E-Companion to Learning Preferences with Side Information

A. Survey of works employing user-item interaction data

Table 2 summarizes the different forms of user-item interaction data that have previously been studied.

Activity	Interaction	Representative articles
Direct Feedback	Numerical ratings	Breese et al. (1998)
		Ansari et al. (2000)
		Herlocker et al. (2002)
	Text reviews	Hu and Liu (2004)
		Das and Chen (2007)
		Archak et al. (2011)
Purchases	In-store sales transactions	Fader and Hardie (1996)
		Chong et al. (2001)
		Chintagunta et al. (2005)
		Chintagunta and Dube (2005)
	Online sales transactions	Bodapati (2008)
		Moon and Russell (2008)
In-site actions	Add to cart, search	Wu and Rangaswamy (2003)
	Product view	Moe (2006)
	Clicks on email links sent by site	Ansari and Mela (2003)
	Product customization	Sismeiro and Bucklin (2004)
	Browsing	Bucklin and Sismeiro (2003)
		Montgomery et al. (2004)
		Besbes et al. (2015)
Outside online	Twitter, Google, Wiki, IMDB	Liu et al. (2016)
	Blogs	Gopinath et al. (2013)
	Browsing	Trusov et al. (2016)
	Tagging	Ghose et al. (2012)
Physical actions	Try-on, facial expressions	Lu et al. (2016)
	Movement, direction faced, gaze	Hui et al. (2013)
Usage	Song listening time	Chung et al. (2009)

Table 2: List of user-item interactions.

A classic user-item interaction is users' direct feedback in the form of numerical ratings. Numerical ratings have been the traditional subject of study for recommender system researchers (Breese et al. (1998), Herlocker et al. (2002) and Ansari et al. (2000)). The prototypical example of this is the Netflix Prize competition (Bennett and Lanning (2007)), where the data consisted of users' movie grades on a scale from 1 to 5. Advances in text mining techniques have also allowed analysis of users' text reviews; see Hu and Liu (2004) and Archak et al. (2011).

Bodapati (2008) points out that 'in real-world systems, explicit self-reports of ratings are not observed as frequently as behavioral data in the form of purchases.' Along these lines, another well-studied type of interaction is purchases. Sales transaction data at the customer-product granularity has existed for some

time now in many forms, e.g. customer transactions have been tracked in brick-and-mortar retail with the use of scanning devices and loyalty programs. There have been studies dealing with this type of data; for example, Chong et al. (2001), Fader and Hardie (1996), Chintagunta et al. (2005), and Chintagunta and Dube (2005) all analyze purchase data among households at brick-and-mortar stores. Online retail has made sales transactions even easier to record. Both Bodapati (2008) and Moon and Russell (2008) use online purchase data to make product recommendations.

Beyond purchases, advanced tracking software allows for all types of online behavior to be recorded. Within a business' website, a variety of user-item interactions may be recorded, including adding an item to a virtual shopping cart (Wu and Rangaswamy (2003)), viewing an item's page (Moe (2006)), and clicking on personalized email links (Ansari and Mela (2003)). These interactions also extend beyond a business' own website into general online behavior, including user-generated content such as blogs (Gopinath et al. (2013)) and social media (Liu et al. (2016)).

Increasingly sophisticated technology has even allowed for collection and analysis of new kinds of data in the brick-and-mortar setting. Data is generated for example through cell phone tracking and in-store video: Macy's encourages shoppers to scan products through their mobile app (MobileCommerceDaily (2013)), and video can be used to record when customers slow down and look at a product (Hui et al. (2013)); Lu et al. (2016) were even able to record and analyze customers' facial expressions while trying on clothing items.

B. Comparison of slice rank to existing tensor ranks

There are many definitions of rank for tensors that have already been studied. The two most common, which tensor recovery has focused on, are referred to here as CP rank and Tucker rank. We review the canonical definitions of these ranks here. See Kolda and Bader (2009) for a more thorough treatment of these concepts.

CP rank The CP rank of a tensor relates to its orthogonal decompositions. A rank-one tensor is any tensor $M \in \mathbb{R}^{m \times m \times n}$ that is the outer product of three vectors, i.e. $M = u \otimes v \otimes w$ for some $u \in \mathbb{R}^m$, $v \in \mathbb{R}^m$, and $w \in \mathbb{R}^n$, or equivalently, $M^k_{i,j} = u_i v_j w_k$. For any tensor M, we denote its CP rank as $\text{CP}(M)$, which is the minimum number r such that M can be expressed as the sum of r rank-one tensors.

Tucker rank The Tucker rank of a tensor M, denoted $\text{Tucker}(M)$, is the vector (r_1, r_2, r_3), where r_d is the rank of its mode-d unfolding. This relates to its higher order singular value decomposition: given a tensor of Tucker rank (r_1, r_2, r_3), there exist vectors $u^1, \dots, u^{r_1} \in \mathbb{R}^m$, $v^1, \dots, v^{r_2} \in \mathbb{R}^m$, and $w^1, \dots, w^{r_3} \in \mathbb{R}^n$, and a smaller tensor $S \in \mathbb{R}^{r_1 \times r_2 \times r_3}$, such that $M = \sum_{\ell_1=1}^{r_1} \sum_{\ell_2=1}^{r_2} \sum_{\ell_3=1}^{r_3} S^{\ell_3}_{\ell_1,\ell_2} u^{\ell_1} \otimes v^{\ell_2} \otimes w^{\ell_3}$.

Proposition 1 establishes that the slice rank is a less restrictive measure of complexity than either of these two rank definitions. Recall that $\mathcal{CP}(r)$ is the set of tensors with CP rank at most r, $\text{Tucker}(r, r, l)$ is the set of tensors whose Tucker rank is component-wise at most (r, r, l), and $\text{Slice}(r)$ is the set of tensors whose slice rank is at most r.

Proof of Proposition 1. We first prove that $\mathcal{CP}(r) \subseteq \text{Slice}(r)$. Suppose $M \in \mathcal{CP}(r)$. By definition there exist vectors $u^1, \dots, u^r \in \mathbb{R}^m$, $v^1, \dots, v^r \in \mathbb{R}^m$, and $w^1, \dots, w^r \in \mathbb{R}^n$, such that each entry of M can be expressed as

$$M^k_{i,j} = \sum_{\ell=1}^{r} u^\ell_i v^\ell_j w^\ell_k.$$

Let U and V be the matrices with columns $u^1, \ldots, u^r$ and $v^1, \ldots, v^r$, respectively. Then we can equivalently write the above expression in matrix form for slices as

$$M^k = \sum_{\ell=1}^{r} w_k^\ell \left(u^\ell v^{\ell\top}\right) = US^kV^\top,$$

where S^k is the diagonal matrix whose diagonal elements are $w_k^1, \ldots, w_k^r$. This equivalent expression for the slices of M reveals that the slice rank of M is at most r, and so $M \in \text{Slice}(r)$.

Now we prove that $\text{Tucker}(r, r, l) \subseteq \text{Slice}(r)$. Suppose $M \in \text{Tucker}(r, r, l)$. By definition there exist vectors $u^1, \ldots, u^r \in \mathbb{R}^m$, $v^1, \ldots, v^r \in \mathbb{R}^m$, and $w^1, \ldots, w^l \in \mathbb{R}^n$, and a tensor $S \in \mathbb{R}^{r\times r\times l}$, such that $M = \sum_{\ell_1=1}^r \sum_{\ell_2=1}^r \sum_{\ell_3=1}^l S_{\ell_1,\ell_2}^{\ell_3} u^{\ell_1} \otimes v^{\ell_2} \otimes w^{\ell_3}$. Equivalently, each entry of M can be expressed as

$$M_{i,j}^k = \sum_{\ell_1=1}^{r} \sum_{\ell_2=1}^{r} \sum_{\ell_3=1}^{l} S_{\ell_1,\ell_2}^{\ell_3} u_i^{\ell_1} v_j^{\ell_2} w_k^{\ell_3}.$$

Let U and V be the matrices with columns $u^1, \ldots, u^r$ and $v^1, \ldots, v^r$, respectively. Then we can equivalently write the above expression in matrix form for slices as

$$M^k = \sum_{\ell_3=1}^{l} w_k^{\ell_3} \sum_{\ell_1=1}^{r} \sum_{\ell_2=1}^{r} S_{\ell_1,\ell_2}^{\ell_3} \left(u^{\ell_1} v^{\ell_2\top}\right) = \sum_{\ell_3=1}^{l} w_k^{\ell_3} (US^{\ell_3}V^\top) = U\left(\sum_{\ell_3=1}^{l} w_k^{\ell_3} S^{\ell_3}\right) V^\top.$$

