vectors in a $\varepsilon$-net $N_\varepsilon$. Finally, we establish the connection between $\sup_{x \in N_\varepsilon} |x^\top Gx|$ and $\|G\|$.

To bound $\mathbb{P}(|x^\top Gx| > \lambda)$ for a fix $x \in S^{n-1}$, we expand $x^\top Gx$ as

$$x^\top Gx = \sigma^2 \left( \|z\|^2 - m \right) + 2\beta\sigma (u^\top x)\gamma,$$

where $z = E^\top x$ and $\gamma = v^\top z$. Since $z \sim \mathcal{N}(0, I)$, we can see that $\|z\|^2$ is $\chi^2$ distributed with $m$ degrees of freedom and $\gamma \sim \mathcal{N}(0, 1)$.

First we bound the deviation of the $\chi^2$ term. By the corollary of Lemma 1 in Laurent and Massart (2000), we have

$$\mathbb{P}(\|z\|^2 - m > \lambda_1) \leq 2e^{-4n},$$

where $\lambda_1 = 2(\sqrt{4mn} + 4n)$. (10)

Next we bound the deviation of the Gaussian term. Using the Gaussian tail inequality, we have

$$\mathbb{P}(|\gamma| > \lambda_2) \leq 2e^{-4n},$$

where $\lambda_2 = \sqrt{8n}$. (11)

Combining inequalities (10) and (11), we have

$$\mathbb{P}(|x^\top Gx| > \lambda) \leq \mathbb{P}(\|z\|^2 - m > \lambda_1 \lor |\gamma| > \lambda_2) \leq \mathbb{P}(\|z\|^2 - m > \lambda_1) + \mathbb{P}(|\gamma| > \lambda_2) \leq 4e^{-4n},$$

where the second to last line follows from the union bound.

Furthermore, using Lemma 5.2 and 5.4 of Vershynin (2010), for any $\varepsilon \in [0, 1)$, it holds that

$$|N_{1/4}| \leq (1 + 2/\varepsilon)^n,$$

and

$$\|G\| \leq (1 - 2\varepsilon)^{-1} \sup_{x \in N_\varepsilon} |x^\top Gx|. $$

Taking the union bound over all the vectors in $N_{1/4}$, we obtain

$$\mathbb{P}\left( \sup_{x \in N_{1/4}} x^\top Gx > \lambda \right) \leq |N_{1/4}| 4e^{-4n} < 4e^{-n}.$$

Finally, the statement is obtained by noticing that $n \leq m$.

### C Proof of Theorem 2

First we decompose the error as

$$\|\mathcal{X}^* - \hat{X}\|_F \leq \|\mathcal{X}^* - \mathcal{X}_p\|_F + \|\mathcal{X}_p - \hat{X}\|_F.$$

The first term is an approximation error that depends on the choice of the subspace $S^{(k)}$. The second term corresponds to an estimation error and we analyze the second term below.

Since $\hat{X}$ is the minimizer of (8) and $\mathcal{X}_p$ is feasible,

$$\frac{1}{2} \|Y - \hat{X}\|^2_F + \lambda \sum_{k=1}^{K} \|M^{(k)}\|_* \leq \frac{1}{2} \|Y - \mathcal{X}_p\|^2_F + \lambda \sum_{k=1}^{K} \|M^{(k)}_p\|_* ,$$

17
from which we have
\[
\frac{1}{2} \| \mathcal{X}_p - \hat{\mathcal{X}} \|_F^2 \leq \| \mathcal{Y} - \mathcal{X}_p \|_s^* \| \mathcal{X}_p - \hat{\mathcal{X}} \|_s + \lambda \sum_{k=1}^{K} \left( \| \mathcal{M}_p^{(k)} \|_s - \| \hat{\mathcal{M}}^{(k)} \|_s \right).
\] (12)

Next we define \( \Delta_k := \hat{\mathcal{M}}^{(k)} - \mathcal{M}_p^{(k)} \in \mathbb{R}^{n_k \times m_k} \) and define its orthogonal decomposition \( \Delta_k = \Delta_k' + \Delta_k'' \) as
\[
\Delta_k' := (I_{n_k} - P_{U_p}) \Delta_k (I_{m_k} - P_{V_p}),
\]
where \( P_{U_p} \) and \( P_{V_p} \) are projection matrices to the column and row spaces of \( \mathcal{M}_p^{(k)} \), respectively, and \( \Delta_k := \Delta_k - \Delta_k'' \).

