Assigning Topics to Documents by Successive Projections

Olga Klopp, Maxim Panov, Suzanne Sigalla, Alexandre B Tsybakov

Assigning Topics to Documents by Successive Projections

Olga Klopp[1,3], Maxim Panov[2], Suzanne Sigalla[3], and Alexandre B. Tsybakov[3]

[1]ESSEC Business School, France
[2]Skoltech, Russia
[3]CREST, ENSAE, France

July 7, 2021

Abstract

Topic models provide a useful tool to organize and understand the structure of large corpora of text documents, in particular, to discover hidden thematic structure. Clustering documents from big unstructured corpora into topics is an important task in various areas, such as image analysis, e-commerce, social networks, population genetics. A common approach to topic modeling is to associate each topic with a probability distribution on the dictionary of words and to consider each document as a mixture of topics. Since the number of topics is typically substantially smaller than the size of the corpus and of the dictionary, the methods of topic modeling can lead to a dramatic dimension reduction. In this paper, we study the problem of estimating topics distribution for each document in the given corpus, that is, we focus on the clustering aspect of the problem. We introduce an algorithm that we call Successive Projection Overlapping Clustering (SPOC) inspired by the Successive Projection Algorithm for separable matrix factorization. This algorithm is simple to implement and computationally fast. We establish theoretical guarantees on the performance of the SPOC algorithm, in particular, near matching minimax upper and lower bounds on its estimation risk. We also propose a new method that estimates the number of topics. We complement our theoretical results with a numerical study on synthetic and semi-synthetic data to analyze the performance of this new algorithm in practice. One of the conclusions is that the error of the algorithm grows at most logarithmically with the size of the dictionary, in contrast to what one observes for Latent Dirichlet Allocation.

Keywords— Topic model, latent variable model, nonnegative matrix factorization, adaptive estimation

1 Introduction

Assigning topics to documents is an important task in several applications. For example, press agencies need to identify articles of interest to readers based on the topics of articles that they have read in the past. Analogous goals are pursued by many other text-mining applications such as, for example, recommending blogs from among the millions of blogs available. A popular approach to the problem of estimating hidden thematic structures in a corpus of documents is based on topic modeling. Topic models have attracted a great deal of attention in the past two decades. Beyond text mining, they were used in areas, such as population genetics [7, 38], social networks [14, 30], image analysis [26, 48], e-commerce [33, 46].

In this paper, we adopt the *probabilistic Latent Semantic Indexing* (pLSI) model introduced in [21]. The pLSI model deals with three types of variables, namely, documents, topics and words. Topics are latent variables, while the observed variables are words and documents. Assume that we have a dictionary of p words and a collection of n documents. Documents are sequences of words from the dictionary. The number of topics is denoted by K. Usually, $K \ll \min(p, n)$. Throughout this paper, we assume that $2 \le K \le \min(p, n)$. The pLSI model assumes that the probability of occurrence of word j in a document discussing topic k is independent of the document. Therefore, by the total probability formula,

$$\mathbb{P}(\text{word } j | \text{document } i) = \sum_{k=1}^{K} \mathbb{P}(\text{topic } k | \text{document } i) \mathbb{P}(\text{word } j | \text{topic } k).$$

Introducing the notation $\Pi_{ij} := \mathbb{P}(\text{word } j | \text{document } i)$, $W_{ik} := \mathbb{P}(\text{topic } k | \text{document } i)$ and $A_{kj} := \mathbb{P}(\text{word } j | \text{topic } k)$ we may write $\Pi_{ij} = W_i^{\mathrm{T}} A_j$, where $W_i = (W_{i1}, \ldots, W_{iK})^{\mathrm{T}} \in [0,1]^K$ is the topic probability vector for document i and $A_j = (A_{1j}, \ldots, A_{Kj})^{\mathrm{T}} \in [0,1]^K$ is the vector of word j probabilities under topics $k = 1, \ldots, K$. In matrix form,

$$\boldsymbol{\Pi} = \boldsymbol{W}\boldsymbol{A}, \tag{1}$$

where $\boldsymbol{\Pi}$ is the document-word matrix of size $n \times p$ with entries Π_{ij}, $\boldsymbol{W} := (W_1, \ldots, W_n)^{\mathrm{T}}$ is the document-topic matrix of size $n \times K$ and $\boldsymbol{A} := (A_1, \ldots, A_p)$ is the topic-word matrix of size $K \times p$. The rows of these matrices are probability vectors,

$$\sum_{m=1}^{K} W_{im} = 1, \ \ \sum_{j=1}^{p} A_{kj} = 1, \ \ \sum_{j=1}^{p} \Pi_{ij} = 1 \text{ for any } i = 1, \ldots, n, \ k = 1, \ldots K. \tag{2}$$

Unless otherwise stated, we will assume throughout the paper that $\boldsymbol{\Pi}, \boldsymbol{W}, \boldsymbol{A}$ are matrices with non-negative entries satisfying (2).

The value Π_{ij} is the probability of occurrence of word j in document i. It is not available but we have access to the corresponding empirical frequency X_{ij}. Thus, we have a document-word matrix $\boldsymbol{X} = (X_{ij})$ of size $n \times p$ such that for each document i in $1, \ldots, n$, and each word j in $1, \ldots, p$, the entry X_{ij} is the observed frequency of word j in document i. Let N_i denote the (non-random) number of sampled words in document i. In what follows, we assume that, for each document-word vector $X_i = (X_{i1}, \ldots, X_{ip})^{\mathrm{T}}$, the corresponding vector of cumulative counts $N_i X_i$ follows a $\text{Multinomial}_p(N_i, \Pi_i)$ distribution, where $\Pi_i := \mathbb{E}(X_i) = (\Pi_{i1}, \ldots, \Pi_{ip})^{\mathrm{T}}$. We also assume that $X_1, \ldots, X_n$ are independent random vectors. We will denote by $\mathbb{P}_{\boldsymbol{\Pi}}$ the probability measure corresponding to the distribution of $\boldsymbol{X}$.

We can write the observation model in a "signal + noise" form:

$$\boldsymbol{X} = \boldsymbol{\Pi} + \boldsymbol{Z} = \boldsymbol{W}\boldsymbol{A} + \boldsymbol{Z}, \tag{3}$$

where $\boldsymbol{Z} := \boldsymbol{X} - \boldsymbol{\Pi}$ is a zero mean noise. The objective in topic modeling is to estimate matrices $\boldsymbol{A}$ and $\boldsymbol{W}$ based on the observed frequency matrix $\boldsymbol{X}$ and on the known $N_1, \ldots, N_n$. The recovery of $\boldsymbol{A}$ and the recovery of $\boldsymbol{W}$ address different purposes. An estimator of matrix $\boldsymbol{A}$ identifies the topic distribution on the dictionary. An estimator of $\boldsymbol{W}$ indicates the topics associated to each document.

Estimation of $\boldsymbol{W}$ has multiple applications and has been extensively discussed in the literature, mainly in the Bayesian perspective. The focus was on Latent Dirichlet Allocation (LDA) and related techniques (see Section 5 for more details and references). These methods are computationally slow and, to the best of our knowledge, no theoretical guarantees on their performance are available.

On the other hand, estimation of matrix $\boldsymbol{A}$ is well-studied in the theory. Several papers provide bounds on the performance of different estimators of $\boldsymbol{A}$. We give a more detailed account on this work in Section 5. Most of the results use the *anchor word assumption* postulating that for every topic there is at least one word, which occurs only in this topic, see [4, 8, 22]. At first sight, it seems that results on estimation of

Table 1: The first ten rows of the estimated matrix $\boldsymbol{W}$ for the Associated Press data set.

Document	Finance	Politics
1	0.248	0.752
2	0.362	0.638
3	0.527	0.473
4	0.357	0.643
5	0.181	0.819
6	0.001	0.999
7	0.773	0.227
8	0.004	0.996
9	0.967	0.033
10	0.147	0.953

matrix $\boldsymbol{A}$ can be applied to estimation of $\boldsymbol{W}$ by simply taking the transpose of (2) and inverting the roles of these two matrices. However, such an argument is not valid since the resulting models are different. Indeed, the rows of matrix $\boldsymbol{X}^{\mathrm{T}}$ are not independent and the rows of matrices $\boldsymbol{\Pi}^{\mathrm{T}}, \boldsymbol{A}^{\mathrm{T}}, \boldsymbol{W}^{\mathrm{T}}$ do not sum up to 1, which leads to a different statistical analysis.

In the present paper, we change the framework compared to [4, 8, 22] by focusing on estimation of matrix $\boldsymbol{W}$ rather than $\boldsymbol{A}$. We introduce the following assumption.

Assumption 1 (Anchor document assumption)**.** *For each topic $k = 1, \ldots, K$, there exists at least one document i (called an anchor document) such that $W_{ik} = 1$ and $W_{il} = 0$ for all $l \neq k$.*

Both anchor word and anchor document assumptions are very relevant in real word applications. Since each document is identified with a mixture of K topics, anchor document assumption means that, for each topic, there is a document devoted solely to this topic. To illustrate the anchor document assumption, consider the Associated Press data set [20], which is a collection of 2246 articles published by this press agency mostly around 1988. An application of the pLSI model fitted via the LDA method with $K = 2$ leads to two well-shaped topics "finance" and "politics". We refer to [41, Section 6] for the details of the analysis. The first 10 rows of the estimator of matrix $\boldsymbol{W}$ are presented in Table 1. Notice that documents 6 and 8 in Table 1 can be considered as anchor documents. For example, document 6 has the weight of the second topic estimated as 0.999. A closer look at the most frequent words in this document (Noriega, Panama, Jackson, Powell, administration, economic, general) tells us that, indeed, this article corresponds solely to the topic "politics" – it is about the relationship between the American government and the Panamanian leader Manuel Noriega.

Our approach to estimation of matrix $\boldsymbol{W}$ that we call Successive Projection Overlapping Clustering (SPOC) is inspired by the Successive Projection Algorithm (SPA) initially proposed for non-negative matrix factorization [3] and further used by [17, 29, 34] in the context of mixed membership stochastic block models. The idea of such methods is to start with the singular value decomposition (SVD) of matrix $\boldsymbol{X}$, and launch an iterative procedure that, at each step, chooses the maximum norm row of the matrix composed of singular vectors and then projects on the linear subspace orthogonal to the selected row. From a geometric perspective, the rows of the matrix composed of singular vectors of $\boldsymbol{\Pi}$ belong to a simplex in $\mathbb{R}^K$. The documents can be identified with some points in this simplex and the anchor documents with its vertices. Our algorithm iteratively finds estimators of the vertices, based on which we finally estimate $\boldsymbol{W}$.

Note that the idea of exploiting simplex structures for estimation of matrix $\boldsymbol{A}$ rather than $\boldsymbol{W}$ was previously developed in, for example, [4, 15, 22], among others. For example, the method to estimate $\boldsymbol{A}$ suggested in [22] is based on an exhaustive search over all size K subsets of $\{1, \ldots, p\}$. Its goal is to select K vertices of a p-dimensional simplex and its computational cost is at least of the order p^K. Our algorithm for

estimating $\boldsymbol{W}$ recovers the vertices of much less complex object, which is a K-dimensional simplex (recall that $K \ll p$), and its computational cost is of the order $\max(p, n)K + nK^2$. Here, $\max(p, n)K$ and nK^2 are the costs of performing a truncated singular value decomposition and the SPA algorithm, respectively. Another important point is that existing simplex-based methods for estimation of matrix $\boldsymbol{A}$ require the number K of topics to be known. In the present paper, we propose a procedure that is adaptive to unknown K.

