

---

# Partitioned Tensor Factorizations for Learning Mixed Membership Models

---

Zilong Tan<sup>1</sup> Sayan Mukherjee<sup>1</sup>

## Abstract

We present an efficient algorithm for learning mixed membership models when the number of variables  $p$  is much larger than the number of hidden components  $k$ . This algorithm reduces the computational complexity of state-of-the-art tensor methods, which require decomposing an  $O(p^3)$  tensor, to factorizing  $O(p/k)$  sub-tensors each of size  $O(k^3)$ . In addition, we address the issue of negative entries in the empirical method of moments based estimators. We provide sufficient conditions under which our approach has provable guarantees. Our approach obtains competitive empirical results on both simulated and real data.

## 1. Introduction

Mixed membership models (Woodbury et al., 1978; Pritchard et al., 2000a;b; Blei et al., 2003; Erosheva, 2005) have been used extensively across applications ranging from modeling population structure in genetics (Pritchard et al., 2000a;b) to topic modeling of documents (Woodbury et al., 1978; Blei et al., 2003; Erosheva, 2005). Mixed membership models use Dirichlet latent variables to define cluster membership where samples can partially belong to each of  $k$  latent components. Parameter estimation for such latent variables models (LVMs) using maximum likelihood methods such as expectation maximization is computationally intensive for large data, for example, if number of samples  $n$  is large.

Parameter estimation using the method of moments for LVMs is an attractive scalable alternative that has been shown to have certain theoretical and computational advantages over maximum likelihood methods in the setting when  $n$  is large. For LVMs, method of moments approaches reduce to tensor methods—the moments of the model parameters are expressed as a function of statistics

of the observations in a tensor form. Inference in this setting becomes a problem of tensor factorization. Computational advantages of using tensor methods have been observed for many popular models, including latent Dirichlet allocation (Anandkumar et al., 2012), spherical Gaussian mixture models (Hsu & Kakade, 2013), hidden Markov models (Anandkumar et al., 2012), independent component analysis (Comon & Jutten, 2010), and multi-view models (Anandkumar et al., 2014). An appealing property of tensor methods is the guarantee of a unique decomposition under mild conditions (cf. Kruskal, 1977; Leurgans et al., 1993).

There are two complications to using standard tensor decomposition methods (Anandkumar et al., 2016; 2014; Gu et al., 2014; Kuleshov et al., 2015; Nicolò Colombo and Nikos Vlassis; Kim et al., 2014; Chi & Kolda, 2012) for LVMs. The first problem is computation and space complexity. Given  $p$  variables in the LVM, parameter inference requires factorizing typically a non-orthogonal estimator tensor of size  $O(p^3)$  (Anandkumar et al., 2014; Kuleshov et al., 2015; Nicolò Colombo and Nikos Vlassis), which is prohibitive for large  $p$ . When the estimator is orthogonal and symmetric, this can be done in  $O(p^2 \log p)$  (Wang et al., 2015). Online tensor decomposition (Huang et al., 2013) uses dimension reduction to instead factorize a reduced  $k$ -by- $k$ -by- $k$  tensor. However, the dimension reduction can be slower than decomposing the estimator directly for large sample sizes, as well as suffer from high variance (Wang et al., 2015). We introduce a simple factorization with improved complexity for the general case where the parameters are not required to be orthogonal.

The second problem arises from negative entries in the empirical moments tensor. LVMs for count data are constrained to have nonnegative parameters. However, the empirical moments tensor computed from the data may contain negative elements due to sampling variation and noise. Indeed, for small sample sizes or data with many small or zero counts, there will be many negative entries in the empirical moments tensor. General tensor decomposition algorithms (Kuleshov et al., 2015; Nicolò Colombo and Nikos Vlassis), including the tensor power method (TPM) (Anandkumar et al., 2014), do not guarantee the nonnegativity of model parameters. Approaches such as positive/nonnegative tensor factorization (Chi & Kolda, 2012; Shashua & Hazan, 2005; Welling & Weber, 2001) also do

---

<sup>1</sup>Duke University, Durham, NC. Correspondence to: Sayan Mukherjee <sayan@stat.duke.edu>.

not address this situation as they require all the elements of the tensor to be factorized to be nonnegative. With robust tensor methods (Anandkumar et al., 2016; Gu et al., 2014), sparse negative entries may potentially be treated as corrupted elements; however, these methods are not applicable in this setting since there can be many negative elements.

In this paper, we introduce a novel parameter inference algorithm called partitioned tensor parallel quadratic programming (PTPQP) that is efficient in the setting where the number of variables  $p$  is much larger than the number of latent components  $k$ . The algorithm is also robust to negative entries in the empirical moments tensor. There are two key innovations in the PTPQP algorithm. The first innovation is a partitioning technique which recovers the parameters through factorizing  $O(p/k)$  much smaller sub-tensors each of size  $O(k^3)$ . The second innovation is a parallel quadratic programming (Brand & Chen, 2011) based algorithm to factor tensors with negative entries under the constraint that the factors are all nonnegative. To the best of our knowledge, this is the first algorithm designed to address the problem of negative entries in empirical estimator tensors. We show that the proposed factorization algorithm converges linearly with respect to each factor matrix. We also provide sufficient conditions under which the partitioned factorization scheme is consistent, the parameter estimates converge to the true parameters.

## 2. Preliminaries

**Notations.** We use bold lowercase letters to represent vectors and bold capital letters for matrices. Tensors are denoted by calligraphic capital letters. The subscript notation  $A_j$  refers to  $j$ -th column of matrix  $A$ . We denote the  $j$ -th column of the identity matrix as  $e_j$  and  $\mathbf{1}$  is a vector of ones. We further write  $\text{diag}(\mathbf{x})$  for a diagonal matrix whose diagonal entries are  $\mathbf{x}$ , and  $\text{diag}(A)$  to mean a vector of the diagonal entries of  $A$ .

Element-wise matrix operators include  $\succ$  and  $\succeq$ , e.g.,  $A \succeq 0$  means that  $A$  has nonnegative entries.  $(\cdot)_+$  refers to element-wise  $\max(\cdot, 0)$ .  $*$  and  $\oslash$  respectively represent element-wise multiplication and division. Moreover,  $\times$  refers to the outer product and  $\odot$  denotes the Khatri-Rao product.  $\|\cdot\|_F$  and  $\|\cdot\|_2$  represent the Frobenius norm and spectral norm, respectively.