Once again, this equivalent expression for the slices of M reveals that the slice rank of M is at most r, and so $M \in \text{Slice}(r)$. ■

C. Proof of Proposition 2

Our proof follows a standard Bayesian argument for minimax lower bounds; for example, see the proof of Theorem 1.2 in Chatterjee (2014). We will separately show that $\text{MSE}(\hat{M}) \geq C(r^2/m^2)$ and $\text{MSE}(\hat{M}) \geq C(r/mn)$. We first give a detailed proof that $\text{MSE}(\hat{M}) \geq C(r^2/m^2)$. For each ground truth slice M^k, let the elements sitting in the first r rows and first r columns be drawn independently from a uniform distribution, and the remaining elements set equal to 0:

$$M_{ij}^k \sim \begin{cases} \text{Uniform}[0,1] & \text{if } i \leq r \text{ and } j \leq r \\ 0 & \text{if } i > r \text{ or } j > r \end{cases}. \tag{7}$$

Note that all slices of M share the same column and row spaces, both with dimension at most r. Finally, conditional on M, each entry $X_{i,j}^k$ of X is drawn from the following two point distribution:

$$X_{ij}^k \sim \text{Ber}(M_{ij}^k). \tag{8}$$

Then for each $i \leq r$ and $j \leq r$, we have

$$\begin{aligned}
\mathsf{E}\left[\mathrm{Var}\left(M_{ij}^k|X\right)\right] &= \mathrm{Var}\left(M_{ij}^k\right) - \mathrm{Var}\left(\mathsf{E}\left[M_{ij}^k|X\right]\right) \\
&= \mathrm{Var}\left(M_{ij}^k\right) - \mathrm{Var}\left(\mathsf{E}\left[M_{ij}^k|X_{ij}^k\right]\right) \\
&= \mathrm{Var}\left(M_{ij}^k\right) - \mathrm{Var}\left(\frac{1+X_{ij}^k}{3}\right) \\
&= \frac{1}{12} - \frac{1}{36} = \frac{1}{18} \qquad (9)
\end{aligned}$$

The first equality is the law of total variance. For the second equality, observe that M_{ij}^k is independent of all entries of X except for its corresponding entry X_{ij}^k. The third equality comes from the fact that, having defined M_{ij}^k to be distributed as Uniform$[0,1]$ (or equivalently Beta$(1,1)$), its distribution conditional on X_{ij}^k is Beta$(1+X_{ij}^k, 2-X_{ij}^k)$.

Then for any estimator $\hat{M}$, the definition of variance implies that

$$\mathsf{E}\left[\left(\hat{M}_{ij}^k - M_{ij}^k\right)^2 \middle| X\right] \geq \mathrm{Var}\left(M_{ij}^k|X\right).$$

Taking expectations of both sides, and applying (9), we have

$$\mathsf{E}\left[\left(\hat{M}_{ij}^k - M_{ij}^k\right)^2\right] \geq \frac{1}{18}.$$

The proof concludes by summing both sides over all entries of M in the first r rows and first r columns (i.e. nr^2 entries in total) and dividing by $m^2 n$.

A nearly identical argument shows that $\mathrm{MSE}(\hat{M}) \geq C(r/mn)$. For the first slice M^1, let the elements in the first r rows be drawn independently from a uniform distribution, and the remaining elements set equal to 0:

$$M_{ij}^1 \sim \begin{cases} \text{Uniform}[0,1] & \text{if } i \leq r \\ 0 & \text{if } i > r \end{cases}. \qquad (10)$$

Set the entries of the remaining slices equal to the corresponding entries in the first slice, i.e. $M_{ij}^k = M_{ij}^1$ for all k. Once again, conditional on M, each entry X_{ij}^k is drawn independently from the distribution in (8), so while the slices of M are copies of each other, the slices of X are not. Then for each $i \leq r$, we have

$$\begin{aligned}
\mathsf{E}\left[\mathrm{Var}\left(M_{ij}^k|X\right)\right] &= \mathrm{Var}\left(M_{ij}^k\right) - \mathrm{Var}\left(\mathsf{E}\left[M_{ij}^k|X\right]\right) \\
&= \mathrm{Var}\left(M_{ij}^k\right) - \mathrm{Var}\left(\mathsf{E}\left[M_{ij}^k|X_{ij}^1, \ldots X_{ij}^n\right]\right) \\
&= \mathrm{Var}\left(M_{ij}^k\right) - \mathrm{Var}\left(\frac{1+X_{ij}^1+\cdots+X_{ij}^n}{n+2}\right) \\
&= \frac{1}{12} - \frac{n}{12(n+2)} = \frac{1}{6(n+2)} \qquad (11)
\end{aligned}$$

Again, the first equality is the law of total variance, and for the second equality, observe that M_{ij}^k is independent of all entries of X except for the $(i,j)^{\text{th}}$ entry of each slice. For the third equality, the distribution of M_{ij}^k conditional on $X_{ij}^1, \ldots, X_{ij}^n$ is Beta$(1+X_{ij}^1+\cdots+X_{ij}^n, n+1-(X_{ij}^1+\cdots+X_{ij}^n))$.

Therefore, for any estimator $\hat{M}$ we have

$$\mathsf{E}\left[\left(\hat{M}_{ij}^k - M_{ij}^k\right)^2\right] \geq \frac{1}{6(n+2)}.$$

The proof concludes by summing both sides over all entries of M in the first r rows (i.e. nmr entries in total) and dividing by $m^2 n$. ■

D. Proofs of Theorems 1 and 2 and Corollary 1

Before proceeding with the proofs, it will be convenient to review and introduce some additional notation. For $n \in \mathbb{Z}^+$, $[n]$ denotes the set $\{1, \ldots, n\}$. If X is a matrix, then $\|X\|_2$, $\|X\|_F$, and $\|X\|_*$ are respectively the operator, frobenius, and nuclear norms of X. $\sigma_i(X)$ is the i^{th} largest singular value of X. For a matrix $U \in \mathbb{R}^{m \times r}$ with orthonormal columns, we will refer to U as a *matrix* and *subspace* interchangeably, where the subspace is the space in $\mathbb{R}^m$ spanned by the columns of U; $P_U = UU^\top$ is the projection operator onto the subspace U. We use $d(U, \hat{U}) = \left\|P_U - P_{\hat{U}}\right\|_F$ as a metric for subspaces.

It will suffice to prove Theorem 2; Theorem 1 is just the special case when $\delta = 0$. The proof of Theorem 2 involves two steps, corresponding to the two stages of the algorithm: learning subspaces and projection. In the first step, we show that we are able to closely estimate the column and row spaces, and in the second step, we show that if our estimates of the 'true' column and row spaces are close, then our estimate of each slice is close.

D.1. Step 1: Column and Row Space Estimation

To estimate the column space (and similarly the row space), we take the top column singular vectors of $X_{(1)} = M_{(1)} + \epsilon_{(1)}$, so it is important to understand the extent to which $\epsilon_{(1)}$ changes the singular vectors of $M_{(1)}$. Lemma 1 bounds the error of this step. The first result in Lemma 1 is an upper bound on $\mathsf{E}\left[d(U, \hat{U})^2\right]$, which is the expected error of our subspace estimate. Because of the decomposition we make later on, we also need to bound $\mathsf{E}\left[\left\|\epsilon^k\right\|_F^2 d(U, \hat{U})^2\right]$ for any slice of the noise tensor ϵ^k, which is complicated by the fact that ϵ^k and $d(U, \hat{U})$ are not independent. The second result in Lemma 1 controls this term.

Lemma 1. *Let $M \in \mathbb{R}^{m \times mn}$ be a matrix with column space $U \in \mathbb{R}^{m \times r}$. Suppose $\epsilon \in \mathbb{R}^{m \times mn}$ is a random matrix with independent elements, where each element ϵ_{ij} is mean-zero, and $\mathsf{E}[\epsilon_{ij}^2]$, $\mathsf{E}[\epsilon_{ij}^4]$, and $\mathsf{E}[\epsilon_{ij}^6]$ are respectively bounded by K_2, K_4, and K_6. Let $X = M + \epsilon$, and let $\hat{U} \in \mathbb{R}^{m \times r}$ be the column singular vectors of X corresponding to its largest r singular values. Then taking expectation over ϵ, we have*

$$\mathsf{E}\left[d(U, \hat{U})^2\right] \leq 24 \frac{4K_2 m \|M\|_F^2 + K_4 m^2 n + K_2^2 m^3 n + \min_\rho \left\|\mathsf{E}[\epsilon\epsilon^\top] - \rho I_m\right\|_F^2}{\sigma_r^4(M)}, \quad \textit{and}$$

$$\begin{aligned}\mathsf{E}\left[\left\|\epsilon^1\right\|_F^2 d(U, \hat{U})^2\right] \leq\ & 24m^2 \frac{4K_2^2 m \|M\|_F^2 + 3K_4 K_2 m^2 n + K_2^3 m^3 n + K_2 \min_\rho \left\|\mathsf{E}[\epsilon\epsilon^\top] - \rho I_m\right\|_F^2}{\sigma_r^4(M)} \\ & + 24m^2 \frac{4K_4 \left\|M^1\right\|_F^2 / m + K_6}{\sigma_r^4(M)},\end{aligned}$$

where M^1 and ϵ^1 are the $m \times m$ submatrices consisting of the first m columns of M and ϵ, respectively.