Our theoretical results deal only with the problem of estimating the topic-document matrix $\boldsymbol{W}$, for which the theory was not developed in prior work. But in practice, our method can be used for estimation of matrix $\boldsymbol{A}$ as well. Based on the SPOC estimator of $\boldsymbol{W}$, we immediately obtain an estimator of matrix $\boldsymbol{A}$ by a computationally fast procedure (see Section 3). Our simulation studies (see Section A of the Appendix) indicate that this estimator exhibits a behavior similar to LDA on average while being more stable.

In this paper, we prove that the SPOC estimator of $\boldsymbol{W}$ converges in the Frobenius norm and in the ℓ_1-norm with the rates $\sqrt{n/N}$ and $n/\sqrt{N}$ (up to a weak factor[1]), respectively, assuming that $N_i = N$ for $i = 1, \ldots, n$. We also prove lower bounds of the order $\sqrt{n/N}$ and $n/\sqrt{N}$, respectively, implying near optimality of the proposed method. One of the conclusions, both from the theory and the numerical experiments, is that the error of the SPOC algorithm does not grow significantly with the size of the dictionary p, in contrast to what one observes for Latent Dirichlet Allocation. We also introduce an estimator for the number K of topics, which is usually unknown in practice. We show that SPOC algorithm using the estimator of K preserves its optimal properties in this more challenging setting.

The rest of the paper is organized as follows. In Section 3, we introduce the SPOC algorithm. Section 4 contains the main results on the convergence rate of the algorithm and the minimax lower bound for estimation of $\boldsymbol{W}$. Section 5 is devoted to discussion of the prior work. In Section 6, we present numerical experiments for synthetic and real-world data in order to illustrate our theoretical findings. Finally, in Section 7 we summarize the outcomes of the study. The proofs are given in the Appendix.

2 Notation

For any matrix $\boldsymbol{M} = (M_{ij}) \in \mathbb{R}^{n \times k}$, we denote by $\|\boldsymbol{M}\|$ its spectral norm, i.e., its maximal singular value, by $\|\boldsymbol{M}\|_F$ its Frobenius norm, and by $\|\boldsymbol{M}\|_1 = \sum_{i=n}^{k} \sum_{j=1}^{k} |M_{ij}|$ its ℓ_1-norm. We also consider the maximum ℓ_1-norm of its rows $\|\boldsymbol{M}\|_{1,\infty} = \max_{1 \leq i \leq n} \sum_{j=1}^{k} |M_{ij}|$. We denote by $\lambda_j(\boldsymbol{M})$ the jth singular value and by $\lambda_{\min}(\boldsymbol{M})$ the smallest singular value of $\boldsymbol{M}$. Assuming that matrix $\boldsymbol{M}$ has rank K we consider its singular values $\lambda_1(\boldsymbol{M}) \geq \lambda_2(\boldsymbol{M}) \geq \cdots \geq \lambda_K(\boldsymbol{M}) > 0$ and its condition number $\kappa(\boldsymbol{M}) = \lambda_1(\boldsymbol{M})/\lambda_K(\boldsymbol{M})$. If J is a non-empty subset of rows of matrix $\boldsymbol{M}$ the notation $\boldsymbol{M}_J$ is used for a matrix in $\mathbb{R}^{|J| \times k}$ obtained from $\boldsymbol{M}$ by keeping only the rows in J. We denote by $\boldsymbol{I}_K$ the $K \times K$ identity matrix, and by $(\mathbf{e}_1, \ldots, \mathbf{e}_n)$ the canonical basis of $\mathbb{R}^n$. For any vector $u \in \mathbb{R}^d$, we denote by $\|u\|_2$ its Euclidean norm. Throughout the paper, we use the notation $\mathbf{O}$ for orthogonal matrices and $\mathbf{P}$ for permutation matrices. We denote by $\mathscr{P}$ the set of all permutation matrices in $\mathbb{R}^{K \times K}$. We denote by c, C positive constants than may vary from line to line

3 Successive Projection Overlapping Clustering

In this section, we introduce the *Successive Projection Overlapping Clustering (SPOC)* algorithm for estimation of matrix $\boldsymbol{W}$. It is an analog, in the context of topic models, of the algorithm proposed in [34] for the problem of parameter estimation in Mixed Membership Stochastic Block Model.

---

[1] In what follows, we mean by *weak factor* a small power of K multiplied by a term logarithmic in the parameters of the problem. We will ignore weak factors when discussing the convergence rates.

In order to explain the main idea of the algorithm, we start with considering the singular value decomposition (SVD) of matrix $\mathbf{\Pi}$:

$$\mathbf{\Pi} = \mathbf{U}\mathbf{L}\mathbf{V}^{\mathrm{T}}, \tag{4}$$

where $\mathbf{U} = [U_1, \ldots, U_K] \in \mathbb{R}^{n \times K}$ and $\mathbf{V} = [V_1, \ldots, V_K] \in \mathbb{R}^{p \times K}$ are the matrices of left and right singular vectors and $\mathbf{L} \in \mathbb{R}^{K \times K}$ is a diagonal matrix of the corresponding singular values. We have used here the fact that the rank of matrix $\mathbf{\Pi}$ is at most K. Recall that we assume $K \leq \min(p, n)$. A key observation is that, if $\lambda_K(\mathbf{\Pi}) > 0$ and Assumption 1 is satisfied, then matrix $\mathbf{U}$ can be represented as

$$\mathbf{U} = \boldsymbol{W}\mathbf{H}, \tag{5}$$

where $\mathbf{H} \in \mathbb{R}^{K \times K}$ is a full rank matrix (cf. Lemma 6 in the Appendix). Thus, the rows of matrix $\mathbf{U}$ belong to a simplex in $\mathbb{R}^K$ with vertices given by the rows of matrix $\mathbf{H}$.

The empirical counterparts of $\mathbf{U}, \mathbf{L}, \mathbf{V}$ are obtained from the SVD of the observed matrix $\boldsymbol{X}$:

$$\boldsymbol{X} = \hat{\mathbf{U}}\hat{\mathbf{L}}\widehat{\mathbf{V}}^{\mathrm{T}} + \hat{\mathbf{U}}_1\hat{\mathbf{L}}_1\widehat{\mathbf{V}}_1^{\mathrm{T}}, \tag{6}$$

where $\hat{\mathbf{U}} = [\widehat{U}_1, \ldots, \widehat{U}_K]$ and $\widehat{\mathbf{V}} = [\widehat{V}_1, \ldots, \widehat{V}_K]$ are, respectively, the matrices of left and right singular vectors of $\boldsymbol{X}$ corresponding to its K leading singular values $\widehat{\lambda}_1 \geq \cdots \geq \widehat{\lambda}_K$; $\widehat{\mathbf{L}} = \mathrm{diag}\{\widehat{\lambda}_1, \ldots, \widehat{\lambda}_K\}$, and $\hat{\mathbf{U}}_1\hat{\mathbf{L}}_1\widehat{\mathbf{V}}_1^{\mathrm{T}}$ is the singular value decomposition of $\boldsymbol{X} - \hat{\mathbf{U}}\hat{\mathbf{L}}\widehat{\mathbf{V}}^{\mathrm{T}}$. Due to matrix perturbation results (see Appendix A.1), there exists an orthogonal matrix $\mathbf{O}$ such that $\hat{\mathbf{U}}$ is a good approximation for $\mathbf{U}\mathbf{O}$. We can write

$$\hat{\mathbf{U}} = \mathbf{U}\mathbf{O} + \mathbf{N} = \boldsymbol{W}\mathbf{H}\mathbf{O} + \mathbf{N}, \tag{7}$$

where $\mathbf{N}$ is a "small enough" noise matrix. Having obtained $\hat{\mathbf{U}}$ from the SVD of $\boldsymbol{X}$, we then apply the *Successive Projection Algorithm* (SPA) [3, 17] to estimate the matrix $\mathbf{H}\mathbf{O}$ in model (7).

Algorithm 1 SPA

Input: Matrix $\boldsymbol{M} \in \mathbb{R}^{n \times K}$ and integer $r \leq n$.
Output: Set of indices $J \subseteq \{1, \ldots, n\}$.
1: Initialize: $\boldsymbol{S}_0 = \boldsymbol{M}^{\mathrm{T}}$, $J_0 = \emptyset$.
2: For $t = 1, \ldots, r$ do:
  – Find $i(t) = \arg\max_{i=1,\ldots,n} \|\mathbf{s}_i\|_2$, where $\mathbf{s}_i$'s are the column vectors of $\boldsymbol{S}_{t-1}$.
  – Set $\boldsymbol{S}_t = \left(\boldsymbol{I}_K - \frac{\mathbf{s}_{i(t)}\mathbf{s}_{i(t)}^{\mathrm{T}}}{\|\mathbf{s}_{i(t)}\|_2^2}\right)\boldsymbol{S}_{t-1}$, $\quad J_t = J_{t-1} \cup \{i(t)\}$.
3: Set $J = J_r$.

Applied to matrix $\boldsymbol{M} = \hat{\mathbf{U}}$ and $r = K$ this algorithm finds the rows of matrix $\hat{\mathbf{U}}$ with the maximum Euclidean norm and then projects on the subspace orthogonal to these rows and repeats the procedure until K rows are selected. The main idea underlying the SPA is that the maximum of the Euclidean norm of a vector on a simplex is attained at one of its vertices.

In the noiseless case when $\mathbf{N} = 0$, it can be shown that if Assumption 1 holds then $\hat{\mathbf{U}}_J = \mathbf{H}\mathbf{O}$, where J is the set of K rows of $\hat{\mathbf{U}}$ selected after K steps of SPA. In the noisy case, we need additional assumptions on the noise level to ensure that SPA extracts documents close to anchor ones, which leads to an accurate enough estimator $\hat{\mathbf{H}}$ of $\mathbf{H}\mathbf{O}$ (see Appendix A.2 for the precise statement). Once we have such an $\hat{\mathbf{H}}$, the final step is to define the following estimator of matrix $\boldsymbol{W}$:

$$\hat{\boldsymbol{W}} = \hat{\mathbf{U}}\hat{\mathbf{H}}^{-1}.$$

This definition is valid only if matrix $\hat{\mathbf{H}}$ is non-degenerate, which is true with high probability under suitable assumptions (cf. Section 4).

An additional potentially useful step is to apply preconditioning to matrix $\hat{\mathbf{U}}$, which leads to improved bounds on the performance of the algorithm in the presence of noise, see [18, 32]. Preconditioned SPA is defined as follows. Let $r = K$ and let $\mathbf{a}_1, \ldots, \mathbf{a}_n$ be the column vectors of matrix $\boldsymbol{M}^{\mathrm{T}}$. Let $\boldsymbol{L}^* \in \mathbb{R}^{K \times K}$ be the solution of the minimization problem

$$\min_{\boldsymbol{L} \succ 0: \ \max_i \mathbf{a}_i^{\mathrm{T}} \boldsymbol{L} \mathbf{a}_i \leq 1} - \log \det \boldsymbol{L}. \tag{8}$$

Matrix $\boldsymbol{L}^*$ defines the minimum volume ellipsoid centered at the origin that contains $\mathbf{a}_1, \ldots, \mathbf{a}_n$. The preconditioned SPA is defined by Algorithm 1 initialized with $\boldsymbol{S}_0 = (\boldsymbol{L}^*)^{1/2} \boldsymbol{M}^{\mathrm{T}}$ rather than with $\boldsymbol{S}_0 = \boldsymbol{M}^{\mathrm{T}}$.