**Tensor basics.** This paper uses similar tensor notations as (Kolda & Bader, 2009). In particular, we are primarily concerned with Kruskal tensors in  $\mathbb{R}^{d_1 \times d_2 \times d_3}$ , which can be expressed in the form of

$$\mathcal{T} = \sum_{j=1}^r \mathbf{A}_j \times \mathbf{B}_j \times \mathbf{C}_j, \quad (1)$$

where  $\mathbf{A}$ ,  $\mathbf{B}$ , and  $\mathbf{C}$  are respectively  $d_1$ -by- $r$ ,  $d_2$ -by- $r$ , and

$d_3$ -by- $r$  factor matrices. The rank of  $\mathcal{T}$  is defined as the smallest  $r$  that admits such a decomposition. The decomposition is known as the CP (CANDECOMP/PARAFAC) decomposition. The  $j$ -mode unfolding of  $\mathcal{T}$ , denoted by  $\mathcal{T}_{(j)}$ , for  $j = 1, 2, 3$  is a  $d_j$ -by- $(\prod_{t \neq j} d_t)$  matrix whose rows are serializations of the tensor fixing the index of the  $j$ -th dimension. The unfoldings have the following well-known compact expressions:

$$\begin{aligned} \mathcal{T}_{(1)} &= \mathbf{A} (\mathbf{C} \odot \mathbf{B})^\top, & \mathcal{T}_{(2)} &= \mathbf{B} (\mathbf{C} \odot \mathbf{A})^\top \\ \mathcal{T}_{(3)} &= \mathbf{C} (\mathbf{B} \odot \mathbf{A})^\top. \end{aligned} \quad (2)$$

## 3. Learning through Method of Moments

### 3.1. Generalized Dirichlet latent variable models

A generalized Dirichlet latent variable model (GLDM) was proposed in (Zhao et al., 2016) for the joint distribution of  $n$  observations  $\mathbf{y}_1, \mathbf{y}_2, \dots, \mathbf{y}_n$ . Each observation  $\mathbf{y}_i$  consists of  $p$  variables  $\mathbf{y}_i = (y_{i1}, y_{i2}, \dots, y_{ip})^\top$ . GLDM assumes a generative process involving  $k$  hidden components. For each observation, sample a random Dirichlet vector  $\mathbf{x}_i = (x_{i1}, x_{i2}, \dots, x_{ik})^\top \in \Delta^{k-1}$  with concentration parameter  $\boldsymbol{\alpha} = (\alpha_1, \alpha_2, \dots, \alpha_k)^\top$ . The elements of  $\mathbf{x}_i$  are the membership probabilities for  $\mathbf{y}_i$  to belong to each of the  $k$  components. Specifically,

$$y_{ij} \sim \sum_{h=1}^k x_{ih} g_j(\theta_{jh}),$$

where  $g_j(\theta_{jh})$  is the density of the  $j$ -th variable specific to component  $h$  with parameter  $\boldsymbol{\theta}_j = (\theta_{j1}, \theta_{j2}, \dots, \theta_{jk})$ . One advantage of GLDM is that  $y_{ij}$  can take categorical values. Let  $d_j$  denote the number of categories for the  $j$ -th variable (set  $d_j = 1$  for scalar variables),  $\boldsymbol{\theta}_j$  becomes a  $d_j$ -by- $k$  probability matrix where the  $c$ -th row corresponds to category  $c$ . We aim to accurately recover  $\boldsymbol{\theta}_j$  from independent copies of  $\mathbf{y}_i$  involving variables of mixed data types, either categorical or non-categorical.

### 3.2. Moment-based estimators

The moment estimators of latent variable models typically take the form of a tensor (Anandkumar et al., 2014). Consider the estimators of GLDM (Zhao et al., 2016) for example. Let  $\mathbf{b}_{ij} = e_{y_{ij}}$  if variable  $j$  is categorical;  $\mathbf{b}_{ij} = y_{ij}$  otherwise. The second- and third- order parameter estimators for variable  $j$ ,  $s$ , and  $t$  are written

$$\begin{aligned} \mathcal{M}^{js} &= \mathbb{E}[\mathbf{b}_{ij} \times \mathbf{b}_{is}] - c_1 \mathbb{E}[\mathbf{b}_{ij}] \mathbb{E}[\mathbf{b}_{is}]^\top \\ \mathcal{M}^{jst} &= \mathbb{E}[\mathbf{b}_{ij} \times \mathbf{b}_{is} \times \mathbf{b}_{it}] + c_2 \mathbb{E}[\mathbf{b}_{ij}] \times \mathbb{E}[\mathbf{b}_{is}] \times \mathbb{E}[\mathbf{b}_{it}] \\ &\quad - c_3 (\mathbb{E}[\mathbb{E}[\mathbf{b}_{ij}] \times \mathbf{b}_{is} \times \mathbf{b}_{it}] + \mathbb{E}[\mathbf{b}_{ij} \times \mathbb{E}[\mathbf{b}_{is}] \times \mathbf{b}_{it}] \\ &\quad + \mathbb{E}[\mathbf{b}_{ij} \times \mathbf{b}_{is} \times \mathbb{E}[\mathbf{b}_{it}]]) . \end{aligned}$$

Alternatively,  $\mathcal{M}^{js}$  and  $\mathcal{M}^{jst}$  have the following CP decomposition into parameters  $\theta_j$ :

$$\mathcal{M}^{js} = \sum_{h \geq 1} \eta_h \theta_{jh} \times \theta_{sh}, \quad \theta_{uv} \in \mathbb{R}^{d_i} \quad (3)$$

$$\mathcal{M}^{jst} = \sum_{h \geq 1} \lambda_h \theta_{jh} \times \theta_{sh} \times \theta_{th}, \quad \theta_{uv} \in \mathbb{R}^{d_i}. \quad (4)$$

Here,  $c_1, c_2, c_3, \eta_h$ , and  $\lambda_h$  depend on only  $\alpha_0 = \sum_{j \geq 1} \alpha_j$ . For the special case of latent Dirichlet allocation,  $\mathcal{M}^{js}$  and  $\mathcal{M}^{jst}$  are scalar joint probabilities.

The parameters  $\theta_j$  are typically obtained by factorizing the block tensor  $\mathcal{M}_2$  whose  $(j, s)$ -th element is the empirical  $\widehat{\mathcal{M}}^{js}$  and/or  $\mathcal{M}_3$  whose  $(j, s, t)$ -th element is the empirical  $\widehat{\mathcal{M}}^{jst}$  (Anandkumar et al., 2016; 2014; Zhao et al., 2016). Note that  $\theta_j$  are generally non-orthogonal, and thus preprocessing steps are needed for orthogonal decomposition methods (Wang et al., 2015; Song et al., 2016; Anandkumar et al., 2014). The preprocessing can be expensive and often leads to suboptimal performance (Souloumiac, 2009; Nicolò Colombo and Nikos Vlassis). Here, we highlight a few relevant observations:

- $\mathcal{M}^{js}$  alone does not yield unique parameters  $\theta_j$  due to the well-known rotation problem. This is true even when enforcing nonnegativity constraints on parameters (Donoho & Stodden, 2004).
- $\mathcal{M}^{jst}$  is sufficient to uniquely recover the parameters under certain mild conditions (Kruskal, 1977); for example, when any two of  $\theta_j, \theta_s$ , and  $\theta_t$  have linearly independent columns and the columns of the third are pair-wise linearly independent (Leurgans et al., 1993).
- The empirical estimator  $\widehat{\mathcal{M}}^{jst}$  generally contains negative entries due to variance and noise. The fraction of negative entries can approach 50%, as we shall see in experiments. We address this issue in § 4.4.
- While the decomposition (4) can be unique up to permutation and rescaling, the correspondence between each column of the factor matrix and each hidden component may not be consistent across multiple decompositions. Techniques for achieving consistency are developed in § 4.2.