Note that M^1 and ϵ^1 in the statement of Lemma 1 can in fact be any $m \times m$ submatrices of M and ϵ; they are taken to be the first M columns to save on notation, and without loss of generality. The proof of Lemma 1 relies on the Davis-Kahan Theorem (Davis and Kahan (1970)), via a recent extension by Yu et al. (2015), which we reproduce as Lemma 2. Note that Lemma 2 is a statement about symmetric matrices, which we adapt to our setting where the matrices are not symmetric or even square; Yu et al. (2015) also show a version of Lemma 2 for rectangular matrices that is a similar modification to Wedin's Theorem (Wedin (1972)), but applying that directly would not yield as strong a bound as Lemma 1. However, this stronger bound requires the noise to be balanced.

Lemma 2 (Davis-Kahan Variant; Yu et al. (2015), Theorem 2)**.** *Suppose S and $\hat{S}$ are symmetric matrices, and let U and $\hat{U}$ be the eigenvectors corresponding to the r largest eigenvalues of S and $\hat{S}$, respectively. Let $\lambda_r(S)$ and $\lambda_{r+1}(S)$ be the r^{th} and $r+1^{th}$ largest eigenvalues of S. Then assuming $\lambda_r(S) \neq \lambda_{r+1}(S)$, we have*

$$d(U, \hat{U}) \leq \frac{2\sqrt{2}\left\|S - \hat{S}\right\|_F}{\lambda_r(S) - \lambda_{r+1}(S)}.$$

Proof of Lemma 1. First note that the column singular vectors of M and X are identical to the eigenvectors of $MM^\top$ and $XX^\top$, respectively, and further, the eigenvectors of $XX^\top - \rho I_m$ are the same for any $\rho \in \mathbb{R}$. Thus, Lemma 2 can be applied directly with $S = MM^\top$, and $\hat{S} = XX^\top - \rho I_m$ for any $\rho \in \mathbb{R}$, and $\lambda_r(MM^\top) - \lambda_{r+1}(MM^\top) = \sigma_r(M)^2 - \sigma_{r+1}(M)^2 = \sigma_r(M)^2$:

$$d(U, \hat{U})^2 \leq \frac{8 \min_\rho \left\|MM^\top - (XX^\top - \rho I_m)\right\|_F^2}{\sigma_r^4(M)}. \tag{12}$$

To upper bound the numerator, we make the following decomposition:

$$\begin{aligned}\min_\rho \left\|MM^\top - (XX^\top - \rho I_m)\right\|_F &\leq 2\left\|M\epsilon^\top\right\|_F + \min_\rho \left\|\epsilon\epsilon^\top - \rho I_m\right\|_F \\ &\leq 2\left\|M\epsilon^\top\right\|_F + \left\|\epsilon\epsilon^\top - \mathsf{E}[\epsilon\epsilon^\top]\right\|_F + \min_\rho \left\|\mathsf{E}[\epsilon\epsilon^\top] - \rho I_m\right\|_F,\end{aligned}$$

and since $(a+b+c)^2 \leq 3(a^2+b^2+c^2)$ for any $a, b, c \in \mathbb{R}$, we have

$$\min_\rho \left\|MM^\top - (XX^\top - \rho I_m)\right\|_F^2 \leq 3\left(4\left\|M\epsilon^\top\right\|_F^2 + \left\|\epsilon\epsilon^\top - \mathsf{E}[\epsilon\epsilon^\top]\right\|_F^2 + \min_\rho \left\|\mathsf{E}[\epsilon\epsilon^\top] - \rho I_m\right\|_F^2\right).$$

We have decomposed the numerator of (12) into three terms. The last term is a deterministic quantity. The proof concludes by bounding the expectation of the first two terms. All of the following calculations proceed in the same manner: the equalities come from rewriting expressions in expanded form and setting any summands with a lone $\mathsf{E}[\epsilon_{ij}]$ to zero, and the inequality applies the assumptions on the moments of the noise terms.

$$\begin{aligned}\mathsf{E}\left[\left\|M\epsilon^\top\right\|_F^2\right] &= \sum_{i_1 \in [m], i_2 \in [m]} \mathsf{E}\left[\left(\sum_{j \in [mn]} M_{i_1 j}\epsilon_{i_2 j}\right)^2\right] \\ &= \sum_{i_1 \in [m], i_2 \in [m]} \sum_{j \in [mn]} M_{i_1 j}^2 \mathsf{E}\left[\epsilon_{i_2 j}^2\right] \leq K_2 m \left\|M\right\|_F^2\end{aligned} \tag{13}$$

$$
\begin{aligned}
(14) \qquad \mathsf{E}\left[\left\|\epsilon^1\right\|_F^2 \left\|M\epsilon^\top\right\|_F^2\right] &= \mathsf{E}\left[\left\|\epsilon^1\right\|_F^2 \sum_{i_1,i_2\in[m]} \left(\sum_{j\in[mn]} M_{i_1 j}\epsilon_{i_2 j}\right)^2\right] \\
&= \sum_{i_1,i_2\in[m]} \sum_{j\in[mn]} M_{i_1 j}^2 \mathsf{E}\left[\epsilon_{i_2 j}^2 \left\|\epsilon^1\right\|_F^2\right] \\
&= \sum_{i_1,i_2\in[m]} \sum_{j\in[m]} M_{i_1 j}^2 \mathsf{E}\left[\epsilon_{i_2 j}^4\right] + \sum_{i_1,i_2,i_3\in[m]} \sum_{\substack{j_1\in[mn],j_2\in[m] \\ (i_2,j_1)\neq(i_3,j_2)}} M_{i_1 j_1}^2 \mathsf{E}\left[\epsilon_{i_2 j_1}^2\right] \mathsf{E}\left[\epsilon_{i_3 j_2}^2\right] \\
&\leq K_4 m \left\|M^1\right\|_F^2 + K_2^2 m^3 \left\|M\right\|_F^2
\end{aligned}
$$

$$
\begin{aligned}
(15) \qquad \mathsf{E}\left[\left\|\epsilon\epsilon^\top - \mathsf{E}[\epsilon\epsilon^\top]\right\|_F^2\right] &= \sum_{i\in[m]} \mathrm{Var}\left(\sum_{j\in[mn]} \epsilon_{ij}^2\right) + \sum_{\substack{i_1\in[m],i_2\in[m] \\ i_1\neq i_2}} \mathsf{E}\left[\left(\sum_{j\in[mn]} \epsilon_{i_1 j}\epsilon_{i_2 j}\right)^2\right] \\
&= \sum_{i\in[m]} \sum_{j\in[mn]} \mathrm{Var}\left[\epsilon_{ij}^2\right] + \sum_{\substack{i_1\in[m],i_2\in[m] \\ i_1\neq i_2}} \sum_{j\in[mn]} \mathsf{E}\left[\epsilon_{i_1 j}^2\right] \mathsf{E}\left[\epsilon_{i_2 j}^2\right] \\
&\leq K_4 m^2 n + K_2^2 m^3 n
\end{aligned}
$$