The above definition allows us to decompose \( \| \hat{\mathcal{M}}^{(k)} \|_s \) as follows:
\[
\| \hat{\mathcal{M}}^{(k)} \|_s = \| \mathcal{M}_p^{(k)} \|_s + \| \Delta_k' \|_s + \| \Delta_k'' \|_s.
\] (13)

Moreover,
\[
\| \mathcal{X}_p - \hat{\mathcal{X}} \|_s \leq \sum_{k=1}^{K} \| \Delta_k \|_s \leq \sum_{k=1}^{K} \left( \| \Delta_k' \|_s + \| \Delta_k'' \|_s \right)
\] (14)

Combining inequalities (12)–(14), we have
\[
\frac{1}{2} \| \mathcal{X}_p - \hat{\mathcal{X}} \|_F^2 \leq \left( \| \mathcal{Y} - \mathcal{X}_p \|_s^* + \lambda \right) \sum_{k=1}^{K} \| \Delta_k \|_s + \left( \| \mathcal{Y} - \mathcal{X}_p \|_s^* - \lambda \right) \sum_{k=1}^{K} \| \Delta_k'' \|_s.
\] (15)

Since
\[
\| \mathcal{Y} - \mathcal{X}_p \|_s^* \leq \sigma \| \mathcal{E} \|_s^* + \| \mathcal{X}^* - \mathcal{X}_p \|_s^*,
\]
if \( \lambda \geq \sigma \| \mathcal{E} \|_s^* + \| \mathcal{X}^* - \mathcal{X}_p \|_s^* \), the second term in the right-hand side of inequality (15) can be ignored and we have
\[
\frac{1}{2} \| \mathcal{X}_p - \hat{\mathcal{X}} \|_F^2 \leq 2 \lambda \sum_{k=1}^{K} \| \Delta_k \|_s
\]
\[
\leq 2 \lambda \sum_{k=1}^{K} \sqrt{2r_k} \| \Delta_k \|_F
\]
\[
\leq 2 \lambda \sum_{k=1}^{K} \sqrt{2r_k} \| \Delta_k \|_F
\]
\[
\leq 2 \sqrt{2} \lambda \sqrt{ \sum_{k=1}^{K} r_k \sum_{k=1}^{K} \| \Delta_k \|_F^2 },
\] (16)

where in the second line we used a simple observation that \( \text{rank}(\Delta_k') \leq 2r_k \).

Next, we relate the norm \( \| \mathcal{X}_p - \hat{\mathcal{X}} \|_F^2 \) to the sum \( \sum_{k=1}^{K} \| \Delta_k \|_F^2 \) in the right-hand side of inequality (16).

First suppose that \( \sum_{k=1}^{K} \| \Delta_k \|_F^2 \leq \| \mathcal{X}_p - \hat{\mathcal{X}} \|_F^2 \). Then from inequality (16), we have
\[
\| \mathcal{X}_p - \hat{\mathcal{X}} \|_F \leq \sqrt{2} \lambda \sqrt{ \sum_{k=1}^{K} r_k \sum_{k=1}^{K} \| \Delta_k \|_F^2 },
\]
by dividing both sides by \( \| \mathcal{X}_p - \hat{\mathcal{X}} \|_F \).

On the other hand, if \( \| \mathcal{X}_p - \hat{\mathcal{X}} \|_F^2 \leq \sum_{k=1}^{K} \| \Delta_k \|_F^2 \), we use the following lemma
Lemma 4. Suppose \( \{M_p^{(k)}\}_{k=1}^K, \{M^{(k)}\}_{k=1}^K \in \mathcal{M}(\rho) \), and \( S^{(k)} \) is constructed as a Kronecker product of \( K-1 \) ortho-normal matrices \( \hat{A}^{(k)} \) as \( S^{(k)} = \hat{A}^{(k-1)} \otimes \cdots \otimes \hat{A}^{(k+1)} \), where \( (\hat{A}^{(k)})^\top \hat{A}^{(k)} = I_m \) for \( \ell = 1, \ldots, K \). Then for \( X_p = \sum_{k=1}^K \text{fold}_k \left( M_p^{(k)} S^{(k)} \right)^\top \) and \( \hat{X} = \sum_{k=1}^K \text{fold}_k \left( M^{(k)} S^{(k)} \right)^\top \), the following inequality holds:

\[
\frac{1}{2} \sum_{k=1}^K \|\Delta_k\|_F^2 \leq \frac{1}{2} \|X_p - \hat{X}\|_F^2 + \rho \max_k (\sqrt{n_k} + \sqrt{m_k}) \sum_{k=1}^K \|\Delta_k\|_*.
\]  