### 3.3. Computational complexity

Tensor methods such as TPM typically decompose the  $O(p^3 d_{max}^3)$  full estimator tensor that includes all variables. More efficient algorithms have been developed for the case that parameters are orthogonal (Wang et al., 2015; Song et al., 2016), and when the sample size is small (Huang et al., 2013). However, these methods do not

apply in the general case where the parameters are non-orthogonal and the sample size can be potentially large. A key insight underlying our approach is that it is sufficient to recover the parameters by factorizing only  $O(p/k)$  much smaller sub-tensors each of size  $O(k^3)$ . This technique can also be combined with the aforementioned methods to further improve the complexity in certain cases.

## 4. An efficient algorithm

In this section, we develop partitioned tensor parallel quadratic programming (PTPQP) an efficient approximate algorithm for learning mixed membership models. We first introduce a novel partitioning-and-matching scheme that reduces parameter estimation to factorizing a sequence of sub-tensors. Then, we develop a nonnegative factorization algorithm that can handle negative entries in the sub-tensors.

### 4.1. Partitioned factorization

Factorizing the full tensor formed by all  $\mathcal{M}^{jst}$  is expensive while a three-variable tensor  $\mathcal{M}^{jst}$  in (4) alone may not be sufficient to determine  $\theta_j$  when  $k$  is large. In this section, we consider factorizing the sub-tensors corresponding to a cover of the set of variables  $[p]$  such that each sub-tensor admits an identifiable CP decomposition (1), i.e. unique up to permutation and rescaling of columns. This gives the parameters for all variables. Suppose that  $p > k$  and the maximum number of categories  $d_{max}$  is a constant, the aggregated size of the sub-tensors can be much smaller, i.e.,  $O(pk^2)$ , than the size  $O(p^3)$  of the full estimator.

Let  $\pi^j, \pi^s$ , and  $\pi^t$  denote ordered subsets  $\subseteq [p]$ , with cardinality  $|\pi_j| = p_j, |\pi_s| = p_s$ , and  $|\pi_t| = p_t$ , respectively. Consider the  $p_j$ -by- $p_s$ -by- $p_t$  block tensor  $\mathcal{M}^{\pi^j \pi^s \pi^t}$  whose  $(u, v, w)$ -th element is the tensor  $\mathcal{M}_{uvw}^{\pi^j \pi^s \pi^t} = \mathcal{M}^{\pi_u^j \pi_v^s \pi_w^t}$ . From (4) the tensor  $\mathcal{M}^{\pi^j \pi^s \pi^t}$  is

$$\sum_{h=1}^k \lambda_h \begin{bmatrix} \theta_{\pi_1^j h} \\ \theta_{\pi_2^j h} \\ \vdots \\ \theta_{\pi_{p_j}^j h} \end{bmatrix} \times \begin{bmatrix} \theta_{\pi_1^s h} \\ \theta_{\pi_2^s h} \\ \vdots \\ \theta_{\pi_{p_s}^s h} \end{bmatrix} \times \begin{bmatrix} \theta_{\pi_1^t h} \\ \theta_{\pi_2^t h} \\ \vdots \\ \theta_{\pi_{p_t}^t h} \end{bmatrix}. \quad (5)$$

Clearly, the block tensor is identifiable if it has an identifiable sub-tensor. Suppose that a sub-tensor  $\mathcal{M}^{\pi^u \pi^v \pi^w}$  is identifiable, then one can construct an identifiable tensor  $\mathcal{M}^{\pi^{j'} \pi^{s'} \pi^{t'}}$  from  $\mathcal{M}^{\pi^j \pi^s \pi^t}$  by setting

$$\pi^{j'} = \pi^j \cup \pi^u, \quad \pi^{s'} = \pi^s \cup \pi^v, \quad \pi^{t'} = \pi^t \cup \pi^w. \quad (6)$$

We further remark that a sub-tensor can be identifiable under mild conditions, for example, if the sum of the Kruskal rank of the three factor matrices is at least than  $2k + 2$  (Kruskal, 1977).

Given an identifiable sub-tensor  $\mathcal{M}^{\pi^u \pi^v \pi^w}$  of *anchor variables* indexed by  $\pi^u$ ,  $\pi^v$ , and  $\pi^w$ , the partitioning produces a set of sub-tensors (partitions) constructed through (6), that includes all variables. Thus,  $\mathcal{M}^{\pi^u \pi^v \pi^w}$  is a common sub-tensor shared across all partitions. We choose anchor variables whose parameter matrices are of full column rank to obtain an identifiable  $\mathcal{M}^{\pi^u \pi^v \pi^w}$ . Finally, one can divide the rest of variables evenly and randomly into the partitions.

## 4.2. Matching parameters with hidden components

Since the factorization of a partition (5) can only be identifiable up to permutation and rescaling of the columns of constituent  $\theta_j$ , the correspondence between the columns of  $\theta_j$  and hidden components can differ across partitions. To enforce consistency, we associate a permutation operator  $\psi^j$  for each variable  $j$  such that  $(\psi^j \theta_j)_h$  are the parameters specific to hidden component  $h$  across all variables  $j$ . Consider the following vector representation of  $\psi$ :

$$\begin{aligned} \psi &= (\psi_1, \psi_2, \dots, \psi_k), \quad \psi_i \in [k] \\ \psi \mathbf{A} &= [\mathbf{A}_{\psi_1}, \mathbf{A}_{\psi_2}, \dots, \mathbf{A}_{\psi_k}]. \end{aligned}$$

Observe that  $\psi^j = \psi^s = \psi^t$  within a factorization of  $\mathcal{M}^{jst}$ , and this also holds for the partitioned factorization (5) of  $\mathcal{M}^{\pi^j \pi^s \pi^t}$  as well, i.e.,  $\psi^x = \psi^y$ ,  $\forall x, y \in \pi^j \cup \pi^s \cup \pi^t$ .

Consider the factorizations of  $\mathcal{M}^{\pi^j \pi^s \pi^t}$  and  $\mathcal{M}^{\pi^u \pi^v \pi^w}$  and suppose that  $\exists x \in (\pi^j \cup \pi^s \cup \pi^t) \cap (\pi^u \cup \pi^v \cup \pi^w)$ . The permutation operator for one factorization is determined given the other by column matching the parameters of variable  $x$  in both factorizations. Thus, an inductive way to achieve a consistent factorization is to start with one factorization, and let its permutation be the identity  $(1, 2, \dots, k)$ , then perform the factorization over new sets of variables with at least one variable in common with the initial factorization. Permutations for the sequential factorizations are determined via column matching parameter matrices of the common variables.

Given two factorized parameter matrices  $\theta_j$  and  $\theta'_j$  of variable  $j$ , our goal is to find a *consistent permutation*  $\psi$  (of  $\theta_j$  with respect to  $\theta'_j$ ) such that  $(\psi \theta_j)_h$  and  $\theta'_{jh}$  correspond to the same hidden component for all  $h \in [k]$ . We now present an algorithm with provable guarantees to compute a consistent permutation.