The SPOC algorithm for topic modeling is summarized in Algorithm 2.

Algorithm 2 SPOC (respectively, preconditioned SPOC)

Input: Observed matrix $\boldsymbol{X}$ and number of topics K.
Output: Estimated document-topic matrix $\hat{\boldsymbol{W}}$.
 1: Get the rank K SVD of $\boldsymbol{X}$: $\hat{\mathbf{U}}\hat{\mathbf{L}}\hat{\mathbf{V}}^{\mathrm{T}}$.
 2: Run SPA (respectively, preconditioned SPA) with input $(\hat{\mathbf{U}}, K)$, which outputs a set of indices J with cardinality K.
 3: Set $\hat{\mathbf{H}} := \hat{\mathbf{U}}_J$.
 4: Set $\hat{\boldsymbol{W}} := \hat{\mathbf{U}}\hat{\mathbf{H}}^{-1}$.

Based on the SPOC estimator $\hat{\boldsymbol{W}}$ of matrix $\boldsymbol{W}$, it is possible to construct an estimator for matrix $\boldsymbol{A}$ in a straightforward way. Indeed, given the decompositions (4) and (5), we can use the definition $\boldsymbol{\Pi} = \boldsymbol{W}\boldsymbol{A}$ and deduce that $\boldsymbol{A} = \mathbf{H}\mathbf{L}\mathbf{V}^{\mathrm{T}}$. A direct sample-based estimator of $\boldsymbol{A}$ is then given by

$$\hat{\boldsymbol{A}} = \hat{\mathbf{H}}\hat{\mathbf{L}}\widehat{\mathbf{V}}^{\mathrm{T}}. \tag{9}$$

In order to illustrate the usefulness of this estimator, we have performed its experimental comparison with LDA, see Section A of the Appendix.

4 Main results

In this section, we provide bounds on the performance of SPOC algorithm. We first prove deterministic bounds assuming that $\boldsymbol{X}$ is some fixed matrix close enough to $\boldsymbol{\Pi}$ in the spectral norm. Next, we combine these results with a concentration inequality for $\|\boldsymbol{X} - \boldsymbol{\Pi}\|$ when $\boldsymbol{X}$ is distributed according to $\mathbb{P}_{\boldsymbol{\Pi}}$ in order to obtain a bound on the estimation error with high probability under our statistical model.

4.1 Deterministic bounds

A key step in analyzing the performance of SPOC algorithm is to show that $\hat{\mathbf{U}}$ is close to an orthogonal transformation $\mathbf{U}\mathbf{O}$ of the population matrix $\mathbf{U}$. The next lemma gives a bound on the maximal ℓ_2-distance between the rows of $\hat{\mathbf{U}}$ and $\mathbf{U}\mathbf{O}$ for some orthogonal matrix $\mathbf{O}$, which will allow us to deduce an upper bound on the error of SPA from the results of [18, 32] (see Appendix A.2 for the details). We state this lemma as a deterministic result where $\boldsymbol{X}$ is some fixed matrix close enough to $\boldsymbol{\Pi}$ in the spectral norm. Recall that $\lambda_1(\boldsymbol{W})$ is the maximum singular value of matrix $\boldsymbol{W}$, $\lambda_K(\boldsymbol{\Pi})$ is the Kth singular value of matrix $\boldsymbol{\Pi}$, and $\kappa(\boldsymbol{W})$, $\kappa(\boldsymbol{\Pi})$ are the condition numbers of matrices $\boldsymbol{W}$ and $\boldsymbol{\Pi}$, respectively. Assuming that $\lambda_K(\boldsymbol{\Pi}) > 0$

(that is, $\boldsymbol{\Pi}$ is a rank K matrix) we define the values

$$\beta_i(\boldsymbol{X}, \boldsymbol{\Pi}) = K^{1/2} \kappa^2(\boldsymbol{\Pi}) \, \frac{\|\mathbf{e}_i^{\mathrm{T}} \boldsymbol{X}\|_2 \|\boldsymbol{X} - \boldsymbol{\Pi}\|}{\lambda_K^2(\boldsymbol{\Pi})} + \frac{\|\mathbf{e}_i^{\mathrm{T}}(\boldsymbol{X} - \boldsymbol{\Pi})\|_2}{\lambda_K(\boldsymbol{\Pi})}, \quad i = 1, \dots, n,$$

where $(\mathbf{e}_1, \dots, \mathbf{e}_n)$ is the canonical basis of $\mathbb{R}^n$.

Lemma 1. *Assume that $\boldsymbol{\Pi} \in \mathbb{R}^{n \times p}$ is a rank K matrix, and $\boldsymbol{X} \in \mathbb{R}^{n \times p}$ is any matrix such that $\|\boldsymbol{X} - \boldsymbol{\Pi}\| \leq \frac{1}{2}\lambda_K(\boldsymbol{\Pi})$. Let $\hat{\mathbf{U}}, \mathbf{U}$ be the $n \times K$ matrices of left singular vectors corresponding to the top K singular values of $\boldsymbol{X}$ and $\boldsymbol{\Pi}$, respectively. Then, there exist an orthogonal matrix $\mathbf{O}$ and a constant $C > 0$ such that, for any $i = 1, \dots, n$,*

$$\|\mathbf{e}_i^{\mathrm{T}}(\hat{\mathbf{U}} - \mathbf{U}\mathbf{O})\|_2 \leq C\beta_i(\boldsymbol{X}, \boldsymbol{\Pi}).$$

Define now

$$\beta(\boldsymbol{X}, \boldsymbol{\Pi}) = \max_{i=1,\dots,n} \beta_i(\boldsymbol{X}, \boldsymbol{\Pi}).$$

We will need the following condition.

Assumption 2. *For a constant $\bar{C} > 0$ we have*

$$\beta(\boldsymbol{X}, \boldsymbol{\Pi}) \leq \frac{\bar{C}}{\lambda_1(\boldsymbol{W})\kappa(\boldsymbol{W})K\sqrt{K}}.$$

Assumption 2 is satisfied with high probability for $\boldsymbol{X} \sim \mathbb{P}_{\boldsymbol{\Pi}}$ when the sample size N is large enough (see Appendix B.3). Under Assumption 2, we can derive from Lemma 1 the following deterministic bound on the error of estimating the topic-document matrix by the SPOC algorithm.

Lemma 2. *Let Assumptions 1 and 2 be satisfied with constant $\bar{C}$ small enough. Assume that $\boldsymbol{\Pi} \in \mathbb{R}^{n \times p}$ is a rank K matrix, and $\boldsymbol{X} \in \mathbb{R}^{n \times p}$ is any matrix such that $\|\boldsymbol{X} - \boldsymbol{\Pi}\| \leq \frac{1}{2}\lambda_K(\boldsymbol{\Pi})$. Then, matrix $\hat{\mathbf{H}}$ is non-degenerate and the preconditioned SPOC algorithm outputs matrix $\hat{\boldsymbol{W}}$ such that*

$$\min_{\mathbf{P} \in \mathscr{P}} \|\hat{\boldsymbol{W}} - \boldsymbol{W}\mathbf{P}\|_F \leq CK^{1/2} \left\{ \lambda_{\max}^2(\boldsymbol{W})\kappa(\boldsymbol{W})\beta(\boldsymbol{X}, \boldsymbol{\Pi}) + \frac{\kappa(\boldsymbol{\Pi})\lambda_1(\boldsymbol{W})\|\boldsymbol{X} - \boldsymbol{\Pi}\|}{\lambda_K(\boldsymbol{\Pi})} \right\},$$

where $\mathscr{P}$ denotes the set of all permutation matrices.

Inspection of the proof shows that, for this lemma to hold, it is enough to choose the constant $\bar{C} \leq \min(C_*, C_0^{-1})$ where C_*, C_0 are the constants from Theorem 4 and Corollary 5.

4.2 Bounds with high probability

Lemma 2 combined with a concentration inequality for $\|\boldsymbol{X} - \boldsymbol{\Pi}\|$ (cf. Lemma 4 in the Appendix) allows us to derive a bound for the estimation error that holds with high probability when $\boldsymbol{X}$ is sampled from distribution $\mathbb{P}_{\boldsymbol{\Pi}}$. Introduce the value

$$\Delta(\boldsymbol{W}, \boldsymbol{\Pi}) = \left(\frac{\lambda_1(\boldsymbol{W})}{\lambda_K(\boldsymbol{\Pi})} \right)^2 \kappa(\boldsymbol{W})\kappa^2(\boldsymbol{\Pi}).$$

The main result is summarized in the next theorem.

7

Theorem 1. *Let Assumption 1 hold, and $N_i = N$ for $i = 1, \ldots, n$. Assume that $N \geq \log(n+p)$ and*

$$\lambda_K(\mathbf{\Pi}) \geq \sqrt{\frac{10}{\overline{C}}} \, K \left(\frac{n \log(n+p)}{N} \right)^{1/4} \kappa(\mathbf{\Pi}) \sqrt{\lambda_1(\boldsymbol{W}) \kappa(\boldsymbol{W})}. \tag{10}$$

Then, with probability at least $1 - 2(n+p)^{-1}$, matrix $\hat{\mathbf{H}}$ is non-degenerate and the output $\hat{\boldsymbol{W}}$ of preconditioned SPOC algorithm satisfies, for some constant $C_1 > 0$,

$$\min_{\mathbf{P} \in \mathscr{P}} \left\| \hat{\boldsymbol{W}} - \boldsymbol{W}\mathbf{P} \right\|_F \leq C_1 K \sqrt{\frac{n \log(n+p)}{N}} \Delta(\boldsymbol{W}, \mathbf{\Pi}),$$

where $\mathscr{P}$ denotes the set of all permutation matrices.

Condition (10) in Theorem 1 guarantees that Assumption 2 is satisfied. This condition holds if N is large enough and it quantifies the separation of the spectrum of matrix $\mathbf{\Pi}$ from zero.

The bound of Theorem 1 depends on the singular values of matrices $\boldsymbol{W}$ and $\mathbf{\Pi}$. We now further detail this bound for the balanced case where matrices $\boldsymbol{W}$ and $\mathbf{\Pi}$ are well conditioned and the smallest non-zero singular value of $\mathbf{\Pi}$ is of the same order as the largest singular value of $\boldsymbol{W}$. It follows from Lemma 7 in the Appendix that in this case both $\lambda_K(\mathbf{\Pi})$ and $\lambda_1(\boldsymbol{W})$ are of the order of $\sqrt{n/K}$. This is coherent with the behavior of the singular values of $\mathbf{\Pi}$ and $\boldsymbol{W}$ that we observed in the simulation study (see Section B). The balanced case is formally described by the following assumption.

Assumption 3. *There exist two constants $C > 1$ and $c > 0$ such that*

$$\lambda_K(\mathbf{\Pi}) \geq C\lambda_1(\boldsymbol{W}) \quad and \quad \max\{\kappa(\mathbf{\Pi}), \kappa(\boldsymbol{W})\} \leq c.$$

The second condition in Assumption 3 is quite standard and just states that matrices $\mathbf{\Pi}$ and $\boldsymbol{W}$ are well-conditioned. The first condition is more restrictive. It holds, in particular, if matrix $\boldsymbol{A}$ is well-conditioned with large enough singular value $\lambda_K(\boldsymbol{A})$. For example, it will be the case if $\boldsymbol{A}$ satisfies the *anchor word assumption* (see Section 1) with the probabilities of anchor words uniformly above the probabilities of other words. This is detailed in Lemma 9 of the Appendix. Noteworthy, the lower bound of Theorem 3 below is attained with such choice of matrix $\boldsymbol{A}$, see the proof of Theorem 3 in Appendix B.5. We can interpret it as the fact that, in a minimax sense, such matrices $\boldsymbol{A}$ are associated with the least favorable models.