(17)

Proof. The proof is presented in Section E.

Combining inequalities (15) and (17), we have

\[
\frac{1}{2} \sum_{k=1}^K \|\Delta_k\|_F^2 \leq \left( \|Y - X_p\|_* + \rho \max_k (\sqrt{n_k} + \sqrt{m_k}) + \lambda \right) \sum_{k=1}^K \|\Delta_k\|_* + \left( \|Y - X_p\|_* - \rho \max_k (\sqrt{n_k} + \sqrt{m_k}) - \lambda \right) \sum_{k=1}^K \|\Delta_k\|_*.
\]

Thus if we take \( \lambda \geq \sigma \|E\|_* + \|X_p\|_* + \rho \max_k (\sqrt{n_k} + \sqrt{m_k}) \), the second term in the right-hand side can be ignored and following the derivation leading to inequality (16) and dividing both sides by \( \sqrt{\sum_{k=1}^K \|\Delta_k\|_F^2} \), we have

\[
\|X_p - \hat{X}\|_F \leq \sqrt{\sum_{k=1}^K \|\Delta_k\|_F^2} \leq 4\sqrt{2\lambda} \sqrt{\sum_{k=1}^K r_k},
\]

where the first inequality follows from the assumption.

The final step of the proof is to bound the norm \( \|E\|_* \) with sufficiently high probability. By Lemma 1,

\[
\|E\|_* = \max_k \|E^{(k)} S^{(k)}\|.
\]

Therefore, taking the union bound, we have

\[
P \left( \max_k \|E^{(k)} S^{(k)}\| \geq t \right) \leq \sum_{k=1}^K P \left( \|E^{(k)} S^{(k)}\| \geq t \right).
\]

(18)

Now since each \( E^{(k)} S^{(k)} \in \mathbb{R}^{n_k \times m_k} \) is a random matrix with i.i.d. standard Gaussian entries,

\[
P \left( \|E^{(k)} S^{(k)}\| \geq \sqrt{n_k} + \sqrt{m_k} + t \right) \leq \exp(-t^2/(2\sigma^2)).
\]

Therefore, choosing \( t = \max_k (\sqrt{n_k} + \sqrt{m_k}) + \sqrt{2\log(K/\delta)} \) in inequality (18), we have

\[
\max_k \|E^{(k)} S^{(k)}\| \leq \max_k (\sqrt{n_k} + \sqrt{m_k}) + \sqrt{2\log(K/\delta)},
\]

with probability at least \( 1 - \delta \). Plugging this into the condition for the regularization parameter \( \lambda \), we obtain what we wanted.

D Proof of Lemma 2

Proof. i) Consider the Khatri-Rao product

\[
U^{(k-1)} \odot U^{(k-2)} = \begin{bmatrix}
    u_1^{(k-1)} \otimes u_1^{(k-2)} & \cdots & u_1^{(k-1)} \otimes u_R^{(k-2)} \\
    u_2^{(k-1)} \otimes u_1^{(k-2)} & \cdots & u_2^{(k-1)} \otimes u_R^{(k-2)} \\
    \vdots & \ddots & \vdots \\
    u_{n_k-1}^{(k-1)} \otimes u_1^{(k-2)} & \cdots & u_{n_k-1}^{(k-1)} \otimes u_R^{(k-2)}
\end{bmatrix}.
\]