(16)

$$
\begin{aligned}
\mathsf{E}\left[\left\|\epsilon^1\right\|_F^2 \left\|\epsilon\epsilon^\top - \mathsf{E}[\epsilon\epsilon^\top]\right\|_F^2\right] &= \mathsf{E}\left[\left\|\epsilon^1\right\|_F^2 \sum_{i\in[m]} \left(\sum_{j\in[mn]} \epsilon_{ij}^2 - \mathsf{E}\left[\epsilon_{ij}^2\right]\right)^2 + \left\|\epsilon^1\right\|_F^2 \sum_{\substack{i_1,i_2\in[m] \\ i_1\neq i_2}} \left(\sum_{j\in[mn]} \epsilon_{i_1 j}\epsilon_{i_2 j}\right)^2\right] \\
&= \sum_{i\in[m]} \sum_{j\in[mn]} \mathsf{E}\left[\left\|\epsilon^1\right\|_F^2 \left(\epsilon_{ij}^2 - \mathsf{E}\left[\epsilon_{ij}^2\right]\right)^2\right] + \sum_{\substack{i_1,i_2\in[m] \\ i_1\neq i_2}} \sum_{j\in[mn]} \mathsf{E}\left[\left\|\epsilon^1\right\|_F^2 \epsilon_{i_1 j}^2 \epsilon_{i_2 j}^2\right] \\
&= \sum_{i\in[m]} \sum_{j\in[m]} \mathsf{E}\left[\epsilon_{ij}^2 \left(\epsilon_{ij}^2 - \mathsf{E}\left[\epsilon_{ij}^2\right]\right)^2\right] + \sum_{i_1,i_2\in[m]} \sum_{\substack{j_1\in[mn],j_2\in[m] \\ (i_1,j_1)\neq(i_2,j_2)}} \mathsf{E}\left[\epsilon_{i_2 j_2}^2\right] \mathsf{E}\left[\left(\epsilon_{i_1 j_1}^2 - \mathsf{E}\left[\epsilon_{i_1 j_1}^2\right]\right)^2\right] \\
&+ \sum_{\substack{i_1,i_2\in[m] \\ i_1\neq i_2}} \sum_{j\in[m]} 2\mathsf{E}\left[\epsilon_{i_1 j}^4\right] \mathsf{E}\left[\epsilon_{i_2 j}^2\right] + \sum_{\substack{i_1,i_2,i_3\in[m] \\ i_1\neq i_2}} \sum_{\substack{j_1\in[mn],j_2\in[m] \\ (i_3,j_2)\neq(i_1,j_1) \\ (i_3,j_2)\neq(i_2,j_1)}} \mathsf{E}\left[\epsilon_{i_3 j_2}^2\right] \mathsf{E}\left[\epsilon_{i_1 j_1}^2\right] \mathsf{E}\left[\epsilon_{i_2 j_1}^2\right] \\
&\leq K_6 m^2 + K_4 K_2 m^4 n + 2K_4 K_2 m^3 + K_2^3 m^5 n \\
&\leq K_6 m^2 + 3K_4 K_2 m^4 n + K_2^3 m^5 n
\end{aligned}
$$

Combining (13) and (15) completes the first result, and combining (14) and (16), along with the fact that $\mathsf{E}\left[\left\|\epsilon^1\right\|_F^2\right] \leq K_2 m^2$, completes the second.

■

D.2. Step 2: Projection onto Estimated Spaces

Lemma 3 decomposes the error of the projection step in terms of the error of our column and row space estimates. For any slice M^k, our estimate of this slice is the projection of X^k onto the estimated subspaces $\hat{U}$ and $\hat{V}$, i.e. $P_{\hat{U}} M^k P_{\hat{V}} + P_{\hat{U}} \epsilon^k P_{\hat{V}}$. If $\hat{U}$ and $\hat{V}$ are close to U and V, then $P_{\hat{U}} M^k P_{\hat{V}} \approx P_U M^k P_V = M^k$.

Furthermore, since $\hat{U}$ and $\hat{V}$ are low-dimensional subspaces, $P_{\hat{U}}\epsilon^k P_{\hat{V}}$ will be small (this argument needs to be made carefully as $\hat{U}$ and $\hat{V}$ depend on ϵ^k).

Lemma 3. *Let $M^1 \in \mathbb{R}^{m\times m}$ be a matrix with column and row spaces $U, V \in \mathbb{R}^{m\times r}$. Let $\epsilon^1 \in \mathbb{R}^{m\times m}$ be a random matrix, and let $\hat{U}, \hat{V} \in \mathbb{R}^{m\times r}$ be random subspaces, where none of these variables are required to be independent. If $\hat{M}^1 = P_{\hat{U}}(M^1 + \epsilon^1)P_{\hat{V}}$, then taking expectation over $\epsilon^1, \hat{U}$, and $\hat{V}$:*

$$\mathsf{E}\left[\left\|\hat{M}^1 - M^1\right\|_F^2\right] \le 9\mathsf{E}\left[\left\|P_U\epsilon^1 P_V\right\|_F^2\right] + 3\left\|M^1\right\|_F^2 \mathsf{E}\left[4d(U,\hat{U})^2 + d(V,\hat{V})^2\right] + 9\mathsf{E}\left[\left\|\epsilon^1\right\|_F^2\left(4d(U,\hat{U})^2 + d(V,\hat{V})^2\right)\right].$$

Proof of Lemma 3. We begin by making the following decomposition, where the first two inequalities rely on the sub-multiplicative and sub-additive properties of the frobenius norm, and the first inequality also relies on the fact that $\left\|P_{\hat{V}} - P_V\right\|_2 \le 1$. The final inequality comes from $(a+b+c)^2 \le 3(a^2+b^2+c^2)$.

$$\begin{aligned}
\left\|\hat{M}^1 - M^1\right\|_F^2 &= \left\|P_{\hat{U}}(M^1+\epsilon^1)P_{\hat{V}} - M^1\right\|_F^2 \\
&= \left\|[P_U + (P_{\hat{U}} - P_U)]M^1[P_V + (P_{\hat{V}} - P_V)] - M^1 + P_{\hat{U}}\epsilon^1 P_{\hat{V}}\right\|_F^2 \\
&= \left\|M^1(P_{\hat{V}} - P_V) + (P_{\hat{U}} - P_U)\left[M^1 + M^1(P_{\hat{V}} - P_V)\right] + P_{\hat{U}}\epsilon^1 P_{\hat{V}}\right\|_F^2 \\
&\le \left(\left\|M^1(P_{\hat{V}} - P_V)\right\|_F + 2\left\|(P_{\hat{U}} - P_U)M^1\right\|_F + \left\|P_{\hat{U}}\epsilon^1 P_{\hat{V}}\right\|_F\right)^2 \\
&\le \left(\left\|M^1\right\|_F\left(2d(U,\hat{U}) + d(V,\hat{V})\right) + \left\|P_{\hat{U}}\epsilon^1 P_{\hat{V}}\right\|_F\right)^2 \\
&\le 3\left\|M^1\right\|_F^2\left(4d(U,\hat{U})^2 + d(V,\hat{V})^2\right) + 3\left\|P_{\hat{U}}\epsilon^1 P_{\hat{V}}\right\|_F^2.
\end{aligned}$$

The last term is decomposed further in a similar way:

$$\begin{aligned}
\left\|P_{\hat{U}}\epsilon^1 P_{\hat{V}}\right\|_F^2 &= \left\|[P_U + (P_{\hat{U}} - P_U)]\epsilon^1[P_V + (P_{\hat{V}} - P_V)]\right\|_F^2 \\
&\le \left(\left\|P_U\epsilon^1 P_V\right\|_F + \left\|\epsilon^1(P_{\hat{V}} - P_V)\right\|_F + 2\left\|(P_{\hat{U}} - P_U)\epsilon^1\right\|_F\right)^2 \\
&\le \left(\left\|P_U\epsilon^1 P_V\right\|_F + \left\|\epsilon^1\right\|_F\left(2d(U,\hat{U}) + d(V,\hat{V})\right)\right)^2 \\
&\le 3\left\|P_U\epsilon^1 P_V\right\|_F^2 + 3\left\|\epsilon^1\right\|_F^2\left(4d(U,\hat{U})^2 + d(V,\hat{V})^2\right).
\end{aligned}$$

We conclude the proof by taking expectations. ■

D.3. Final Steps

We are now ready to conclude the proof. Fix any slice $k \in [n]$. Lemma 3 first gives us:

$$\mathsf{E}\left[\left\|\hat{M}^k - M^k\right\|_F^2\right] \le 9\mathsf{E}\left[\left\|P_U\epsilon^k P_V\right\|_F^2\right] + 3\left\|M^k\right\|_F^2 \mathsf{E}\left[4d(U,\hat{U})^2 + d(V,\hat{V})^2\right] + 9\mathsf{E}\left[\left\|\epsilon^k\right\|_F^2\left(4d(U,\hat{U})^2 + d(V,\hat{V})^2\right)\right].$$

The first term is bounded as follows:

$$\mathsf{E}\left[\left\|P_U\epsilon^k P_V\right\|_F^2\right] = \mathsf{E}\left[\left\|UU^\top\epsilon^k VV^\top\right\|_F^2\right] = \mathsf{E}\left[\left\|U^\top\epsilon^k V\right\|_F^2\right] = \sum_{i_1\in[r], i_2\in[r]} \mathsf{E}\left[\left(U_{i_1}^\top\epsilon^k V_{i_2}\right)^2\right] \le r^2 K_2.$$