The following corollary quantifies the behavior of SPOC estimator in the balanced case.

Corollary 1 (Upper bound in the balanced case). *Let Assumptions 1 and 3 hold, and $N_i = N$ for $i = 1, \ldots, n$. Let also*

$$N \geq CK^5 \log(n+p) \tag{11}$$

for some $C > 0$ large enough. Then, with probability at least $1 - 2(n+p)^{-1}$, matrix $\hat{\mathbf{H}}$ is non-degenerate and the output $\hat{\boldsymbol{W}}$ of preconditioned SPOC algorithm satisfies, for some constant $C_2 > 0$,

$$\min_{\mathbf{P} \in \mathscr{P}} \left\| \hat{\boldsymbol{W}} - \boldsymbol{W}\mathbf{P} \right\|_F \leq C_2 K \sqrt{\frac{n \log(n+p)}{N}},$$

where $\mathscr{P}$ denotes the set of all permutation matrices.

To prove Corollary 1, it is enough to notice that under Assumption 3 we have $\Delta(\boldsymbol{W}, \mathbf{\Pi}) \leq C'$ for some constant $C' > 0$, and condition (10) follows from (11), Assumption 3 and the inequality $\lambda_1(\boldsymbol{W}) \geq \sqrt{n/K}$ (see Lemma 7 in the Appendix).

Note that from Theorem 1 and Corollary 1 we can derive bounds in other norms. Thus, using the inequalities $\left\| \hat{\boldsymbol{W}} - \boldsymbol{W}\mathbf{P} \right\|_1 \leq \sqrt{Kn} \left\| \hat{\boldsymbol{W}} - \boldsymbol{W}\mathbf{P} \right\|_F$ and $\left\| \hat{\boldsymbol{W}} - \boldsymbol{W}\mathbf{P} \right\|_{1,\infty} \leq \sqrt{K} \left\| \hat{\boldsymbol{W}} - \boldsymbol{W}\mathbf{P} \right\|_F$, we obtain the following corollary:

Corollary 2. *If the assumptions of Theorem 1 are satisfied then, with probability at least $1 - 2(n+p)^{-1}$, matrix $\hat{\mathbf{H}}$ is non-degenerate and the output $\hat{\boldsymbol{W}}$ of preconditioned SPOC algorithm satisfies*

$$\min_{\mathbf{P} \in \mathscr{P}} \left\| \hat{\boldsymbol{W}} - \boldsymbol{W}\mathbf{P} \right\|_{1,\infty} \leq \mathrm{C}_1 K^{3/2} \sqrt{\frac{n \log(n+p)}{N}} \Delta(\boldsymbol{W}, \boldsymbol{\Pi}) \quad and$$

$$\min_{\mathbf{P} \in \mathscr{P}} \left\| \hat{\boldsymbol{W}} - \boldsymbol{W}\mathbf{P} \right\|_{1} \leq \mathrm{C}_1 K^{3/2} n \sqrt{\frac{\log(n+p)}{N}} \Delta(\boldsymbol{W}, \boldsymbol{\Pi}).$$

If the assumptions of Corollary 1 are satisfied then, with probability at least $1 - 2(n+p)^{-1}$, matrix $\hat{\mathbf{H}}$ is non-degenerate and the output $\hat{\boldsymbol{W}}$ of preconditioned SPOC algorithm satisfies

$$\min_{\mathbf{P} \in \mathscr{P}} \left\| \hat{\boldsymbol{W}} - \boldsymbol{W}\mathbf{P} \right\|_{1,\infty} \leq \mathrm{C}_2 K^{3/2} \sqrt{\frac{n \log(n+p)}{N}} \quad and \quad \min_{\mathbf{P} \in \mathscr{P}} \left\| \hat{\boldsymbol{W}} - \boldsymbol{W}\mathbf{P} \right\|_{1} \leq \mathrm{C}_2 K^{3/2} n \sqrt{\frac{\log(n+p)}{N}}$$

It follows from Corollaries 1 and 2 that, for all the considered norms, the rate of estimating $\boldsymbol{W}$ (to within a weak factor) is determined by two parameters, which are the number of documents n and the sample size N. The dependency on the size of the dictionary p is weak. This is confirmed by the numerical experiments, see Section 6.

4.3 Adaptive procedure when K is unknown.

We now propose an adaptive variant of the SPOC algorithm when the number of topics K is unknown. It is obtained by replacing K in Algorithm 2 by the estimator

$$\hat{K} = \max \left\{ j : \ \lambda_j(\boldsymbol{X}) > 4\sqrt{\frac{n \log(n+p)}{N}} \right\}.$$

In the sequel, the resulting procedure will be called the adaptive (preconditioned) SPOC algorithm. The following analogs of Theorem 1 and Corollary 1 hold.

Theorem 2. *Let the assumptions of Theorem 1 be satisfied and*

$$\lambda_1(\boldsymbol{W}) > \frac{32\bar{C}}{5K^2} \sqrt{\frac{n \log(n+p)}{N}}. \tag{12}$$

Then, with probability at least $1 - 2(n+p)^{-1}$, matrix $\hat{\mathbf{H}}$ is non-degenerate, $\hat{K} = K$, and the output $\hat{\boldsymbol{W}}$ of the adaptive preconditioned SPOC algorithm satisfies

$$\min_{\mathbf{P} \in \mathscr{P}} \left\| \hat{\boldsymbol{W}} - \boldsymbol{W}\mathbf{P} \right\|_{F} \leq \mathrm{C}_1 K \sqrt{\frac{n \log(n+p)}{N}} \Delta(\boldsymbol{W}, \boldsymbol{\Pi}).$$

Corollary 3. *Let the assumptions of Corollary 1 and (12) be satisfied. Then, with probability at least $1 - 2(n+p)^{-1}$, matrix $\hat{\mathbf{H}}$ is non-degenerate, $\hat{K} = K$, and the output $\hat{\boldsymbol{W}}$ of the adaptive preconditioned SPOC algorithm satisfies*

$$\min_{\mathbf{P} \in \mathscr{P}} \left\| \hat{\boldsymbol{W}} - \boldsymbol{W}\mathbf{P} \right\|_{F} \leq \mathrm{C}_2 K \sqrt{\frac{n \log(n+p)}{N}}.$$

Note that condition (12) introduced in Theorem 2 and Corollary 3 additionally to the conditions of Theorem 1 and Corollary 1 is rather mild. Indeed, due to inequality (42) proved in the Appendix we have $\lambda_1(\boldsymbol{W}) \geq \sqrt{n/K}$. Therefore, it is sufficient that $N > \frac{C \log(n+p)}{K^3}$ to grant (12).

4.4 Minimax lower bound

The following lower bound shows that the rate obtained in Corollary 1 is near minimax optimal. Denote by $\mathscr{M}$ the class of all matrices $\mathbf{\Pi}$ satisfying the assumptions stated in the Section 1 and Assumption 3.

Theorem 3 (Lower bound)**.** *Assume that $N_i = N$ for $i = 1, \ldots, n$ and $2 \leq K \leq \min(p/4, N/2, n/2)$. Then, there exist two constants $C > 0$ and $c \in (0, 1)$ such that, for any estimator and $\overline{\mathbf{W}}$ of $\mathbf{W}$ we have*

$$\sup_{\mathbf{\Pi} \in \mathscr{M}} \mathbb{P}_{\mathbf{\Pi}} \left\{ \min_{\mathbf{P} \in \mathscr{P}} \|\overline{\mathbf{W}} - \mathbf{W}\mathbf{P}\|_F \geq C\sqrt{\frac{n}{N}} \right\} \geq c, \tag{13}$$

and

$$\sup_{\mathbf{\Pi} \in \mathscr{M}} \mathbb{P}_{\mathbf{\Pi}} \left\{ \min_{\mathbf{P} \in \mathscr{P}} \|\overline{\mathbf{W}} - \mathbf{W}\mathbf{P}\|_1 \geq Cn\sqrt{\frac{K}{N}} \right\} \geq c, \tag{14}$$

where $\mathscr{P}$ denotes the set of all permutation matrices.

Combining Corollary 1 and (13) we find that, to within a weak factor, the minimax optimal rate of estimation of $\mathbf{W}$ on the class $\mathscr{M}$ in the Frobenius norm scales as $\sqrt{n/N}$. On the other hand, (14) and Corollary 2 imply that, again to within a weak factor, the minimax optimal rate of estimation of $\mathbf{W}$ in the l_1-norm on $\mathscr{M}$ scales as $n/\sqrt{N}$.

Remark 1. *Inspection of the proof of Theorem 3 shows that the lower bound is in fact established for a subset of $\mathscr{M}$ composed of matrices satisfying both anchor word and anchor document assumptions.*

Remark 2. *It is proved in [8, 22] that, under the same model of observations and anchor word assumption, the minimax optimal rate for estimation of matrix $\mathbf{A}$ in the ℓ_1-norm scales as $\sqrt{\frac{p}{nN}}$ (to within weak factors). Note that this rate is determined by all the three main parameters of the problem - the size of the dictionary p, the number of documents n and the sample size N. This is quite different from the minimax ℓ_1-rate $n/\sqrt{N}$ of estimation $\mathbf{W}$, which remains valid under anchor word assumption, cf. Remark 1 and the remark after Assumption 3. It shows that there is a significant difference between the problems of estimating matrices $\mathbf{A}$ and $\mathbf{W}$ in topic models.*

5 Related Work

There exists an extensive literature on topic modeling and several algorithms have been proposed for estimation of matrices $\mathbf{A}$ and $\mathbf{W}$. As the problem of recovering these two matrices when there is no noise is an instance of non-negative matrix factorisation several papers propose algorithms based on minimization of a regularized cost function, see, e.g., [13, 16, 24, 40]. Such methods result in non-convex optimization and often fail when many words do not appear in a single document, that is when $N \ll p$.

Another approach is to use Bayesian methods such as the popular *Latent Dirichlet Allocation* (LDA) introduced in [11]. LDA proceeds by imposing a Dirichlet prior on $\mathbf{A}$ and then computing an estimator of $\mathbf{W}$ by a variational EM-algorithm. The original paper [11] and the subsequent line of work do not provide statistical guarantee on the recovery of $\mathbf{W}$. In [11], the authors argue that LDA avoids two issues of the pLSI that are the risk of overfitting and the difficulty of classifying a new document outside the corpus. Yet, LDA is computationally slow and makes the assumption that topics are uncorrelated, which may be not realistic [10, 27]. This last issue has been addressed in [23] by introducing *Correlated topic models*. LDA has been extended to relax some assumptions such as the *bag-of-words* hypothesis ("order of words does not matter") [45], the exchangeability of documents ("topics do not vary in time") [9], the assumption that the number of topics is known [42]. Also, to recover $\mathbf{W}$ in the LDA setting, some papers used Gibbs-sampling [37, 39] or variational Bayes techniques [12, 47] rather than the EM-algorithm. However, these

works do not provide statistical guarantees on the estimation of W and the associated algorithms are computationally slow.