**Smallest angle matching** A simple matching algorithm is to match the two columns of the two parameter matrices that have the smallest angle between them. Consider the factorizations of  $\mathcal{M}^{jst}$  and  $\mathcal{M}^{juv}$  which yield respectively parameters  $\theta_j$  and  $\theta'_j$  for the common variable  $j$ . Given the permutation  $\psi^j$  for  $\mathcal{M}^{jst}$ , the permutation  $\psi^u$  for  $\mathcal{M}^{juv}$  is

computed by:

$$\psi_s^u = \arg \max_t (\bar{\theta}'_j{}^\top \psi^j \bar{\theta}_j)_{ts}. \quad (7)$$

Here,  $\bar{\theta}_j$  and  $\bar{\theta}'_j$  represent respectively the normalized  $\theta_j$  and  $\theta'_j$  with each column having unit Euclidean norm.

There are cases that  $\psi^u$  computed via Equation (7) is not consistent: 1)  $\psi^u$  contains duplicate entries and hence is ineligible; and 2) since  $\theta_j$  and  $\theta'_j$  are the factorized parameter matrices which are generally perturbed from the ground-truth, the resulting  $\psi^u$  may differ from the consistent permutation. To cope with these cases, we establish in § 5 the sufficient conditions for  $\psi^u$  to be consistent.

**Orthogonal Procrustes matching** One issue with the smallest angle matching is that each column is paired independently. It is easy for multiple columns to be paired with a common nearest neighbor. We describe a more robust algorithm based on the orthogonal Procrustes problem, and show improved guarantees. Since a consistent permutation is orthogonal, a natural relaxation is to only require the operator to be orthogonal. This is an orthogonal Procrustes problem, formulated in the same settings as § 4.2

$$\min_{\Psi} \|\bar{\theta}'_j \Psi - \psi^j \bar{\theta}_j\|_F^2, \quad \text{s.t. } \Psi^\top \Psi = \mathbf{I}. \quad (8)$$

Let  $\bar{\theta}'_j{}^\top \psi^j \bar{\theta}_j = \mathbf{U} \Sigma \mathbf{V}^\top$  be the singular value decomposition (SVD), the solution  $\Psi^*$  is given by the polar factor (Schönemann, 1966)

$$\Psi^* = \mathbf{U} \mathbf{V}^\top. \quad (9)$$

Here,  $\Psi^*$  is orthogonal and does not immediately imply the desired permutation  $\psi^u$ . To compute  $\psi^u$ , one can additionally restrict  $\Psi$  to be a permutation matrix, and solve for  $\psi^u$  using linear programming (Gower & Dijkstra, 1969). Aside from efficiency, one fundamental question is that under what assumptions the objective (8) yields the consistent permutation.

Given the solution  $\Psi^*$  to the Procrustes problem, we propose the following simple algorithm for computing  $\psi^u$ :

$$\psi_s^u = \arg \max_t \Psi_{ts}^*. \quad (10)$$

We first establish through Theorem 1 that if  $\psi^u$  obtained using (10) is a valid permutation, i.e., no duplicate entries, then it is optimal in terms of the objective (8).

**Theorem 1.** *The  $\psi^u$  obtained using (10) satisfies*

$$\|\psi^u \bar{\theta}'_j - \psi^j \bar{\theta}_j\|_F^2 \leq \|\bar{\theta}'_j - \psi^j \bar{\theta}_j\|_F^2$$

for all permutations  $\psi$ .

In section § 5 we state sufficient conditions under which the objective (8) yields a consistent permutation.



### 4.3. Approximate nonnegative factorization

In previous sections, we reduced the inference problem to factorizing partitioned sub-tensors. We now present a factorization algorithm for the sub-tensors that contain negative entries. Our goal is to approximate a sub-tensor  $\mathcal{M}$  by a sub-tensor  $\widetilde{\mathcal{M}} = \sum_j \mathbf{A}_j \times \mathbf{B}_j \times \mathbf{C}_j$  where the factors  $\mathbf{A}$ ,  $\mathbf{B}$ , and  $\mathbf{C}$  are nonnegative. The Frobenius norm is used to quantify the approximation

$$\min_{\mathbf{A}, \mathbf{B}, \mathbf{C} \succeq 0} \|\mathcal{M} - \widetilde{\mathcal{M}}\|_F. \quad (11)$$

Note that we do not assume that  $\mathcal{M} \succeq 0$  in (11) which distinguishes our optimization problem from other approximate factorization algorithms (Welling & Weber, 2001; Chi & Kolda, 2012; Kim et al., 2014; Shashua & Hazan, 2005). In § 4.4, we provide some details as to why negative entries are problematic for standard approximate factorization algorithms. We can rewrite (11) using the 1-mode unfolding as

$$\min_{\mathbf{A}, \mathbf{B}, \mathbf{C} \succeq 0} \|\mathcal{M}_{(1)} - \mathbf{A}(\mathbf{C} \odot \mathbf{B})^\top\|_F. \quad (12)$$

Equivalent formulations with respect to the 2-mode and 3-mode unfoldings can be readily obtained from (2).

We point out that another widely-used error measure — the I-divergence (Finesso & Spreij, 2006; Chi & Kolda, 2012) — may not be suitable for our learning problem. The optimization using I-divergence is given by

$$\min_{\mathbf{A}, \mathbf{B}, \mathbf{C} \succeq 0} \sum_{u,v,w} \left[ \mathcal{M}_{uvw} \log \frac{\mathcal{M}_{uvw}}{\widetilde{\mathcal{M}}_{uvw}} - \mathcal{M}_{uvw} + \widetilde{\mathcal{M}}_{uvw} \right].$$

This optimization is useful for nonnegative  $\mathcal{M}$  when each entry follows a Poisson distribution. In this case, the objective is equivalent to the sum of Kullback-Leibler divergence across all entries of  $\mathcal{M}$ :

$$\sum_{u,v,w} D_{\text{KL}} \left( \text{Pois}(x; \mathcal{M}_{uvw}) \parallel \text{Pois}(x; \widetilde{\mathcal{M}}_{uvw}) \right).$$

However, the Poisson assumption does not generally hold for the estimator tensor (4).

### 4.4. Handling negative entries in empirical estimators

We first illustrate that factorizing a tensor with negative entries using either positive tensor factorization (Welling & Weber, 2001) or nonnegative tensor factorization (Chi & Kolda, 2012; Shashua & Hazan, 2005) will either result in factors that violate the nonnegativity constraint or the result of the algorithm diverges. In addition, we show that general tensor decompositions cannot enforce the factor nonnegativity even after rounding the negative entries to zero.

We then present a simple method based on weighted nonnegative matrix factorization (WNMF) (Zhang et al., 1996) that enforce the factor nonnegativity constraint. We further generalize this method using parallel quadratic programming (PQP) (Brand & Chen, 2011) to obtain a method with a provable convergence rate.