The remaining terms are bounded by applying Lemma 1 directly. Replacing running constants with c, we have

$$\begin{aligned}\left\|M^k\right\|_F^2 \mathsf{E}\left[d(U,\hat{U})^2\right] &\leq 24\left\|M^k\right\|_F^2 \frac{4K_2 m\left\|M_{(1)}\right\|_F^2 + K_4 m^2 n + K_2^2 m^3 n + m\delta^2}{\sigma_r^4(M_{(1)})} \\ &\leq cm^2 \frac{(K_2+K_2^2)m^3 n + K_4 m^2 n + m\delta^2}{\gamma_M^2 m^4 n^2/r^2},\end{aligned}$$

$$\begin{aligned}\mathsf{E}\left[\left\|\epsilon^k\right\|_F^2 d(U,\hat{U})^2\right] &\leq 24m^2 \frac{4K_4\left\|M^k\right\|_F^2/m + 4K_2^2 m\left\|M_{(1)}\right\|_F^2 + K_6 + 3K_4K_2m^2n + K_2^3m^3n + K_2m\delta^2}{\sigma_r^4(M_{(1)})} \\ &\leq cm^2 \frac{K_4 m + (K_2^2+K_2^3)m^3 n + K_6 + K_4K_2m^2n + K_2m\delta^2}{\gamma_M^2 m^4 n^2/r^2},\end{aligned}$$

and similarly for $\mathsf{E}\left[d(V,\hat{V})^2\right]$ and $\mathsf{E}\left[\left\|\epsilon^k\right\|_F^2 d(V,\hat{V})^2\right]$. Note that in both calculations above, in the second inequality, we plug in our definition of γ_M, $\min_{i=1,2}\{\sigma_r^2(M_{(i)})\} \geq \gamma_M m^2 n/r$, and use the facts that $\left\|M^k\right\|_F^2 \leq m^2$ and $\left\|M_{(1)}\right\|_F^2 \leq m^2 n$.

Putting all of this together and rearranging terms, we have

$$\begin{aligned}\frac{1}{m^2}\mathsf{E}\left[\left\|\hat{M}^k - M^k\right\|_F^2\right] &\leq c\left[\frac{K_2 r^2}{m^2} + \frac{(K_2+K_2^2+K_2^3)m^3 n + K_6 + K_4(K_2+1)m^2 n + (K_2+1)m\delta^2}{\gamma_M^2 m^4 n^2/r^2}\right] \\ &\leq c\left[\frac{K_2 r^2}{m^2} + \frac{(K_2+K_2^2+K_2^3)r^2}{\gamma_M^2 mn} + \frac{K_6 r^2}{\gamma_M^2 m^4 n^2} + \frac{K_4(K_2+1)r^2}{\gamma_M^2 m^2 n} + \frac{(K_2+1)r^2\delta^2}{\gamma_M^2 m^3 n^2}\right] \\ &\leq c\left[\frac{K^2 r^2}{m^2} + \frac{K^2(K^4+1)r^2}{\gamma_M^2 mn} + \frac{(K^2+1)r^2\delta^2}{\gamma_M^2 m^3 n^2}\right].\end{aligned} \tag{17}$$

In the last step above, we consolidated terms by using the fact that K_2, K_4 and K_6 are respectively bounded by K^2, K^4 and K^6. Note that this entire analysis holds for any $k \in [n]$, so

$$\mathrm{SMSE}(\hat{M}) = \max_{k\in[n]} \frac{1}{m^2}\mathsf{E}\left[\left\|\hat{M}^k - M^k\right\|_F^2\right] \leq c\left[\frac{K^2 r^2}{m^2} + \frac{K^2(K^4+1)r^2}{\gamma_M^2 mn} + \frac{(K^2+1)r^2\delta^2}{\gamma_M^2 m^3 n^2}\right].$$

This concludes the proofs of Theorems 1 and 2. Corollary 1 follows by returning to (17), and applying (22) from Section G:

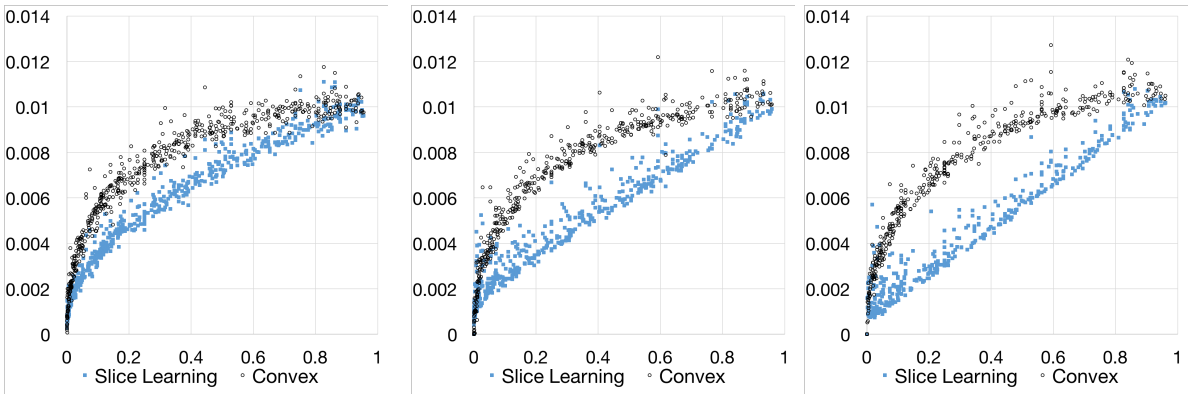
$$\begin{aligned}\frac{1}{m^2}\mathsf{E}\left[\left\|\hat{M}^k - M^k\right\|_F^2\right] &\leq c\left[\frac{K_2 r^2}{m^2} + \frac{(K_2+K_2^2+K_2^3)r^2}{\gamma_M^2 mn} + \frac{K_6 r^2}{\gamma_M^2 m^4 n^2} + \frac{K_4(K_2+1)r^2}{\gamma_M^2 m^2 n} + \frac{(K_2+1)r^2\delta^2}{\gamma_M^2 m^3 n^2}\right] \\ &\leq c\left[\frac{r^2}{m^2 p} + \frac{r^2}{\gamma_M^2 mnp^3} + \frac{r^2}{\gamma_M^2 m^4 n^2 p^5} + \frac{r^2}{\gamma_M^2 m^2 np^4} + \frac{r^2\delta^2}{\gamma_M^2 m^3 n^2 p}\right].\end{aligned}$$

■

E. Additional Synthetic Experiments

E.1. Additional Results for Section 6.1.2

Figure 7 depicts results from an extension of the experiment described in Section 6.1.2. This extension is meant to show the behavior of the slice learning and convex algorithms when the number of slices n falls in between the cases depicted in Figures 5a and 5b. We used the exact same experimental setup for three more values of n: 2, $\sqrt{r}$, and $r-1$. Figure 7a shows that the slice learning algorithm starts to show improvement even when $n = 2$, i.e. a single additional slice. Figures 7b and 7c reveal a progression as n increases until the slice learning algorithm achieves the $(r/m)^2$ rate.



(a) $n = 2$ **(b)** $n = \sqrt{r}$ **(c)** $n = r - 1$

Figure 7: Comparison of slice learning and convex approach for noisy slice recovery of low slice rank tensors with varying numbers of slices. SMSE vs. $(r/m)^2$ is plotted for each replication.

E.2. The Effect of Unbalanced Noise

To test the effect of unbalanced noise, we randomly generated ground truth tensors with varying sizes and low slice rank exactly as described in Section 6.1.2. Each tensor had $n = m$ slices. We again added mean-zero gaussian noise, but this time with varying standard deviations. In all cases, the total noise was kept constant, i.e. the noise model ϵ satisfied

$$\mathsf{E}\left[\left\|\epsilon^k\right\|_F^2\right] = .01m^2, \quad k = 1, \ldots, n. \tag{18}$$

For example, the experiments in Section 6.1.2 had all noise terms set with standard deviation 0.1, which satisfies (18).