For the problem of estimation of matrix A, papers [1, 2, 4–6, 8, 15, 22], to mention but a few, provided algorithms with provable statistical guarantees under *anchor word assumption.* They proposed various techniques based, for example, on analyzing co-occurrence matrices, tensors, or on recovering vertices of a simplex using SVD. Most of these papers, except for [8, 22], do not work under the same statistical model as ours (cf. Section 1). Thus, [4, 5, 15] assume that topic-document matrix W is randomly generated from some prior distribution. For a setting with no randomness, [31] proposes ellipsoidal rounding algorithm with application to topic models. Paper [28] develops a generalized method to bind overlapping clustering models, including topic models. Moreover, for some classes of matrices, papers [8, 22] proposed minimax optimal algorithms of estimating A. Both [8, 22] impose *anchor word assumption* but their estimators are different. Thus, [22] performs SVD on properly normalized matrix X followed by an exhaustive search over a p-dimensional simplex, while [8] proceeds by first recovering the anchor words and then deriving estimators of A from a scaled version of matrix XX^{T}.

6 Simulations

6.1 Synthetic Data

We first present the results of experiments on synthetic data. We have performed simulations with different values of the parameters n, p, N and the number of topics K. Our objective was to observe the effect of each parameter on the Frobenius error between W and its estimator $\hat{W}$ obtained by the SPOC algorithm. We report the results for the SPOC algorithm without preconditioning step as it had a negligible impact on the performance of the method while being computationally demanding. As a benchmark, we use the LDA algorithm [11]. For the experiments we use the Python implementation of SPOC[2] and an implementation of the LDA algorithm available in Sklearn [35].

Figures 1-4 present an example of results that we have typically obtained in simulations. In Figures 1-3 we take $K = 3$ and the matrix W has the following structure: K rows of W are canonical basis vectors, each of the remaining $n-K$ rows is generated independently using the Dirichlet distribution with parameter $\alpha = (0.1, 0.15, 0.2)$. In Figure 4, where K must vary, we define W in a different way. Namely, for the $n-K$ rows that are not canonical basis vectors, each element W_{kj} is generated from the uniform distribution on $[0, 1]$ and then each row of the matrix is normalized so as to have $\sum_{k=1}^{K} W_{ik} = 1$. For the matrix A, we take K columns proportional to canonical basis vectors. The elements A_{kj} of matrix A in the remaining $p - K$ columns are obtained by generating numbers from the uniform distribution on $[0, 1]$ and then normalizing each row of the matrix to have $\sum_{j=1}^{p} A_{kj} = 1$. For given W and A, the data matrix X is generated according to the pLSI model defined in Section 1. For each value on the x-axes of the figures, we present the averaged result over 10 simulations.

We clearly retrieve the patterns indicated in Theorem 1, Corollary 1 and (13). Thus, the plots have a near $\sqrt{n}$ behaviour in Figure 1 and a near $1/\sqrt{N}$ behaviour in Figure 2. Figure 3 shows weak dependence of the error of the SPOC algorithm on the size of the dictionary p, which agrees with the bound obtained in Corollary 1. This can be interpreted as one of the advantages of our method over LDA. Indeed, for LDA we observe that the error increases significantly as p grows. Finally, we notice the linear dependence of the error on K as predicted by Corollary 1. In all the experiments we observe that SPOC algorithm is very competitive with LDA while being much more stable.

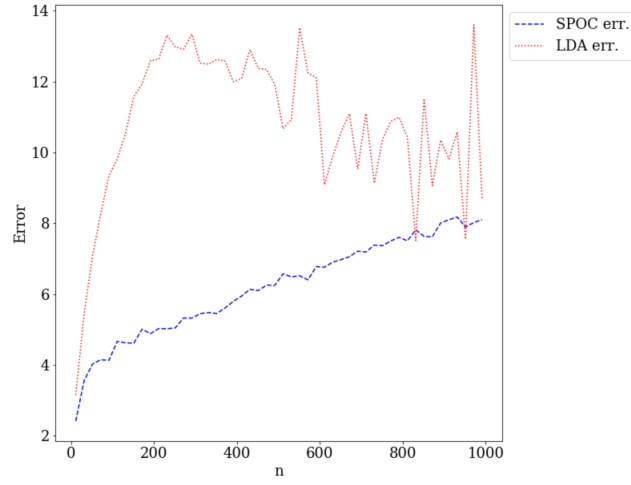[2]The code of SPOC algorithm is available at https://github.com/stat-ml/SPOC

Figure 1: The n-dependency of $\min_{\mathbf{P} \in \mathscr{P}} \|\boldsymbol{W} - \hat{\boldsymbol{W}}\mathbf{P}\|_F$ using SPOC and LDA algorithms. Total number of words $p = 5000$, number of sampled words in each document $N = 200$.
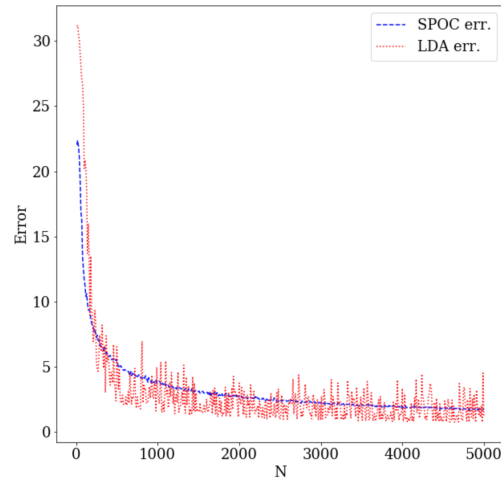
Figure 2: The N-dependency of $\min_{\mathbf{P} \in \mathscr{P}} \|\boldsymbol{W} - \hat{\boldsymbol{W}}\mathbf{P}\|_F$ using SPOC and LDA algorithms. Total number of words $p = 5000$, number of documents $n = 1000$.

Figure 3: The p-dependency of $\min_{\mathbf{P} \in \mathscr{P}} \|\boldsymbol{W} - \hat{\boldsymbol{W}}\mathbf{P}\|_F$ using SPOC and LDA algorithms. Number of documents $n = 1000$, number of sampled words in each document $N = 200$.

Figure 4: The K-dependency of $\min_{\mathbf{P} \in \mathscr{P}} \|\boldsymbol{W} - \hat{\boldsymbol{W}}\mathbf{P}\|_F$ using SPOC and LDA algorithms. Total number of words $p = 5000$, number of sampled words in each document $N = 5000$.
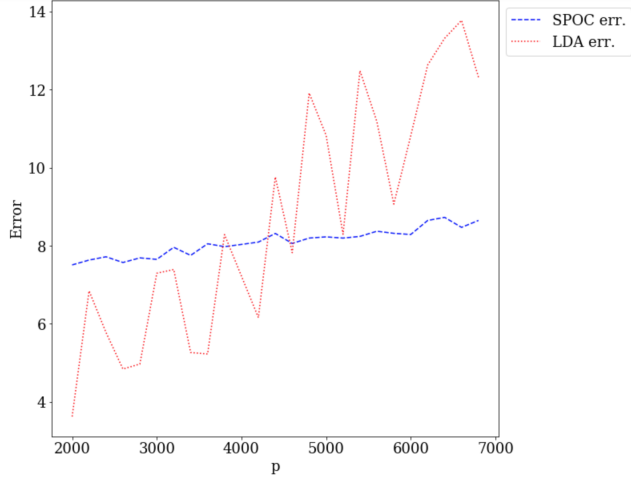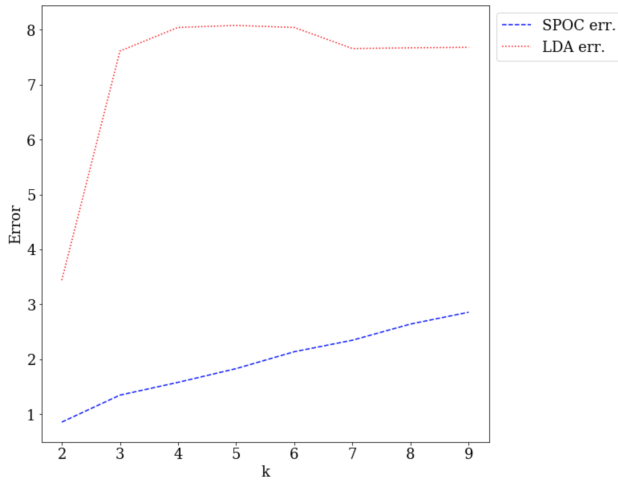
A numerical study of the SPOC estimator $\hat{\boldsymbol{A}}$ of matrix $\boldsymbol{A}$ is deferred to Section A of the Appendix. It shows that $\hat{\boldsymbol{A}}$ behaves similarly to the corresponding LDA estimator while being more stable.

6.2   Corpus of NIPS abstracts

We now illustrate the performance of our algorithm applying it to the data set of full texts of NIPS papers[3] [36]. This data set contains the distribution of words in the full text of the NIPS conference papers

---

[3]The link to the dataset: https://archive.ics.uci.edu/ml/datasets/NIPS+Conference+Papers+1987-2015

published from 1987 to 2015. The data set has the form of a 11463×5811 matrix of word counts containing 11463 words and 5811 NIPS conference papers. Each column contains the number of times each word appears in the corresponding document.

We start by pre-processing the data. We first remove all the documents with less than 150 words. Then we remove from the resulting dictionary the stop words and the words that appear in less that 150 documents. This results in a database of 5801 documents with a dictionary of 6380 words. Note that we do not actually have access to the true topics of each document. We also do not have access to the true number of topics. In order to compare our method to LDA, we proceeded as follows. For each value of $K = 3, \ldots, 10$, we first computed the LDA estimator $\tilde{W}$ of the document-topic matrix and the LDA estimator $\tilde{A}$ of the topic-word matrix. Next, with the underlying matrix $\tilde{\Pi} = \tilde{W}\tilde{A}$, for each value of K we simulated 10 matrices $\tilde{X}$ with $N = 200$ sampled words according to pLSI model. For each matrix $\tilde{X}$, we estimated $\tilde{W}$ using both LDA and SPOC algorithms. Finally, for each K we computed the mean error over 10 simulations. The resulting comparison as function of K is presented in Figure 5. We can observe that SPOC systematically outperforms the LDA algorithm, except for $K = 2$.
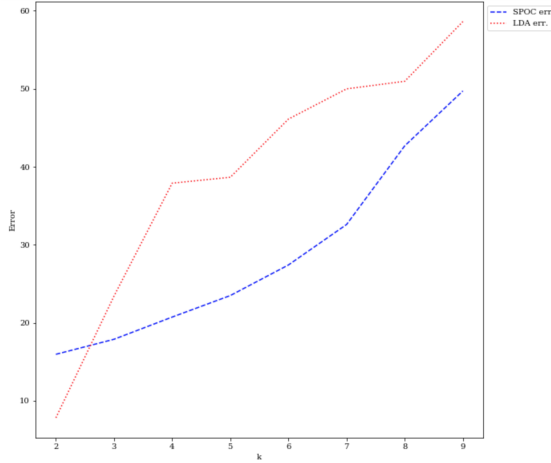

Figure 5: The K-dependency of $\min_{\mathbf{P} \in \mathscr{P}} \|\mathbf{W} - \hat{\mathbf{W}}\mathbf{P}\|_F$ using SPOC and LDA algorithms on semi-synthetic data. Matrix $\tilde{W}$ is the LDA estimator on the NIPS data set ($n = 5081$ documents, $p = 6380$ words), and $\hat{W}$ is the LDA or SPOC estimator on data simulated from $\tilde{\Pi}$.