**Issue of negative entries** If the tensor is strictly nonnegative, the optimization specified in (11) can be reduced to nonnegative matrix factorization (NMF). Solvers abound for NMF including the celebrated Lee-Seung’s multiplicative updates (Lee & Seung, 2001). The reduction is done by viewing (12) as  $\|Y - WH\|_F^2$  with  $Y = \mathcal{M}_{(1)}^{jst}$ ,  $W = \mathbf{A}$ , and  $H = (\mathbf{C} \odot \mathbf{B})^\top$ , and alternating

$$W_{st} \leftarrow W_{st} \frac{(YH^\top)_{st}}{(WHH^\top)_{st}}, \quad (13)$$

over each unfolding and factor matrix  $W$ . Obviously, the updates may yield negative entries in  $W$  when the unfolding contains negative entries. In addition, convergence relies on the nonnegativity of the unfolding (cf. Lee & Seung, 2001). This issue extends to their tensor factorization variants (Welling & Weber, 2001; Chi & Kolda, 2012; Kim et al., 2014) known as the positive tensor factorization and nonnegative tensor factorization. For these approaches, a naive resolution is to round negative entries of  $\widehat{\mathcal{M}}^{jst}$  to 0, this however lacks theoretical guarantees.

It is important to note that the rounding does not help general tensor decompositions like TPM. The following example illustrates that the unique decomposition (up to permutation and rescaling) of a positive tensor can contain negative entries. Consider a 2-by-2-by-2 positive tensor, whose 1-mode unfolding is given by

$$\left[ \begin{array}{cc|cc} 1 & 3 & 2 & 2 \\ 2 & 2 & 2 & 2 \end{array} \right],$$

where the vertical bar separates two frontal slices. It has the following decomposition, written in the form of (1):

$$\mathbf{A} = \mathbf{C} = \begin{bmatrix} 1 & 1 \\ 1 & 0 \end{bmatrix}, \quad \mathbf{B} = \begin{bmatrix} 2 & -1 \\ 2 & 1 \end{bmatrix}.$$

Since all factors are of full-rank, the decomposition is unique up to permutation and rescaling of columns (Kruskal, 1977). Thus, a general tensor decomposition yields a  $\mathbf{B}$  with negative entries regardless of rescaling.

### 4.5. Factorization via WNMF

Since the ground-truth  $\mathcal{M}^{jst}$  are nonnegative, we may “ignore” the negative entries of  $\widehat{\mathcal{M}}^{jst}$  by treating them as missing values. This idea leads to the following modified objective:

$$\min_{\mathbf{W}, \mathbf{H} \succeq 0} \|\Omega * (Y - WH)\|_F^2 \quad (14)$$

where  $Y, W, H$  are chosen identically as (13), and define the boolean  $\Omega_{uv} := Y_{uv} \geq 1$ . The optimization can be carried out using WNMf. Here, we modify the original updates by introducing a positive constant  $\epsilon$  to ensure that the updates are well-defined:

$$W_{uv} \leftarrow W_{uv} \frac{[(\Omega * Y) H^\top]_{uv} + \epsilon}{[(WH) * \Omega]_{uv} + \epsilon}. \quad (15)$$

It can be easily shown that the presence of  $\epsilon$  does not affect the solution accuracy. In addition, having  $\epsilon$  in both the numerator and denominator of (15) guarantees that the objective (14) is non-increasing under the updates.

#### 4.6. Parallel quadratic programming

We now generalize the WNMf approach using parallel quadratic programming to obtain a convergence rate. Let  $\mathbb{S}_{++}$  denote the set of symmetric positive definite matrices, we consider the following optimization problem

$$\min_x \frac{1}{2} x^\top Q x + z^\top x \quad \text{s.t.} \quad x \geq 0, Q \in \mathbb{S}_{++}, \quad (16)$$

which can be solved by iterating multiplicative updates (Brand & Chen, 2011; Sha et al., 2003). We use the parallel quadratic programming (PQP) algorithm (Brand & Chen, 2011; Brand et al., 2011) to solve (16), partly because it has a provable linear convergence rate. The PQP multiplicative update for (16) takes the following simple form:

$$x \leftarrow x * (Q^- x + z^-) \oslash (Q^+ x + z^+), \quad (17)$$

with

$$\begin{aligned} Q^+ &= (Q)_+ + \text{diag}(\gamma), & Q^- &= (-Q)_+ + \text{diag}(\gamma) \\ z^+ &= (z)_+ + \phi, & z^- &= (-z)_+ + \phi. \end{aligned}$$

Here  $\gamma$  and  $\phi$  are arguments to PQP, we will discuss these arguments in section § 5.2. The update maintains nonnegativity since all items are nonnegative. We make the following observation.

**Theorem 2.** *The multiplicative updates for Lee-Seung and WNMf are special cases of PQP.*

We can now solve the approximate nonnegative factorization problem stated in (11) using (17). Theorem 3 states the multiplicative updates. A more detailed discussion of  $\Phi$  is included in § 5.2. We present pseudo-code in Algorithm 1.

**Theorem 3.** *For optimization (11), the following update converges linearly to a local optimum*

$$A \leftarrow A * [(-Z)_+ + \Phi] \oslash [AQ + (Z)_+ + \Phi] \quad (18)$$

---

#### Algorithm 1 Factorize $(\mathcal{M}, k, d)$

---

```

 $\mathcal{M} \leftarrow \mathcal{M} / \max_{uvw} |\mathcal{M}_{uvw}|, \quad \epsilon \leftarrow 10^{-10}$ 
% Initialize with random nonnegative matrices:
 $A \leftarrow \text{rand}(d_j, k), B \leftarrow \text{rand}(d_s, k), C \leftarrow \text{rand}(d_t, k)$ 
% Create a set of alternating variable tuples:
 $F \leftarrow \{ [A, (C^\top C) * (B^\top B), -\mathcal{M}_{(1)}(C \odot B)] \}$ 
 $F \leftarrow F \cup \{ [B, (C^\top C) * (A^\top A), -\mathcal{M}_{(2)}(C \odot A)] \}$ 
 $F \leftarrow F \cup \{ [C, (B^\top B) * (A^\top A), -\mathcal{M}_{(3)}(B \odot A)] \}$ 
repeat
  for each  $[X, Q, Z]$  in  $F$  do
     $\Phi \leftarrow \lambda_{\min}^{-1/2}(Q) \sqrt{\text{diag}(ZQ^{-1}Z^\top) \text{diag}(Q)^\top}$ 
     $\Phi \leftarrow (\Phi - |Z|)_+ / 2 + \epsilon \mathbf{1}^\top$ 
     $X \leftarrow X * [(-Z)_+ + \Phi] \oslash [XQ + (Z)_+ + \Phi]$ 
  end for
until  $X$  ceased to change, or reached max #iterations
  Normalize the columns of  $A, B, C$  to sum to 1.
return  $A, B, C$ 

```

---

with

$$\begin{aligned} Q &= (C^\top C) * (B^\top B), & Z &= -\mathcal{M}_{(1)}(C \odot B) \\ \Phi &\succ \frac{1}{2} \left( \sqrt{\frac{\text{diag}(ZQ^{-1}Z^\top)}{\lambda_{\min}(Q)} \text{diag}(Q)^\top} - |Z| \right)_+ \end{aligned}$$

where  $\lambda_{\min}(\cdot)$  is the smallest eigenvalue. Similar updates for  $B$  and  $C$  are obtained using (2).