In our first experiment, the top half of each slice ϵ^k had variance 0.2 and the bottom half had zero variance, i.e.

$$\mathsf{E}\left[\left(\epsilon_{ij}^k\right)^2\right] = \begin{cases} .02 & \text{if } i \leq m/2 \\ 0 & \text{if } i > m/2 \end{cases}.$$

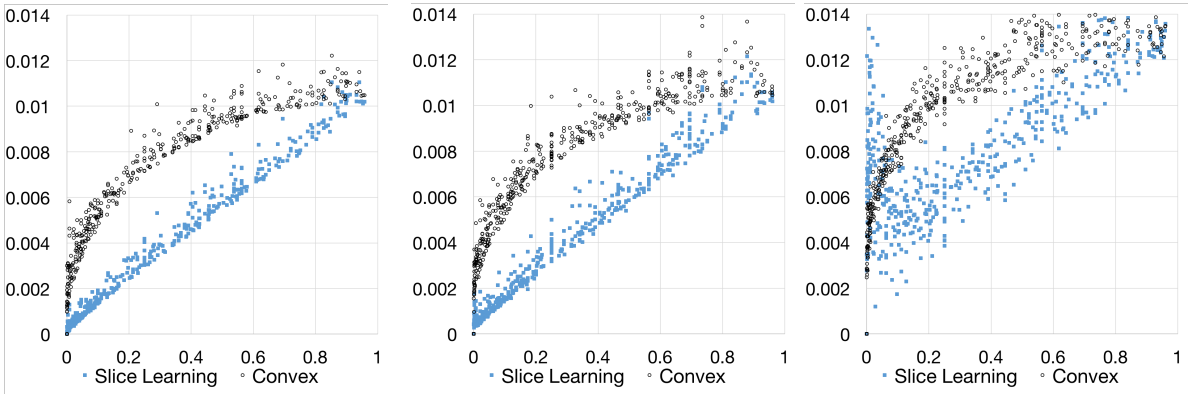
In Section 5.1, we defined an unbalance term δ, and here this corresponds to $\delta^2 = O(m^2 n^2)$, which is the highest scaling for δ when the variances are bounded. The results are summarized in Figure 8b, and compared with Figure 8a, which is a reproduction of the corresponding balanced noise experiment from Section 6.1.2,

they show that the slice learning algorithm still performs well, even though Theorem 2 no longer makes such a guarantee.

For our second experiment, only the top two rows have noise, but this noise is allowed to grow unbounded:

$$\mathsf{E}\left[\left(\epsilon_{ij}^{k}\right)^{2}\right] = \begin{cases} .005m & \text{if } i = 1, 2 \\ 0 & \text{if } i \geq 3 \end{cases}.$$

This case corresponds to $\delta^2 = O(m^3 n^2)$, and is summarized in Figure 8c. At this point, the slice learning algorithm performs poorly when r/m is small.



(a) $\delta = 0$ **(b)** $\delta = O(m^2 n^2)$ **(c)** $\delta = O(m^3 n^2)$

Figure 8: Comparison of slice learning and convex approach for noisy slice recovery of low slice rank tensors with varying levels of unbalanced noise. SMSE vs. $(r/m)^2$ is plotted for each replication. Display (a) is a reproduction of Figure 5c.

For another view into what is going on here, we fixed a particular tensor size and rank, and varied the unbalance level further. We randomly generated tensors of size $30 \times 30 \times 30$ with slice rank 5. For the added noise, as before, the bottom rows had zero variance, and the variance of the top rows were set to satisfy (18), so that the unbalance level can be varied by changing the number of these non-zero variance rows. For each level of unbalance, we ran 15 replications. The results shown in Figure 9 reveal that the performance of both algorithms worsens as the unbalance increases, and that the slice learning algorithm outperforms the convex algorithm for lower levels.

E.3. The Effect of Correlated Noise

The goal of our final set of synthetic experiments is to evaluate what happens when the independence assumption is relaxed. Following the same experimental setup as in the previous section, we randomly generated tensors of size $30 \times 30 \times 5$ with slice rank 5. We will focus on the setting where the noise between slices is correlated. To do so, each noise term was a zero-mean gaussian with standard deviation 0.1, but the covariances between corresponding entries across slices were allowed to vary:

$$\mathrm{Cov}(\epsilon_{ij}^{k}, \epsilon_{ij}^{\ell}) = c, \quad k \neq \ell,$$

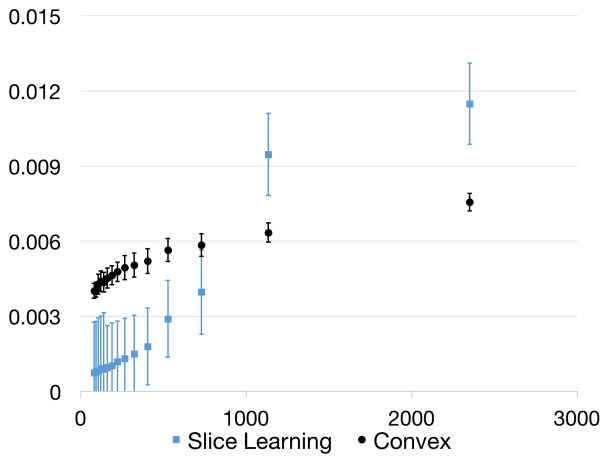
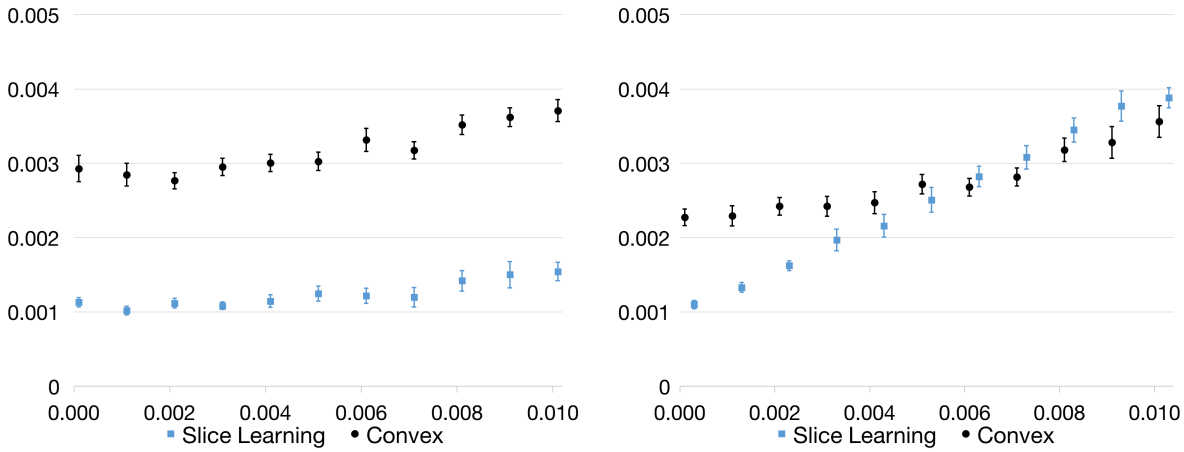


Figure 9: Comparison of slice learning and convex approach for tensors of size $30 \times 30 \times 30$ with slice rank 5 and varying levels of unbalanced noise. SMSE vs. δ^2 is plotted. Each point is the aggregate of 15 replications.

where c takes values between 0 and its maximum possible value of 0.01. Put another way, for each i, j, the vector of corresponding noise terms $(\epsilon_{ij}^1, \ldots, \epsilon_{ij}^n)$ was an independently generated mean-zero multivariate gaussian, with the above covariance structure.



(a) Different Slices

(b) Identical Slices

Figure 10: Comparison of slice learning and convex approach for tensors of size $30 \times 30 \times 5$ with slice rank 5 and varying levels correlated noise. SMSE vs. covariance is plotted for each replication. Each point is the aggregate of 15 replications, with 95% error bars. Points in Figure (b) are slightly offset horizontally to show overlapping error bars more clearly.

The results in Figure 10a show that both algorithms are only slightly affected by increasing covariance; note that at the most extreme case of $c = .01$, the noise terms across slices are identical! However, this does not necessarily mean that correlated noise is always innocuous, at least without further assumptions beyond slice rank. For example, for the experiment shown in Figure 10b, we used randomly generated tensors whose slices are identical. These tensors still have low slice rank, but the effect of correlation on noise is much

stronger. In particular, the extreme case $c = .01$ is now identical to having only a single slice of data.

F. Additional Results for Xiami Experiments

Tables 3 and 4 summarize the results of the experiments using data from `Xiami.com`, in terms of recovering the Download and Listen slices. See Section 6.2 for a detailed description of the experiment.