Next, we investigate whether our estimator of matrix $\mathbf{W}$ helps to well classify the documents in the NIPS corpus. We apply the simplest possible classifier based on the obtained SPOC estimator $\hat{\mathbf{W}}$, namely, we classify article i to the topic that has the maximum value of $\hat{W}_{ik}$ for $k = 1, \ldots, K$. We consider the number of topics $K = 3$. While apparently for NIPS papers some words (such as "learning" or "model") are more frequent than others in the full corpus, other words can be more frequent for particular topics. Therefore, for each topic we present the words that have the highest difference between their frequency for this topic and their maximum frequency in other topics. We clearly see that the obtained topics are semantically well separated, cf. Table 2.

14

Table 2: Top 10 words, which have the highest difference in frequency for each topic compared to other topics. The three topics were identified by SPOC method.

	"Neural networks"	"Statistical learning"	"Algorithms and theory"
1	network	model	algorithm
2	input	data	learning
3	neural	image	function
4	neurons	distribution	problem
5	units	inference	set
6	output	likelihood	theorem
7	layer	latent	bound
8	neuron	prior	matrix
9	system	Gaussian	loss
10	synaptic	parameters	error

7   Conclusion

In the present paper, we proposed the SPOC algorithm, which is a computationally efficient procedure to estimate the document-topic matrix in topic models with known or unknown number of topics K. It is based on the Successive Projection Algorithm used to recover the vertices of a K-dimensional simplex in the context of separable matrix factorization. We developed the statistical analysis of SPOC algorithm under the *anchor document assumption* requiring that, for each topic, there is a document devoted solely to this topic. We proved that the proposed method is near minimax optimal for estimation of the document-topic matrix under the Frobenius norm and the ℓ_1-norm. As an element of our analysis, we derived a bound on concentration of matrices with independent multinomial columns that may be of independent interest. The theoretical results are supported by empirical evidence demonstrating a good performance of the SPOC algorithm and its advantages compared to LDA.

A   Tools

A.1   Matrix Perturbation Bounds

In this section, we provide some facts about matrix perturbation that will be used in the proofs. We start with the following lemma, which is a variant of Davis-Kahan theorem.

Proposition 1 (Lemma 5.1 in [25])**.** *Let* $\mathbf{M} \in \mathbb{R}^{n \times n}$ *be a rank K symmetric matrix with smallest nonzero eigenvalue* $\lambda_K(\mathbf{M})$*, and let* $\hat{\mathbf{M}} \in \mathbb{R}^{n \times n}$ *be any symmetric matrix. Let* $\hat{\mathbf{U}}(\hat{\mathbf{M}}) \in \mathbb{R}^{n \times K}$ *and* $\mathbf{U}(\mathbf{M}) \in \mathbb{R}^{n \times K}$ *be the matrices of K leading eigenvectors of* $\hat{\mathbf{M}}$ *and* $\mathbf{M}$*, respectively. Then there exists a $K \times K$ orthogonal matrix* $\mathbf{O}$ *such that*

$$\|\hat{\mathbf{U}}(\hat{\mathbf{M}}) - \mathbf{U}(\mathbf{M})\mathbf{O}\|_F \leq \frac{2\sqrt{2K}\|\hat{\mathbf{M}} - \mathbf{M}\|}{\lambda_K(\mathbf{M})}.$$

15

Corollary 4. *Let $\mathbf{\Pi}$ and $\boldsymbol{X}$ be matrices with singular value decompositions given by* (4) *and* (6). *Then there exist $K \times K$ orthogonal matrices $\mathbf{O}$ and $\widetilde{\mathbf{O}}$ such that*

$$\|\hat{\mathbf{U}} - \mathbf{U}\mathbf{O}\|_F \leq \frac{2\sqrt{2K}(\|\boldsymbol{X}\| + \|\mathbf{\Pi}\|)\|\boldsymbol{X} - \mathbf{\Pi}\|}{\lambda_K^2(\mathbf{\Pi})} \tag{15}$$

and

$$\|\hat{\mathbf{V}} - \mathbf{V}\widetilde{\mathbf{O}}\|_F \leq \frac{2\sqrt{2K}(\|\boldsymbol{X}\| + \|\mathbf{\Pi}\|)\|\boldsymbol{X} - \mathbf{\Pi}\|}{\lambda_K^2(\mathbf{\Pi})}. \tag{16}$$

Furthermore, if $\|\boldsymbol{X} - \mathbf{\Pi}\| \leq \frac{1}{2}\lambda_K(\mathbf{\Pi})$ then

$$\max\left(\|\hat{\mathbf{U}} - \mathbf{U}\mathbf{O}\|_F, \|\hat{\mathbf{V}} - \mathbf{V}\widetilde{\mathbf{O}}\|_F\right) \leq \frac{5\sqrt{2K}\kappa(\mathbf{\Pi})\|\boldsymbol{X} - \mathbf{\Pi}\|}{\lambda_K(\mathbf{\Pi})}. \tag{17}$$

Proof. Applying Proposition 1 to matrices $\mathbf{\Pi}\mathbf{\Pi}^{\mathrm{T}}$ and $\boldsymbol{X}\boldsymbol{X}^{\mathrm{T}}$ we get

$$\|\hat{\mathbf{U}} - \mathbf{U}\mathbf{O}\|_F \leq \frac{2\sqrt{2K}\|\mathbf{\Pi}\mathbf{\Pi}^{\mathrm{T}} - \boldsymbol{X}\boldsymbol{X}^{\mathrm{T}}\|}{\lambda_K(\mathbf{\Pi}\mathbf{\Pi}^{\mathrm{T}})} \leq \frac{2\sqrt{2K}(\|\boldsymbol{X}\| + \|\mathbf{\Pi}\|)\|\boldsymbol{X} - \mathbf{\Pi}\|}{\lambda_K^2(\mathbf{\Pi})}.$$

Similarly, inequality (16) is obtained by applying Proposition 1 to matrices $\mathbf{\Pi}^{\mathrm{T}}\mathbf{\Pi}$ and $\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}$. Next, if $\|\boldsymbol{X} - \mathbf{\Pi}\| \leq \frac{1}{2}\lambda_K(\mathbf{\Pi})$ then due to the triangle inequality we have $\|\boldsymbol{X}\| \leq \|\mathbf{\Pi}\| + \frac{1}{2}\lambda_K(\mathbf{\Pi}) \leq \frac{3}{2}\|\mathbf{\Pi}\|$. Combining this fact with (15) and (16) we obtain (17). $\qquad\square$

We will also need the following bounds for matrices of singular values $\hat{\mathbf{L}}$ and $\mathbf{L}$.

Lemma 3. *Let the assumptions of Corollary 4 hold. Let $\hat{\mathbf{L}}$ and $\mathbf{L}$ be diagonal $K \times K$-matrices of K largest singular values of $\boldsymbol{X}$ and $\mathbf{\Pi}$, respectively, cf.* (4) *and* (6). *If $\|\boldsymbol{X} - \mathbf{\Pi}\| \leq \frac{1}{2}\lambda_K(\mathbf{\Pi})$ then*

$$\|\hat{\mathbf{L}} - \mathbf{O}^{\mathrm{T}}\mathbf{L}\widetilde{\mathbf{O}}\| \leq C\kappa^2(\mathbf{\Pi})\sqrt{K}\,\|\boldsymbol{X} - \mathbf{\Pi}\|$$

and

$$\|\hat{\mathbf{L}}^{-1} - \widetilde{\mathbf{O}}^{\mathrm{T}}\mathbf{L}^{-1}\mathbf{O}\| \leq C\kappa^2(\mathbf{\Pi})\sqrt{K}\frac{\|\boldsymbol{X} - \mathbf{\Pi}\|}{\lambda_K^2(\mathbf{\Pi})},$$

where the orthogonal matrices $\mathbf{O}$, $\widetilde{\mathbf{O}}$ are the same as in Corollary 4.

Proof. Applying Weyl's inequality [19, Theorem C.6] we get

$$\|\hat{\mathbf{U}}\hat{\mathbf{L}}\hat{\mathbf{V}}^{\mathrm{T}} - \mathbf{U}\mathbf{L}\mathbf{V}^{\mathrm{T}}\| \leq 2\|\mathbf{\Pi} - \boldsymbol{X}\|$$

and further

$$\|\hat{\mathbf{U}}\hat{\mathbf{L}}\hat{\mathbf{V}}^{\mathrm{T}} - \mathbf{U}\mathbf{L}\mathbf{V}^{\mathrm{T}}\| \geq \|\mathbf{U}\mathbf{O}(\hat{\mathbf{L}} - \mathbf{O}^{\mathrm{T}}\mathbf{L}\widetilde{\mathbf{O}})\hat{\mathbf{V}}^{\mathrm{T}}\|$$
$$-\|(\hat{\mathbf{U}} - \mathbf{U}\mathbf{O})\hat{\mathbf{L}}\hat{\mathbf{V}}^{\mathrm{T}}\| - \|\mathbf{U}\mathbf{L}\widetilde{\mathbf{O}}(\hat{\mathbf{V}} - \mathbf{V}\widetilde{\mathbf{O}})^{\mathrm{T}}\|.$$

Therefore

$$\|\hat{\mathbf{L}} - \mathbf{O}^{\mathrm{T}}\mathbf{L}\widetilde{\mathbf{O}}\| \le \|\mathbf{X} - \mathbf{\Pi}\| + \|(\hat{\mathbf{U}} - \mathbf{UO})\hat{\mathbf{L}}\hat{\mathbf{V}}^{\mathrm{T}}\| + \|\mathbf{UL}\widetilde{\mathbf{O}}(\hat{\mathbf{V}} - \mathbf{V}\widetilde{\mathbf{O}})^{\mathrm{T}}\|$$

$$\le \|\mathbf{X} - \mathbf{\Pi}\| + \|\mathbf{X}\|\|\hat{\mathbf{U}} - \mathbf{UO}\| + \|\mathbf{\Pi}\|\|\hat{\mathbf{V}} - \mathbf{V}\widetilde{\mathbf{O}}\|$$

$$\le \left(\frac{2\sqrt{2K}(\|\mathbf{X}\| + \|\mathbf{\Pi}\|)^2}{\lambda_K^2(\mathbf{\Pi})} + 1 \right) \|\mathbf{X} - \mathbf{\Pi}\|,$$

where the last inequality is due to Corollary 4. Next,

$$\|\hat{\mathbf{L}}^{-1} - \widetilde{\mathbf{O}}^{\mathrm{T}}\mathbf{L}^{-1}\mathbf{O}\| = \|\hat{\mathbf{L}}^{-1}(\mathbf{O}^{\mathrm{T}}\mathbf{L}\widetilde{\mathbf{O}} - \hat{\mathbf{L}})\widetilde{\mathbf{O}}^{\mathrm{T}}\mathbf{L}^{-1}\mathbf{O}\|$$

$$\le \|\hat{\mathbf{L}}^{-1}\| \, \|\hat{\mathbf{L}} - \mathbf{O}^{\mathrm{T}}\mathbf{L}\widetilde{\mathbf{O}}\| \, \|\mathbf{L}^{-1}\|$$

$$\le \left(\frac{2\sqrt{2K}(\|\mathbf{X}\| + \|\mathbf{\Pi}\|)^2}{\lambda_K^2(\mathbf{\Pi})} + 1 \right) \frac{\|\mathbf{X} - \mathbf{\Pi}\|}{\lambda_K(\mathbf{X})\lambda_K(\mathbf{\Pi})},$$

where $\lambda_K(\mathbf{X})$ is the K-th largest singular value of matrix $\mathbf{X}$. Due to Weyl's inequality and the fact that $\|\mathbf{X} - \mathbf{\Pi}\| \le \frac{1}{2}\lambda_K(\mathbf{\Pi})$ we have $\lambda_K(\mathbf{X}) \ge \frac{1}{2}\lambda_K(\mathbf{\Pi})$ and $\|\mathbf{X}\| \le \|\mathbf{\Pi}\| + \frac{1}{2}\lambda_K(\mathbf{\Pi}) \le \frac{3}{2}\|\mathbf{\Pi}\|$. Plugging these inequalities in the last two displays we obtain the lemma. $\square$

A.2 Noisy Separable Matrix Factorization

In this section, we give a bound on the error of preconditioned SPA in Noisy Separable Matrix Factorization model. Assume that we observe

$$\tilde{\mathbf{G}} = \mathbf{G} + \mathbf{N} = \mathbf{W}\mathbf{Q} + \mathbf{N},$$

where $\mathbf{N} \in \mathbb{R}^{n \times K}$ is a perturbation (noise) matrix, and

$$\mathbf{G} = \mathbf{W}\mathbf{Q},$$

where $\mathbf{W} \in \mathbb{R}_+^{n \times K}$ and $\mathbf{Q} \in \mathbb{R}^{K \times K}$. If we assume that $\mathbf{W}$ satisfies Assumption 1 then we obtain the setting usually referred to as Noisy Separable Matrix Factorization (NSMF).