#### 4.7. Proposed approach

To summarize, the proposed approach, referred to as PT-PQP, consists of three steps. Given the indexes of anchor variables  $\pi^u \cup \pi^v \cup \pi^w$ , the variables  $[p] \setminus (\pi^u \cup \pi^v \cup \pi^w)$  are first evenly divided into  $r$  partitions, and the anchor variables are added to each partition. The second step consists of forming and factorizing the sub-tensor of each partition using Algorithm 1, this step can be parallelized. Third, normalize the anchor matrix  $[\theta^{\pi^u \top}, \theta^{\pi^v \top}, \theta^{\pi^w \top}]^\top$  formed by the anchor variable parameters to have unit column Euclidean norm, and then use either (7) or (10) to match over the anchor matrix.

**Efficiency.** Most of the computational cost is in the factorization. Consider one partition, and let  $\mathcal{M}^{\pi_j \pi_s \pi_t}$  be the corresponding sub-tensor, the sub-tensor size is  $\prod_{\pi \in \{\pi^j, \pi^s, \pi^t\}} \sum_{h \in \pi} d_h$ . The maximum number of categories for a variable is generally a constant for the GDLM. Under smallest partitioning, this size is determined by the sub-tensor of anchor variables, i.e.,  $O(k^3)$ , which corresponds to  $(p/k)$  partitions. One benefit of PTPQP is that the number of sub-tensor factorizations is linear in  $p$  due to the partitioned factorization, this results in significant efficiency gains when  $p \gg k$ . Furthermore, PTPQP is easy to

be parallelized across multiple CPUs and machines, since the computation as well as data are not distributed across partitions.

## 5. Provable Guarantees

In this section, we state the main theoretical results of the proposed partitioned factorization and tensor PQP factorization. The proofs can be found in the longer version of this paper (Tan & Mukherjee, 2017).

### 5.1. Sufficient conditions for guaranteed matching

Theorem 4 and Theorem 5 state that when the anchor parameter matrices from two factorizations are “close”, the proposed matching algorithms obtain a consistent permutation.

**Theorem 4.** *Suppose that  $\theta_j$  is the ground-truth matrix for variable  $j$ . Solving Equation (7) results in a consistent permutation if for all factors  $\hat{\theta}_j$  of variable  $j$*

$$\frac{\|\theta_{jh} - \hat{\theta}_{jh}\|_2}{\|\theta_{jh}\|_2} < 1 - \sqrt{\frac{1}{2} + \sqrt{\frac{1}{8} \left(1 + \max_{u < v} (\bar{\theta}_j^\top \bar{\theta}_j)_{uv}\right)}}$$

for all  $h \in [k]$ , where  $\bar{\theta}_{jh} = \theta_{jh} / \|\theta_{jh}\|_2$ .

Theorem 4 states that one obtains a consistent permutation by solving Equation (7) in the columns of the ground-truth parameter matrix are distinct from each other in angles and the factorized parameter matrix is near the ground-truth in Frobenius norm. Thus, a good anchor variable for the partitioned factorization (5) is one whose parameter matrix has distant columns in angles.

The bound in Theorem 4 can be made sharp for certain  $\theta_j$ , and thus the smallest angle matching algorithm has general guarantees only when the perturbation is small, i.e., the relative error ratio is less than  $1 - \sqrt{2} + \sqrt{2}/2 \approx 1/13$ .

**Theorem 5.** *Suppose that  $\theta$  and  $\theta'$  are two factorized parameter matrices for a variable. Solving (10) results in a consistent permutation  $\psi$ , if  $\|\mathbf{E}\|_2 < \sigma_k(\theta^\top \theta)$  and*

$$-\frac{\|\mathbf{E}\|_2}{\rho} \log\left(1 - \frac{\rho}{\nu}\right) < \frac{2 - \sqrt{2}}{4}$$

with

$$\rho = \sigma_1(\mathbf{E}) + \sigma_2(\mathbf{E}), \quad \nu = \sigma_k(\theta^\top \theta) + \sigma_{k-1}(\theta^\top \theta)$$

where the error matrix is define as  $\mathbf{E} = (\psi\theta)^\top (\theta' - \psi\theta)$ , and  $\sigma_j(\cdot)$  denotes the  $j$ -th largest singular value.

The first condition in Theorem 5 requires that at least one of  $\theta$  and  $\theta'$  must have full column rank. We may exchange

$\theta$  and  $\theta'$  in Theorem 5 to first obtain the consistent permutation of  $\theta'$  with respect to  $\theta$ ,  $\psi$  then follows immediately.

Theorem 5 states that solving (10) recovers a consistent permutation whenever the error spectral norm is small as compared to the smallest singular value of  $\theta^\top \theta$ . This is especially useful for  $\theta \in \mathbb{R}^{d \times k}$  with the number of rows  $d$  much larger than the number of columns  $k$ . In particular, for  $\theta$  with independent and identically distributed subgaussian entries,  $\sigma_k(\theta^\top \theta)$  is at least of the order  $(\sqrt{d} - \sqrt{k-1})^2$  (Rudelson & Vershynin, 2009).

### 5.2. Convergence

The following theorem states a sufficient condition for PQP to achieve linear convergence rate. The theorem statement and proof is an adaptation of results stated in (Brand & Chen, 2011)—the proof in (Brand & Chen, 2011) overlooks a required condition on  $\phi$  and the condition  $\gamma \succeq \text{diag}(Q_{jj})$  in the original proof is unnecessary.

**Theorem 6.** *The PQP algorithm given by (17) monotonically decreases the objective (16) and has linear convergence, if  $\gamma \succeq (-\mathbf{Q})_+ \mathbf{1}$  and*

$$\phi \succ \frac{1}{2} \left( \sqrt{\frac{\mathbf{z}^\top \mathbf{Q}^{-1} \mathbf{z}}{\lambda_{\min}(\mathbf{Q})} \text{diag}(\mathbf{Q}) - |\mathbf{z}|} \right)_+, \quad (19)$$

where  $\lambda_{\min}(\cdot)$  is the smallest eigenvalue.

## 6. Results on real and simulated data

We compare the proposed algorithm **ptpqp** with state-of-the-art approaches including: 1) the tensor power method **tpm** (Anandkumar et al., 2014) and matrix simultaneous diagonalization, **nojd0** and **nojd1** (Kuleshov et al., 2015)—two general tensor decomposition methods; 2) nonnegative tensor factorization **hals** (Kim et al., 2014); and 3) generalized method of moments **meld** (Zhao et al., 2016). We use the online code provided by the corresponding authors. The code to reproduce the experiments is available at: <https://goo.gl/3DBXIo>.