Users	Songs	Sparsity	Naive	Matrix	Slice
2,412	1,541	9.6	0.84 (11)	0.87 (7)	0.91 (12)
4,951	2,049	7.9	0.83 (14)	0.85 (9)	0.91 (12)
27,411	3,472	3.2	0.83 (11)	0.86 (8)	0.91 (14)
23,300	10,106	14.2	0.94 (18)	0.93 (13)	0.94 (18)
53,713	10,199	8.2	0.93 (10)	0.93 (7)	0.94 (20)

Table 3: Summary of experiments on Xiami data for recovering the Download slice. Each row corresponds to an experiment on a subset of the data. Columns 'Users' and 'Songs' show the number of users and songs in each experiment, and 'Sparsity' gives the average number of downloads per user in the data. Results for the naive benchmark, the matrix-based benchmark, and the slice learning algorithm are shown in the last three columns. The average AUC over 10 replications is reported, along with the rank in parentheses.

Users	Songs	Sparsity	Naive	Matrix	Slice
2,412	1,541	14.8	0.88 (6)	0.88 (7)	0.91 (11)
4,951	2,049	12.6	0.88 (7)	0.87 (11)	0.91 (11)
27,411	3,472	7.5	0.87 (6)	0.87 (3)	0.90 (9)
23,300	10,106	21.3	0.94 (7)	0.92 (8)	0.94 (15)
53,713	10,199	14.1	0.92 (5)	0.92 (12)	0.93 (7)

Table 4: Summary of experiments on Xiami data for recovering the Listen slice. Each row corresponds to an experiment on a subset of the data. Columns 'Users' and 'Songs' show the number of users and songs in each experiment, and 'Sparsity' gives the average number of listens per user in the data. Results for the naive benchmark, the matrix-based benchmark, and the slice learning algorithm are shown in the last three columns. The average AUC over 10 replications is reported, along with the rank in parentheses.

G. Commentary on Tensor Completion

To this point, our work has applied to the problem of noisy tensor recovery, a framework that addresses settings such as the retail example and our experiment with music streaming data. As discussed in Section 1, there are settings and applications where the existing data instead can be represented as a partially observed tensor, i.e. the tensor completion problem. To model the tensor completion setting mathematically, let Ω be the set of indices (i, j, k) of the observed entries, so that our data consists of the set of values $\{M^k_{ij} : (i, j, k) \in \Omega\}$, where $M \in \mathbb{R}^{m \times m \times n}$ is the ground truth tensor; for this discussion, we are assuming that the observed entries are observed without noise. Moreover, assume that Ω is generated randomly such

that each entry is observed independently with probability $p > 0$. The goal then is to design an estimator $\hat{M}$, which is now a function of only the observed entries.

Theoretical guarantees in this area should address performance in terms of four parameters now: m, n, r and p. In particular, the quantity of m^2p is of interest as it is equal to the expected number of observed entries per slice. Consider that a necessary condition for any algorithm to complete a tensor is

$$m^2p = \Omega(r^2 + mr/n). \tag{19}$$

This lower bound comes from the fact that having slice rank r allows for $O(nr^2 + mr)$ degrees of freedom, and that the quantity m^2np is the expected total number of observed entries. Dividing through by n gives the lower bound in terms of the number of observed entries per slice. Achieving this bound would mean that (1) in the best case, the per-slice data requirement is r^2, which does not depend on the size of the slices, and (2) as sources of side information are added, the per-slice data requirement decreases linearly until r^2. In contrast, a matrix-based approach such as using a matrix completion algorithm on each slice separately would reduce to completing n matrices of size $m \times m$ and rank r. In that case, the best known guarantees (e.g. Gross (2011)), which have matching lower bounds, are exact recovery with high probability given that

$$m^2p = \Omega(mr),$$

where we have omitted polylogarithmic factors. Another matrix-based approach of applying a matrix completion algorithm on a single unfolding would not improve on this guarantee.

There has been much work in designing algorithm tailored to the tensor completion problem. The strongest existing guarantee is by Yuan and Zhang (2015), who propose an algorithm that recovers M exactly, with high probability, when

$$m^2p = \Omega(r^2 + mr^2/n + m\sqrt{r}/\sqrt{n}).$$

This result makes significant progress toward the lower bound (19), but unfortunately the proposed algorithm is computationally intractable. Huang et al. (2015) analyze a tractable algorithm and show that a sufficient condition for recovery is

$$m^2p = \Omega(mr). \tag{20}$$

Now, when entries are observed uniformly at random, it is possible to map completion problems to noisy recovery problems by dividing the observed entries by p and treating unobserved entries as zero. This technique has been applied in the matrix completion setting (Achlioptas and McSherry (2007), Keshavan et al. (2010), Chatterjee (2014)). Following the same arguments, we could use the slice learning algorithm for tensor completion: we (1) divide each observed entry by the proportion of entries observed, i.e. $|\Omega|/m^2n$, (2) treat all hidden entries as observations of the value zero, and (3) execute the slice learning algorithm as usual on this modified tensor. In other words, we execute the slice learning algorithm on tensor $(m^2n/|\Omega|)M_\Omega$, where M_Ω is the tensor defined as

$$(M_\Omega)^k_{ij} = \begin{cases} M^k_{ij} & \text{if } (i,j,k) \in \Omega \\ 0 & \text{if } (i,j,k) \notin \Omega \end{cases}.$$

The modified tensor $(1/p)M_\Omega$ is a noisy observation of M. To see this, note that we can define the additive

noise term of the $(i,j,k)^{\text{th}}$ entry as

$$\epsilon_{ij}^k \sim \frac{\text{Ber}(p)}{p} M_{ij}^k - M_{ij}^k. \tag{21}$$

These noise terms are independent with mean zero, and $(1/p)M_\Omega = M + \epsilon$. It follows that M could be estimated by applying our slice learning algorithm to $(1/p)M_\Omega$. Since we do now know p, we use the proportion of observed entries as an estimate of p.

Since we can reformulate the completion problem as a noisy recovery problem, a natural question then is what Theorems 1 and 2 tell us about the performance of the slice learning algorithm, where again performance is measured in terms of p. To analyze the slice learning algorithm as in Theorems 1 and 2, assume that the entries of M lie in $[-1,1]$. It follows directly from (21) that

$$\mathsf{E}[(\epsilon_{ij}^k)^d] = (M_{ij}^k)^d \left(\frac{(1-p)^d}{p^{d-1}} + (1-p) \right) \leq \frac{1}{p^{d-1}}, \tag{22}$$

for even values of d. Now in the statements of Theorems 1 and 2, the guarantee is parameterized by K, where it is assumed that $\mathsf{E}[(\epsilon_{ij}^k)^6] \leq K^6$. This is actually a compact version of a more specific guarantee we show (see (17) in the Appendix) that is parameterized by the second, fourth, and sixth moments of the noise terms. Combining (22) and that guarantee, we have the following result:

Corollary 1. *Assume the entries of M lie in $[-1,1]$. Suppose Ω is randomly chosen such that each index is included independently with probability $p > 0$. Let $\hat{M}$ be the result of applying the slice learning algorithm to $(1/p)M_\Omega$. Then there exists a constant c such that*

$$\text{SMSE}(\hat{M}) \leq c \left[\frac{r^2}{m^2 p} + \frac{r^2}{\gamma_M^2 mnp^3} + \frac{r^2}{\gamma_M^2 m^4 n^2 p^5} + \frac{r^2}{\gamma_M^2 m^2 np^4} + \frac{r^2 \delta^2}{\gamma_M^2 m^3 n^2 p} \right].$$

Corollary 1 implies that we can expect to recover M using the slice learning algorithm as long as the denominator of each term in the guarantee is much larger than the numerator. To make more sense of this sufficient condition, let us assume that the noise is balanced ($\delta = 0$) and that γ_M scales as a constant, in which case after some algebraic contortion, the condition can be expressed as

$$m^2 p = \Omega(r^2 + m^{5/3} r^{2/3}/n^{1/3} + m^{3/2} r^{1/2}/n^{1/4}).$$

Unlike the guarantee (20) of Huang et al. (2015), this guarantee decreases with n and achieves the final r^2 value for sufficiently large n. On the other hand, the scaling with m is worse, and so is only an improvement when n grows sufficiently faster than m. Finally, our algorithm is dominated by that of Yuan and Zhang (2015), but is computationally efficient. The guarantees we have described here are summarized in Table 5.