It is easy to see this matrix has full column rank if both $U^{(k-1)}$ and $U^{(k-2)}$ have full rank. By applying this to $U^{(k-1)} \odot \cdots \odot U^{(k+1)}$ recursively, one can verify this sequence of Khatri-Rao products gives a matrix of rank $R$. Note that $C$ is a diagonal matrix with non-zero diagonal entries, so that $P^{(k)} = C \left( U^{(k-1)} \odot \cdots \odot U^{(k+1)} \right)^\top$ is rank $R$ and has a Moore–Penrose pseudo inverse $P^{(k)^\dagger}$ such that $P^{(k)} P^{(k)^\dagger} = I$. Therefore,

$$U^{(k)} = A^{(k)} \Lambda^{(k)} B^{(k)} P^{(k)^\dagger}.$$  

ii) Let $U^{(k-1)} = A^{(k-1)} Q^{(k-1)}$ and $U^{(k-2)} = A^{(k-2)} Q^{(k-2)}$. The columns of $U^{(k-1)} \odot U^{(k-2)}$ is a subset of columns of $A^{(k-1)} \odot Q^{(k-2)}$. The latter is in the span of $A^{(k-1)} \odot A^{(k-2)}$ because $A^{(k-1)} Q^{(k-1)} \odot A^{(k-2)} Q^{(k-2)}$. Using this recursively proves the claim.

\[ \square \]

E Proof of Lemma 4

Expanding $X_p$ and $\tilde{X}$, we have

$$\|X_p - \tilde{X}\|_F^2 = \| \sum_{k=1}^K \text{fold}_k (\Delta_k S^{(k)^\top}) \|_F^2$$

$$\begin{align*}
&= \sum_{k=1}^K \| \Delta_k \|_F^2 + \sum_{k \neq \ell} \langle \text{fold}_k (\Delta_k S^{(k)^\top}), \text{fold}_{\ell} (\Delta_{\ell} S^{(\ell)^\top}) \rangle \\
&= \sum_{k=1}^K \| \Delta_k \|_F^2 + \sum_{k \neq \ell} \langle \text{fold}_k (\Delta_k) \times_{k' \neq k} \tilde{A}^{(k')}, \text{fold}_{\ell} (\Delta_{\ell}) \times_{\ell' \neq \ell} \tilde{A}^{(\ell')} \rangle \\
&= \sum_{k=1}^K \| \Delta_k \|_F^2 + \sum_{k \neq \ell} \langle \text{fold}_k (\Delta_k) \times_{\ell} \tilde{A}^{(\ell)}, \text{fold}_{\ell} (\Delta_{\ell}) \times_k \tilde{A}^{(k)} \rangle \\
&= \sum_{k=1}^K \| \Delta_k \|_F^2 + \sum_{k \neq \ell} \langle \Delta_k (I_{\tilde{m}_1} \otimes \cdots \otimes \tilde{A}^{(\ell)} \otimes \cdots \otimes I_{\tilde{m}_K})^\top, \tilde{A}^{(k)} (\text{fold}_{\ell} (\Delta_{\ell}))_{(k)} \rangle \\
&\geq \sum_{k=1}^K \| \Delta_k \|_F^2 - \sum_{k \neq \ell} \| \Delta_k \|_2 \| (\text{fold}_{\ell} (\Delta_{\ell}))_{(k)} \| \\
&\geq \sum_{k=1}^K \| \Delta_k \|_F^2 - 2p \max_k (\sqrt{m_k} + \sqrt{n_k}) \sum_{k=1}^K \| \Delta_k \|_2,
\end{align*}$$

from which the lemma holds. Here we regarded $\text{fold}_k (\Delta_k S^{(k)})$ as a Tucker decomposition with the core tensor $\text{fold}_k (\Delta_k)$ and factor matrices $\tilde{A}^{(k')}$ for $k' \neq k$. Most of the factors except for $k$ and $\ell$ cancel out when calculating the inner product between two such tensors in the third line, because $(\tilde{A}^{(k')})^\top \tilde{A}^{(k')} = I_{\tilde{m}_k}$. After unfolding the inner product at the $k$th mode in the fifth line, we notice that a multiplication by an ortho-normal matrix does not affect the nuclear norm or the spectral norm. In the last line we used $\{ \Delta_k \}_{k=1}^K \in \mathcal{M}(2p)$, which follows from the assumption that both $\{ M_p^{(k)} \}_{k=1}^K, \{ \tilde{M}^{(k)} \}_{k=1}^K \in \mathcal{M}(p)$. 

\[ \square \]