### 6.1. Learning GDLMs on simulated data

We adapt a simulation study from (Zhao et al., 2016) to compare runtime and accuracy of parameter estimation. We consider a GDLM where each variable takes categorical values  $\{0, 1, 2, 3\}$  and the parameters of the Dirichlet mixing distribution are  $\{\alpha_j = 0.1\}_{j=1}^k$ . We initially consider 25 variables. The true parameters for each hidden component  $h$  are drawn from the Dirichlet distribution  $\text{Dir}(0.5, 0.5, 0.5, 0.5)$ . The resulting moment estimator is a 100-by-100-by-100 tensor. We vary the number of components  $k$  and add noise by replacing a fraction  $\delta$

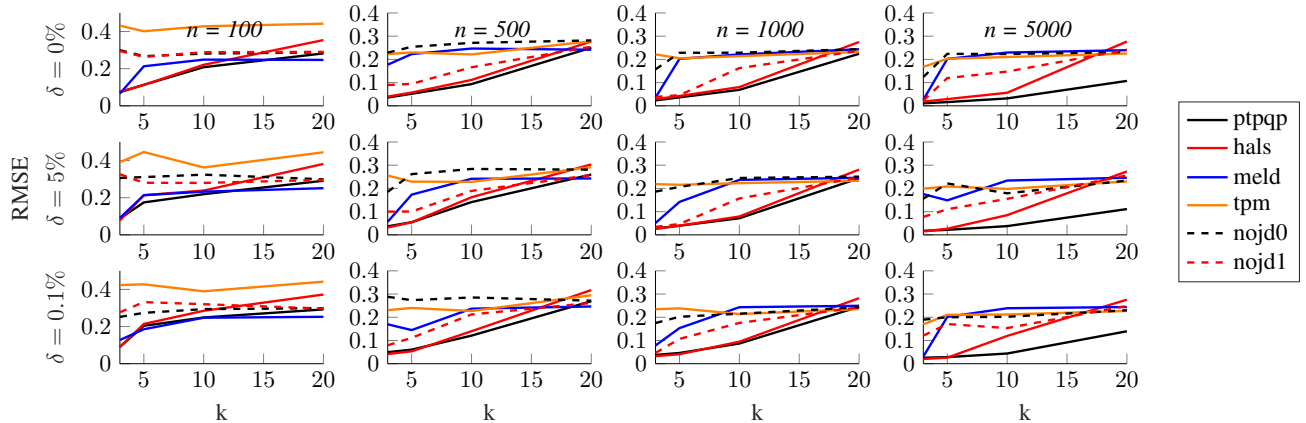


Figure 1. RMSE between inferred and true parameters.

of the observations with draws from a discrete uniform distribution. We also vary the number of samples  $n = 100, 500, 1000, 5000$ , number of clusters  $k = 3, 5, 10, 20$ , and contamination  $\delta = 0, 0.05, 0.1$ . Across these settings we found that the empirical third-order estimator typically exhibits between 20% and 50% negative entries.

**Accuracy of inference:** Accuracy is measured by root-mean-square error (RMSE) which we compare across algorithms as a function of the number of components for various sample sizes and levels of contamination, see Figure 1. Both **hals** and **ptpqp** are consistently among the top estimators, and **ptpqp** outperforms **hals** as  $n$  grows. For small sample sizes and many hidden components **meld** achieves the smallest RMSE. The RMSE of **tpm** is relatively large, probably due to the whitening technique used to approximately transform the nonorthogonal factorization into an orthogonal one, see (Souloumiac, 2009; Nicolò Colombo and Nikos Vlassis). The most relevant observation is that **ptpqp** outperforms other methods for large, noisy data.

**Computational cost:** We examined how runtime scales as a function of the number of partitions. For the same model we set  $p = 1000$  variables and  $n = 1000$  samples. The tensor is now 4000-by-4000-by-4000. We evaluated the runtime of **ptpqp** (without parallelization) with the number of partitions set to  $\{30, 40, 50, 100, 200\}$ . On a laptop with Intel i7-4702HQ@2.20GHz CPU and 8GB memory, **ptpqp** with 100 partitions completes within 3.5 min, 4 min, and 5 min for  $k = 4, 8, 12$ , respectively. In addition, the runtime monotonically decreases with the number of partitions. Further speedups can be obtained by parallelizing the factorization of partitions across multiple CPUs or machines.

## 6.2. Predicting crowdsourced labels

In (Zhang et al., 2014), a combination of EM and tensor decompositions was used to predict crowdsourcing annotations. The task is to predict the true label given incom-

plete and noisy observations from a set of workers, this is a mixed membership problem (Dawid & Skene, 1979). In (Zhang et al., 2014) a third-order tensor estimator was proposed to obtain an initial estimate for the EM algorithm. We compare the predictive performance on five data sets of several tensor decomposition methods as well as the EM algorithm initialized with majority voting by the workers (**MV+EM**). The fraction of incorrect predictions and the size of each dataset are in the table below. Note that **ptpqp** matches or outperforms the other tensor methods on all but one dataset, and even outperforms **MV+EM** on two datasets.

Table 1. Incorrectly predicted labels (%)

DATASET	BIRDS	RTE	TREC	DOGS	WEB
PTPQP	<b>11.11</b>	7.75	30.81	<b>15.37</b>	<b>14.44</b>
HALS	12.96	7.75	31.47	20.57	26.84
TPM	<b>11.11</b>	7.62	31.87	15.49	14.70
NOJD0	12.04	8.00	32.97	15.49	18.39
NOJD1	12.04	8.00	35.91	15.86	25.97
MV+EM	<b>11.11</b>	<b>7.12</b>	<b>30.20</b>	15.86	15.91
SIZE	108	800	19033	807	2665

## 7. Conclusions

We proposed an efficient algorithm for learning mixed mixture models based on the idea of partitioned factorizations. The key challenge is to consistently match the partitioned parameters with the hidden components. We provided sufficient conditions to ensure consistency. In addition, we have also developed a nonnegative approximation to handle the negative entries in the empirical method of moments estimators, a problem not addressed by several recent tensor methods. Results on synthetic and real data corroborate that the proposed approach achieves improved inference accuracy as well as computational efficiency than state-of-the-art methods.



## Acknowledgements

Z.T. would like to thank Rong Ge for sharing helpful insights. S.M. would like to thank Lek-Heng Lim for insights. Z.T. would like to acknowledge the support of grants NSF CNS-1423128, NSF IIS-1423124, and NSF CNS-1218981. S.M. would like to acknowledge the support of grants NSF IIS-1546331, NSF DMS-1418261, NSF IIS-1320357, NSF DMS-1045153, and NSF DMS-1613261.