G.1. Experiment

We ran a set of synthetic experiments to test the performance of the slice learning algorithm in the tensor completion setting. Our first experiment is an exact replication of one of the experiments from Gandy et al. (2011): we randomly generated tensors of size $20 \times 30 \times 40$ with Tucker rank $(2,2,2)$, using the same procedure described in §6.1.1, and observed each entry with probability 0.6. We benchmarked against the

Method	Per-Slice Observations
Lower Bound	$r^2 + mr/n$
Matrix-Complete Slices	mr
Matrix-Complete Unfolding	mr
Huang et al. (2015)	mr
Yuan and Zhang (2015)	$r^2 + mr^2/n + mr^{1/2}/n^{1/2}$
Slice Learning	$r^2 + m^{5/3}r^{2/3}/n^{1/3} + m^{3/2}r^{1/2}/n^{1/4}$

Table 5: Comparison of guarantees for tensor completion algorithms, in terms of the number of per-slice observations sufficient for completion. Logarithmic terms are omitted.

convex algorithm

$$\underset{Y}{\operatorname{argmin}} \sum_{i=1}^{3} \lambda \left\| Y_{(i)} \right\|_* + \|Y_\Omega - X_\Omega\|_F^2 , \tag{23}$$

which we solved via the Douglas-Rachford splitting method described in Gandy et al. (2011), using their recommended values for the step size and index of the proximal operator, and solving (23) multiple times with increasing values of λ until the solutions converged. This procedure was repeated for 60 replications, and the results are summarized in Figure 11a, where we give the root mean squared error (RMSE, square root of MSE) averaged over the 60 replications for both the slice learning algorithm and the convex algorithm (23). Since Douglas-Rachford is an iterative method, we report the average RMSE at various points in the procedure, i.e. various numbers of iterations. On the other hand, the slice learning algorithm consists of only a single 'iteration', and so a single value is reported. We also performed a second experiment that is a larger version of the first. We randomly generated tensors of size $200 \times 200 \times 200$ with Tucker rank $(10, 10, 10)$, with the rest of the experiment remaining the same. The results are summarized in Figure 11b.

Method	Iterations	RMSE
Slice Learning	1	1.6×10^{-3}
Convex	1	1.5×10^{-2}
	10	5.7×10^{-3}
	*24	1.2×10^{-3}
	100	8.2×10^{-6}

(a) $20 \times 30 \times 40$ tensors of Tucker rank $(2, 2, 2)$

Method	Iterations	RMSE
Slice Learning	1	3.9×10^{-4}
Convex	1	9.7×10^{-3}
	10	4.8×10^{-3}
	*22	3.2×10^{-4}
	50	1.1×10^{-4}

(b) $200 \times 200 \times 200$ tensors of Tucker rank $(10, 10, 10)$

Figure 11: Results of synthetic completion experiments. Entries were observed with probability 0.6. RMSE and iteration count are reported, averaged over 60 replications. Starred (*) rows correspond to the iteration of the convex algorithm in which, for the first time, the average RMSE falls below that of the slice learning algorithm.

The slice learning algorithm performs reasonably well, achieving RMSE on the order of 10^{-3} to 10^{-4}. To put this into perspective, the size of the elements of the randomly generated tensors is on the order of 10^{-2},

so this RMSE amounts to a relative error of about 1% to 10%. Ignoring computational costs, the convex algorithm outperforms the slice learning algorithm, achieving a lower average RMSE in both experiments after approximately 20 iterations. Moreover, we observe the RMSE continuing to decrease with each iteration, and indeed with enough iterations the RMSE may go to zero (or machine precision), corresponding to *exact* recovery of the original tensor. On the other hand, when factoring in computational costs, the slice learning algorithm performs very well. Each Douglas-Rachford iteration requires singular value decompositions of all three (dense) unfoldings of the tensor, which means each iteration is more computationally expensive than the entire slice learning algorithm, which only requires SVDs of two (sparse) unfoldings. In concrete terms, this meant that for our larger experiment, the slice learning algorithm ran in less than a minute, while the convex approach took upwards of one hour to reach the same level of accuracy.

.

References

Achlioptas D, McSherry F (2007) Fast computation of low-rank matrix approximations. *Journal of the ACM (JACM)* 54(2):9.

Ansari A, Essegaier S, Kohli R (2000) Internet recommendation systems. *Journal of Marketing research* 37(3):363–375.

Ansari A, Mela CF (2003) E-customization. *Journal of marketing research* 40(2):131–145.

Archak N, Ghose A, Ipeirotis PG (2011) Deriving the pricing power of product features by mining consumer reviews. *Management Science* 57(8):1485–1509.

Bennett J, Lanning S (2007) The netflix prize. *Proceedings of KDD cup and workshop*, volume 2007, 35.

Besbes O, Gur Y, Zeevi A (2015) Optimization in online content recommendation services: Beyond click-through rates. *Manufacturing & Service Operations Management* 18(1):15–33.

Bodapati AV (2008) Recommendation systems with purchase data. *Journal of marketing research* 45(1):77–93.

Breese JS, Heckerman D, Kadie C (1998) Empirical analysis of predictive algorithms for collaborative filtering. *Proceedings of the Fourteenth conference on Uncertainty in artificial intelligence*, 43–52 (Morgan Kaufmann Publishers Inc.).

Bucklin RE, Sismeiro C (2003) A model of web site browsing behavior estimated on clickstream data. *Journal of marketing research* 40(3):249–267.

Chatterjee S (2014) Matrix estimation by universal singular value thresholding. *The Annals of Statistics* 43(1):177–214.

Chintagunta P, Dubé JP, Goh KY (2005) Beyond the endogeneity bias: The effect of unmeasured brand characteristics on household-level brand choice models. *Management Science* 51(5):832–849.

Chintagunta PK, Dube JP (2005) Estimating a stockkeeping-unit-level brand choice model that combines household panel data and store data. *Journal of Marketing Research* 42(3):368–379.

Chong JK, Ho TH, Tang CS (2001) A modeling framework for category assortment planning. *Manufacturing & Service Operations Management* 3(3):191–210.

Chung TS, Rust RT, Wedel M (2009) My mobile music: An adaptive personalization system for digital audio players. *Marketing Science* 28(1):52–68.

Das SR, Chen MY (2007) Yahoo! for amazon: Sentiment extraction from small talk on the web. *Management Science* 53(9):1375–1388.

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

Fader PS, Hardie BG (1996) Modeling consumer choice among skus. *Journal of marketing Research* 442–452.

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

Ghose A, Ipeirotis PG, Li B (2012) Designing ranking systems for hotels on travel search engines by mining user-generated and crowdsourced content. *Marketing Science* 31(3):493–520.

Gopinath S, Chintagunta PK, Venkataraman S (2013) Blogs, advertising, and local-market movie box office performance. *Management Science* 59(12):2635–2654.

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

Herlocker J, Konstan JA, Riedl J (2002) An empirical analysis of design choices in neighborhood-based collaborative filtering algorithms. *Information retrieval* 5(4):287–310.

Hu M, Liu B (2004) Mining and summarizing customer reviews. *Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining*, 168–177 (ACM).

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

Hui SK, Huang Y, Suher J, Inman JJ (2013) Deconstructing the "first moment of truth": Understanding unplanned consideration and purchase conversion using in-store video tracking. *Journal of Marketing Research* 50(4):445–462.

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

Kolda TG, Bader BW (2009) Tensor decompositions and applications. *SIAM review* 51(3):455–500.

Liu X, Singh PV, Srinivasan K (2016) A structured analysis of unstructured big data by leveraging cloud computing. *Marketing Science* 35(3):363–388.

Lu S, Xiao L, Ding M (2016) A video-based automated recommender (var) system for garments. *Marketing Science* 35(3):484–510.

MobileCommerceDaily (2013) Macy's exec underpins navigation, in-store scanning for holiday marketing success. `http://www.mobilecommercedaily.com/macy's-exec-underpins-navigation-in-store-scanning-for-holiday-marketing-success`.

Moe WW (2006) An empirical two-stage choice model with varying decision rules applied to internet clickstream data. *Journal of Marketing Research* 43(4):680–692.

Montgomery AL, Li S, Srinivasan K, Liechty JC (2004) Modeling online browsing and path analysis using clickstream data. *Marketing Science* 23(4):579–595.

Moon S, Russell GJ (2008) Predicting product purchase from inferred customer similarity: An autologistic model approach. *Management Science* 54(1):71–82.

Sismeiro C, Bucklin RE (2004) Modeling purchase behavior at an e-commerce web site: A task-completion approach. *Journal of marketing research* 41(3):306–323.

Trusov M, Ma L, Jamal Z (2016) Crumbs of the cookie: User profiling in customer-base analysis and behavioral targeting. *Marketing Science* 35(3):405–426.

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

Wu J, Rangaswamy A (2003) A fuzzy set model of search and consideration with an application to an online market. *Marketing Science* 22(3):411–434.

Yu Y, Wang T, Samworth RJ (2015) A useful variant of the davis–kahan theorem for statisticians. *Biometrika* 102(2):315–323.

Yuan M, Zhang CH (2015) On tensor completion via nuclear norm minimization. *Foundations of Computational Mathematics* 1–38.