The following theorem holds for preconditioned SPA in the NSMF model, see [18, 32].

Theorem 4. *Let $K \ge 2$ and let Assumption 1 hold. Assume that matrix $\mathbf{Q}$ is non-degenerate and the entries W_{im} of matrix $\mathbf{W}$ satisfy the condition $\sum_{m=1}^{K} W_{im} \le 1$ for $i = 1, \dots, n$. Moreover, assume that for any $i = 1, \dots, n$, the norms of the rows of matrix $\mathbf{N}$ satisfy $\|\mathbf{e}_i^{\mathrm{T}}\mathbf{N}\|_2 \le \epsilon$ with*

$$\epsilon \le C_* \frac{\lambda_{\min}(\mathbf{Q})}{K\sqrt{K}}$$

for some constant $C_* > 0$ small enough. Let J be the set of indices returned by the preconditioned SPA with input $(\tilde{\mathbf{G}}, K)$. Then, there exist a constant $C_0 > 0$ and a permutation π such that, for all $j \in J$,

$$\|\tilde{\mathbf{g}}_j - \mathbf{q}_{\pi(j)}\|_2 \le C_0 \kappa(\mathbf{Q})\epsilon,$$

where $\tilde{\mathbf{g}}_k$ and $\mathbf{q}_k$ are the k-th rows of matrices $\tilde{\mathbf{G}}$ and $\mathbf{Q}$, respectively.

Note that this error bound depends on the upper bound on the individual errors $\|\mathbf{e}_i^{\mathrm{T}}\mathbf{N}\|_2$. From the statistical point of view, one might expect that there should be an algorithm, which improves upon this error bound if there are many nearly "pure" rows in matrix $\mathbf{G}$, so that the value of the error is diminished by averaging. However, to the best of our knowledge, no such algorithm complemented with a performance analysis can be found in the literature.

We now consider a specific instance of NSMF model given by (7). In this case, $\tilde{\mathbf{G}} = \hat{\mathbf{U}}$ and $\mathbf{Q} = \mathbf{HO}$ for an orthogonal matrix $\mathbf{O}$. Specifically, $\mathbf{O}$ is the orthogonal matrix, for which (15) holds (it is the same matrix $\mathbf{O}$, for which the bound of Lemma 1 is valid). Combining Theorem 4 with Lemma 1 we get the following corollary.

Corollary 5. *Let Assumptions 1 and 2 be satisfied with constant $\bar{C} \leq C_*$. Consider the matrices $\mathbf{\Pi}$, $\mathbf{X}$, $\mathbf{H}$, $\hat{\mathbf{U}}$ as in (4) – (6) such that $\lambda_K(\mathbf{\Pi}) > 0$ and $\|\mathbf{X} - \mathbf{\Pi}\| \leq \frac{1}{2}\lambda_K(\mathbf{\Pi})$. Let $\mathbf{O}$ be the orthogonal matrix, for which (15) holds. Let J be the set of indices returned by the preconditioned SPA with input $(\hat{\mathbf{U}}, K)$, and let $\hat{\mathbf{H}} = \hat{\mathbf{U}}_J$. Then, there exist a constant $C_0 > 0$ and a permutation π such that, for all $j = 1, \ldots, K$,*

$$\left\| \hat{\mathbf{h}}_j - \mathbf{q}_{\pi(j)} \right\|_2 \leq C_0 \kappa(\mathbf{H})\beta(\mathbf{X}, \mathbf{\Pi}), \tag{18}$$

where $\hat{\mathbf{h}}_k$ and $\mathbf{q}_k$ are the k-th rows of matrices $\hat{\mathbf{H}}$ and $\mathbf{HO}$, respectively. Furthermore,

$$\left\| \hat{\mathbf{H}} - \tilde{\mathbf{P}}\mathbf{HO} \right\|_F \leq C_0 K^{1/2} \kappa(\mathbf{W})\beta(\mathbf{X}, \mathbf{\Pi}), \tag{19}$$

where $\tilde{\mathbf{P}}$ is a permutation matrix corresponding to the permutation π.

Proof. Taking into account equations (4) – (7) we apply Theorem 4 with $\mathbf{Q} = \mathbf{HO}$, $\mathbf{N} = \hat{\mathbf{U}} - \mathbf{UO}$. By Lemma 1,

$$\|\mathbf{e}_i^{\mathrm{T}}\mathbf{N}\|_2 = \|\mathbf{e}_i^{\mathrm{T}}(\hat{\mathbf{U}} - \mathbf{UO})\|_2 \leq \epsilon, \quad i = 1, \ldots, n,$$

where $\epsilon = \beta(\mathbf{X}, \mathbf{\Pi})$. Therefore, using Assumption 3, (39), and the fact that $\bar{C} \leq C^*$ we have

$$\epsilon \leq \frac{\bar{C}}{\lambda_1(\mathbf{W})K\sqrt{K}} = \frac{\bar{C}\lambda_{\min}(\mathbf{H})}{K\sqrt{K}} \leq \frac{C_*\lambda_{\min}(\mathbf{HO})}{K\sqrt{K}}.$$

Thus, the assumptions of Theorem 4 are satisfied and we deduce from Theorem 4 that

$$\left\| \hat{\mathbf{h}}_j - \mathbf{q}_{\pi(j)} \right\|_2 \leq C_0 \kappa(\mathbf{HO})\epsilon = C_0 \kappa(\mathbf{H})\beta(\mathbf{X}, \mathbf{\Pi}),$$

where $\hat{\mathbf{h}}_k$ and $\mathbf{q}_k$ are the k-th rows of matrices $\hat{\mathbf{H}}$ and $\mathbf{HO}$, respectively. Thus, (18) follows. Inequality (19) is an immediate consequence of (18) and of the equality $\kappa(\mathbf{H}) = \kappa(\mathbf{W})$ (cf. (40)). $\qquad\square$

A.3 Concentration Bounds for Multinomial Matrices

In this section, we provide a bound with high probability on the spectral norm of matrix $\mathbf{X} - \mathbf{\Pi}$. Recall that, by definition, $\mathbf{X}^{\mathrm{T}} = [X_1, \ldots, X_n]$ is such that $NX_i \in \mathbb{R}^p$ are independent random vectors distributed according to p-dimensional multinomial distribution with parameters (N, Π_i). We will use matrix Bernstein inequality in the following form (cf. Theorem 6.1.1 in [43]).

Proposition 2 (Matrix Bernstein inequality). *Let $\mathbf{Z}_1, \ldots, \mathbf{Z}_N$ be independent zero-mean $n \times p$ random matrices such that $\|\mathbf{Z}_m\| \leq L$ for $m = 1, \ldots, N$. Then, for all $t > 0$ we have*

$$\mathbb{P}\left(\left\| \frac{1}{N} \sum_{m=1}^{N} \mathbf{Z}_m \right\| \geq t \right) \leq (n + p)\exp\left(-\frac{t^2 N^2}{2(\sigma^2 + LtN/3)} \right),$$

where

$$\sigma^2 = \max\left\{ \left\| \sum_{m=1}^{N} \mathbb{E}\left(\mathbf{Z}_m \mathbf{Z}_m^{\mathrm{T}} \right) \right\|, \left\| \sum_{m=1}^{N} \mathbb{E}\left(\mathbf{Z}_m^{\mathrm{T}} \mathbf{Z}_m \right) \right\| \right\}.$$

18

Applying Proposition 2 to our setting we obtain the following result.

Proposition 3. *Let $\boldsymbol{X}^{\mathrm{T}} = [X_1, \ldots, X_n]$ be such that $NX_i \in \mathbb{R}^p$ are independent random vectors distributed according to p-dimensional multinomial distribution with parameters (N, Π_i). Then, for all $t > 0$ we have*

$$\mathbb{P}\left(\|\boldsymbol{X} - \boldsymbol{\Pi}\| \geq t\right) \leq (n+p)\exp\left(-\frac{t^2 N}{2\sqrt{2}(n + t\sqrt{n}/3)}\right). \tag{20}$$

Proof. We prove (20) for $\boldsymbol{X}^{\mathrm{T}} - \boldsymbol{\Pi}^{\mathrm{T}}$ rather than $\boldsymbol{X} - \boldsymbol{\Pi}$, which is equivalent. Matrix $\boldsymbol{Z}^{\mathrm{T}} = \boldsymbol{X}^{\mathrm{T}} - \boldsymbol{\Pi}^{\mathrm{T}}$ has the form $\boldsymbol{Z}^{\mathrm{T}} = [Z_1, \ldots, Z_n]$ with independent column vectors $Z_i = \frac{1}{N}\sum_{m=1}^{N}(T_{im} - \mathbb{E}(T_{im}))$, where vectors T_{im} are distributed according to p-dimensional multinomial distribution with parameters $(1, \Pi_i)$ and independent over m for any fixed i. Here, we have used the fact that $\text{Multinomial}_p(N, \Pi_i)$ is a sum of N independent $\text{Multinomial}_p(1, \Pi_i)$ vectors. We also have $\boldsymbol{\Pi}^{\mathrm{T}} = [\Pi_1, \ldots, \Pi_n]$. Thus, we can write

$$\boldsymbol{Z}^{\mathrm{T}} = \frac{1}{N}\sum_{m=1}^{N}(\boldsymbol{T}_m - \mathbb{E}(\boldsymbol{T}_m)) = \frac{1}{N}\sum_{m=1}^{N}\boldsymbol{Z}_m, \tag{21}$$

where $\boldsymbol{T}_m = [T_{1m}, \ldots, T_{nm}]$ and $\boldsymbol{Z}_m = \boldsymbol{T}_m - \mathbb{E}(\boldsymbol{T}_m)$ are independent zero-mean random matrices.