## References

- Anandkumar, Anima, Foster, Dean P, Hsu, Daniel J, Kakade, Sham M, and kai Liu, Yi. A spectral algorithm for latent dirichlet allocation. In *NIPS*, pp. 917–925. 2012.
- Anandkumar, Anima, Jain, Prateek, Shi, Yang, and Niranjan, U. N. Tensor vs. matrix methods: Robust tensor decomposition under block sparse perturbations. In *AIS-TATS*, pp. 268–276, 2016.
- Anandkumar, Animashree, Ge, Rong, Hsu, Daniel, Kakade, Sham M., and Telgarsky, Matus. Tensor decompositions for learning latent variable models. *JMLR*, 15(1):2773–2832, January 2014.
- Blei, David M., Ng, Andrew Y., and Jordan, Michael I. Latent Dirichlet allocation. *JMLR*, 3:993–1022, 2003.
- Brand, M. and Chen, D. Parallel quadratic programming for image processing. In *IEEE International Conference on Image Processing (ICIP)*, pp. 2261–2264, September 2011.
- Brand, M., Shilpiekandula, V., and Bortoff, S.A. A parallel quadratic programming algorithm for model predictive control. In *World Congress of the International Federation of Automatic Control (IFAC)*, volume 18, August 2011.
- Chi, Eric C. and Kolda, Tamara G. On tensors, sparsity, and nonnegative factorizations. *SIAM Journal on Matrix Analysis and Applications*, 33(4):1272–1299, December 2012.
- Comon, Pierre and Jutten, Christian (eds.). *Handbook of blind source separation : independent component analysis and applications*. Communications engineering. Elsevier, 2010.
- Dawid, A. P. and Skene, A. M. Maximum likelihood estimation of observer error-rates using the em algorithm. *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, 28(1):20–28, 1979.
- Donoho, David and Stodden, Victoria. When does non-negative matrix factorization give a correct decomposition into parts? In *NIPS*, pp. 1141–1148. 2004.
- Erosheva, Elena A. Comparing latent structures of the grade of membership, rasch, and latent class models. *Psychometrika*, 70(4):619–628, 2005.
- Finesso, Lorenzo and Spreij, Peter. Nonnegative matrix factorization and  $I$ -divergence alternating minimization. *Linear Algebra and its Applications*, 416(2):270–287, July 2006.
- Gower, J.C. and Dijksterhuis, G.B. *Procrustes Problems*. Oxford Statistical Science Series. OUP Oxford. ISBN 9780198510581.
- Gu, Quanquan, Gui, Huan, and Han, Jiawei. Robust tensor decomposition with gross corruption. In *NIPS*, pp. 1422–1430. 2014.
- Hsu, Daniel and Kakade, Sham M. Learning mixtures of spherical gaussians: Moment methods and spectral decompositions. In *Innovations in Theoretical Computer Science (ITCS)*, pp. 11–20, 2013.
- Huang, Furong, Niranjan, U. N., Hakeem, Mohammad Umar, Verma, Prateek, and Anandkumar, Animashree. Fast detection of overlapping communities via online tensor methods on gpus. *CoRR*, abs/1309.0787, 2013. URL <http://arxiv.org/abs/1309.0787>.
- Kim, Jingu, He, Yunlong, and Park, Haesun. Algorithms for nonnegative matrix and tensor factorizations: a unified view based on block coordinate descent framework. *Journal of Global Optimization*, 58(2):285–319, 2014.
- Kolda, Tamara G. and Bader, Brett W. Tensor decompositions and applications. *SIAM Rev.*, 51(3):455–500, August 2009.
- Kruskal, Joseph B. Three-way arrays: Rank and uniqueness of trilinear decompositions, with application to arithmetic complexity and statistics. *Linear Algebra and Its Applications*, 18(2):95–138, 1977.
- Kuleshov, V., Chaganty, A., and Liang, P. Tensor factorization via matrix factorization. In *AISTATS*, 2015.
- Lee, Daniel D. and Seung, H. Sebastian. Algorithms for non-negative matrix factorization. In *NIPS*, pp. 556–562, 2001.
- Leurgans, S. E., Ross, R. T., and Abel, R. B. A decomposition for three-way arrays. *SIAM Journal on Matrix Analysis and Applications*, 14:1064–1083, 1993.

- Nicolò Colombo and Nikos Vlassis, title = Tensor Decomposition via Joint Matrix Schur Decomposition, booktitle = ICML pages = 2820–2828 year = 2016.
- Pritchard, Jonathan K., Stephens, Matthew, and Donnelly, Peter. Inference of population structure using multilocus genotype data. *Genetics*, 155(2):945–959, 2000a.
- Pritchard, Jonathan K., Stephens, Matthew, Rosenberg, Noah A., and Donnelly, Peter. Association mapping in structured populations. *The American Journal of Human Genetics*, 67(1):170–181, 2000b.
- Rudelson, Mark and Vershynin, Roman. Smallest singular value of a random rectangular matrix. *Comm. Pure Appl. Math*, pp. 1707–1739, 2009.
- Schönemann, Peter. A generalized solution of the orthogonal procrustes problem. *Psychometrika*, 31(1):1–10, 1966.
- Sha, Fei, Saul, Lawrence K., and Lee, Daniel D. Multiplicative updates for nonnegative quadratic programming in support vector machines. In *NIPS*, pp. 1065–1072. 2003.
- Shashua, Amnon and Hazan, Tamir. Non-negative tensor factorization with applications to statistics and computer vision. In *ICML*, pp. 792–799, 2005.
- Song, Zhao, Woodruff, David P., and Zhang, Huan. Sub-linear time orthogonal tensor decomposition. In *NIPS*, pp. 793–801, 2016.
- Souloumiac, A. Joint diagonalization: is non-orthogonal always preferable to orthogonal? In *3rd International Workshop on Computational Advances in Multi-Sensor Adaptive Processing*, pp. 305–308, 2009.
- Tan, Zilong and Mukherjee, Sayan. Efficient learning of graded membership models. *CoRR*, abs/1702.07933, 2017. URL <http://arxiv.org/abs/1702.07933>.
- Wang, Yining, Tung, Hsiao-Yu, Smola, Alexander J, and Anandkumar, Anima. Fast and guaranteed tensor decomposition via sketching. In *NIPS*, pp. 991–999. 2015.
- Welling, Max and Weber, Markus. Positive tensor factorization. *Pattern Recognition Letters*, 22:1255–1261, 2001.
- Woodbury, Max A., Clive, Jonathan, and Garson, Arthur, Jr. Mathematical typology: A grade of membership technique for obtaining disease definition. *Computers and Biomedical Research*, 11:277–298, 1978.
- Zhang, Sheng, Wang, Weihong, Ford, James, and Makedon, Fillia. Learning from incomplete ratings using non-negative matrix factorization. In *In Proceedings of the 6th SIAM Conference on Data Mining (SDM)*, pp. 549–553, 1996.
- Zhang, Yuchen, Chen, Xi, Zhou, Dengyong, and Jordan, Michael I. Spectral methods meet EM: A provably optimal algorithm for crowdsourcing. In *NIPS*, pp. 1260–1268, 2014.
- Zhao, Shiwen, Engelhardt, Barbara E., Mukherjee, Sayan, and Dunson, David B. Fast moment estimation for generalized latent dirichlet models. *CoRR*, abs/1603.05324, 2016. URL <http://arxiv.org/abs/1603.05324>.