We apply Proposition 2 to the sum (21). The first step is to evaluate $\left\|\sum_{m=1}^{N}\mathbb{E}\left(\boldsymbol{Z}_m\boldsymbol{Z}_m^{\mathrm{T}}\right)\right\|$. Let $T_{im}(k)$ denote the k-th component of T_{im}, $k = 1, \ldots, p$. We have $\mathbb{E}(T_{im}(k)) = \Pi_{ik}$, $\text{Var}(T_{im}(k)) = \Pi_{ik}(1 - \Pi_{ik})$, $\text{Cov}(T_{im}(k), T_{im}(j)) = -\Pi_{ik}\Pi_{ij}$ for $i \neq j$. Therefore,

$$\mathbb{E}\left(\boldsymbol{Z}_m\boldsymbol{Z}_m^{\mathrm{T}}\right) = \mathbb{E}\left(\boldsymbol{T}_m\boldsymbol{T}_m^{\mathrm{T}}\right) - \mathbb{E}\left(\boldsymbol{T}_m\right)\mathbb{E}\left(\boldsymbol{T}_m^{\mathrm{T}}\right)$$
$$= \mathbb{E}\sum_{i=1}^{n}T_{im}T_{im}^{\mathrm{T}} - \boldsymbol{\Pi}\boldsymbol{\Pi}^{\mathrm{T}} = \sum_{i=1}^{n}\boldsymbol{Y}_i,$$

where

$$\boldsymbol{Y}_i = \text{diag}(\Pi_{i1}, \ldots, \Pi_{ip}) - \Pi_i\Pi_i^{\mathrm{T}}.$$

The spectral norm of $\boldsymbol{Y}_i$ satisfies

$$\|\boldsymbol{Y}_i\|^2 \leq \|\boldsymbol{Y}_i\|_F^2 = \sum_{k=1}^{p}\Pi_{ik}^2 + \left(\sum_{k=1}^{p}\Pi_{ik}^2\right)^2 - 2\sum_{k=1}^{p}\Pi_{ik}^3 \leq 2,$$

where we have used the fact that $\sum_{k=1}^{p}\Pi_{ik}^2 \leq \sum_{k=1}^{p}\Pi_{ik} = 1$. Thus, $\left\|\mathbb{E}\left(\boldsymbol{Z}_m\boldsymbol{Z}_m^{\mathrm{T}}\right)\right\| \leq \sqrt{2}n$ and

$$\left\|\sum_{m=1}^{N}\mathbb{E}\left(\boldsymbol{Z}_m\boldsymbol{Z}_m^{\mathrm{T}}\right)\right\| \leq \sqrt{2}Nn. \tag{22}$$

Next, we derive an upper bound on $\left\|\sum_{m=1}^{N}\mathbb{E}\left(\boldsymbol{Z}_m^{\mathrm{T}}\boldsymbol{Z}_m\right)\right\|$. Note that $\mathbb{E}(\boldsymbol{T}_m^{\mathrm{T}}\boldsymbol{T}_m)$ is a matrix with diagonal entries $\mathbb{E}(T_{im}^{\mathrm{T}}T_{im}) = \sum_{k=1}^{p}\Pi_{ik} = 1$ while its off-diagonal entries are $\mathbb{E}(T_{im}^{\mathrm{T}}T_{jm}) = [\mathbb{E}(T_{im})]^{\mathrm{T}}\mathbb{E}(T_{jm}) = \Pi_i^{\mathrm{T}}\Pi_j$ due to independence between T_{im} and T_{jm} for $i \neq j$. Also, $\mathbb{E}(\boldsymbol{T}_m^{\mathrm{T}})\mathbb{E}(\boldsymbol{T}_m) = \boldsymbol{\Pi}^{\mathrm{T}}\boldsymbol{\Pi}$ is a matrix with entries $\Pi_i^{\mathrm{T}}\Pi_j$. Hence,

$$\mathbb{E}\left(\boldsymbol{Z}_m^{\mathrm{T}}\boldsymbol{Z}_m\right) = \mathbb{E}(\boldsymbol{T}_m^{\mathrm{T}}\boldsymbol{T}_m) - \mathbb{E}(\boldsymbol{T}_m^{\mathrm{T}})\mathbb{E}(\boldsymbol{T}_m) = \text{diag}\left(1 - \|\Pi_1\|_2^2, \ldots, 1 - \|\Pi_n\|_2^2\right). \tag{23}$$

It follows that $\left\|\mathbb{E}\left(\boldsymbol{Z}_m^{\mathrm{T}}\boldsymbol{Z}_m\right)\right\| \leq 1$, and thus $\left\|\sum_{m=1}^{N}\mathbb{E}\left(\boldsymbol{Z}_m^{\mathrm{T}}\boldsymbol{Z}_m\right)\right\| \leq N$. Combining this inequality with (22) we obtain that σ^2 defined in Proposition 2 satisfies $\sigma^2 \leq \sqrt{2}Nn$.

Finally, we specify the constant L that gives an upper bound on $\|\boldsymbol{Z}_m\|$. Let $u \in S^{p-1}$ be an element of the unit sphere in $\mathbb{R}^p$. Since for any i vector T_{im} has only one component equal to 1 and all other components 0 we have $\|T_{im} - \mathbb{E}(T_{im})\|_2^2 = \|T_{im} - \Pi_i\|_2^2 \leq 2$ and thus

$$\left| u^{\mathrm{T}}(T_{im} - \mathbb{E}(T_{im})) \right| \leq \sqrt{2}.$$

It follows that

$$\|\boldsymbol{T}_m - \mathbb{E}(\boldsymbol{T}_m)\|^2 = \sup_{u \in S^{p-1}} \left\| u^{\mathrm{T}}(\boldsymbol{T}_m - \mathbb{E}(\boldsymbol{T}_m)) \right\|_2^2$$

$$= \sup_{u \in S^{p-1}} \sum_{i=1}^{n} \left| u^{\mathrm{T}}(T_{im} - \mathbb{E}(T_{im})) \right|^2 \leq 2n$$

and we get $\|\boldsymbol{Z}_m\| \leq \sqrt{2n} =: L$ for any $m = 1, \ldots, n$. The desired result now follows by applying Proposition 2 with $\sigma^2 \leq \sqrt{2}Nn$ and $L = \sqrt{2n}$. $\qquad\square$

The next lemma is a corollary of Proposition 3:

Lemma 4. *Let the assumptions of Proposition 3 be satisfied. Assume that $N \geq \log(n+p)$ and $\min(n, p) \geq 2$. Then*

$$\mathbb{P}\left(\|\boldsymbol{X} - \boldsymbol{\Pi}\| \geq 4\sqrt{\frac{n\log(n+p)}{N}} \right) \leq (n+p)^{-1}. \tag{24}$$

Furthermore,

$$\mathbb{P}\left(\max_{1 \leq i \leq n} \|\mathbf{e}_i^{\mathrm{T}}(\boldsymbol{X} - \boldsymbol{\Pi})\|_2 \geq 5\sqrt{\frac{\log(n+p)}{N}} \right) \leq (n+p)^{-1}. \tag{25}$$

Proof. Inequality (24) follows easily from Proposition 3 by setting $t = 4\sqrt{\frac{n\log(n+p)}{N}}$ and using the assumptions $N \geq \log(n+p)$. In order to prove (25), we bound each probability $\mathbb{P}\left(\|\mathbf{e}_i^{\mathrm{T}}(\boldsymbol{X} - \boldsymbol{\Pi})\|_2 \geq 5\sqrt{\frac{\log(n+p)}{N}} \right)$ via Proposition 3 with $n = 1$ (that is, we apply Proposition 3 to $1 \times p$ matrices $\mathbf{e}_i^{\mathrm{T}}\boldsymbol{X}$, $\mathbf{e}_i^{\mathrm{T}}\boldsymbol{\Pi}$) and then use the union bound. This yields

$$\mathbb{P}\left(\max_{1 \leq i \leq n} \|\mathbf{e}_i^{\mathrm{T}}(\boldsymbol{X} - \boldsymbol{\Pi})\|_2 \geq 5\sqrt{\frac{\log(n+p)}{N}} \right) \leq n(p+1)\exp\left(-\frac{75\log(n+p)}{16\sqrt{2}} \right).$$

The right hand side of this inequality does not exceed $(n+p)^{-1}$. $\qquad\square$

B   Proofs of the Main Results

B.1   Proof of Lemma 1

Using the fact that $\widehat{\mathbf{V}}_1^{\mathrm{T}}\widehat{\mathbf{V}}^{\mathrm{T}} = 0$ we obtain

$$\|\mathbf{e}_i^{\mathrm{T}}(\hat{\mathbf{U}} - \mathbf{U}\mathbf{O})\|_2 = \|\mathbf{e}_i^{\mathrm{T}}(\boldsymbol{X}\hat{\mathbf{V}}\hat{\mathbf{L}}^{-1} - \boldsymbol{\Pi}\mathbf{V}\mathbf{L}^{-1}\mathbf{O})\|_2$$

$$= \|\mathbf{e}_i^{\mathrm{T}}\boldsymbol{X}\hat{\mathbf{V}}(\hat{\mathbf{L}}^{-1} - \widetilde{\mathbf{O}}^{\mathrm{T}}\mathbf{L}^{-1}\mathbf{O}) + \mathbf{e}_i^{\mathrm{T}}\boldsymbol{X}(\hat{\mathbf{V}} - \mathbf{V}\widetilde{\mathbf{O}})\widetilde{\mathbf{O}}^{\mathrm{T}}\mathbf{L}^{-1}\mathbf{O} + \mathbf{e}_i^{\mathrm{T}}(\boldsymbol{X} - \boldsymbol{\Pi})\mathbf{V}\mathbf{L}^{-1}\mathbf{O}\|_2$$

$$\leq \|\mathbf{e}_i^{\mathrm{T}}\boldsymbol{X}\hat{\mathbf{V}}(\hat{\mathbf{L}}^{-1} - \widetilde{\mathbf{O}}^{\mathrm{T}}\mathbf{L}^{-1}\mathbf{O})\|_2 + \|\mathbf{e}_i^{\mathrm{T}}\boldsymbol{X}(\hat{\mathbf{V}} - \mathbf{V}\widetilde{\mathbf{O}})\widetilde{\mathbf{O}}^{\mathrm{T}}\mathbf{L}^{-1}\mathbf{O}\|_2 + \|\mathbf{e}_i^{\mathrm{T}}(\boldsymbol{X} - \boldsymbol{\Pi})\mathbf{V}\mathbf{L}^{-1}\mathbf{O}\|_2$$

$$= G_1 + G_2 + G_3.$$

20

We now bound the values G_1, G_2 and G_3 separately. We have

$$G_1 = \|\mathbf{e}_i^{\mathrm{T}} \boldsymbol{X} \hat{\mathbf{V}} (\hat{\mathbf{L}}^{-1} - \widetilde{\mathbf{O}}^{\mathrm{T}} \mathbf{L}^{-1} \mathbf{O})\|_2 \leq \|\mathbf{e}_i^{\mathrm{T}} \boldsymbol{X}\|_2 \|\hat{\mathbf{V}}\| \|\hat{\mathbf{L}}^{-1} - \widetilde{\mathbf{O}}^{\mathrm{T}} \mathbf{L}^{-1} \mathbf{O}\|$$

$$\leq CK^{1/2} \kappa^2(\boldsymbol{\Pi}) \frac{\|\mathbf{e}_i^{\mathrm{T}} \boldsymbol{X}\|_2 \|\boldsymbol{X} - \boldsymbol{\Pi}\|}{\lambda_K^2(\boldsymbol{\Pi})},$$

where the last inequality is due to Lemma 3. The values G_2 and G_3 can be controlled using the bounds for the norm of matrix product and Corollary 4:

$$G_2 = \|\mathbf{e}_i^{\mathrm{T}} \boldsymbol{X} (\hat{\mathbf{V}} - \mathbf{V} \widetilde{\mathbf{O}}) \widetilde{\mathbf{O}}^{\mathrm{T}} \mathbf{L}^{-1} \mathbf{O}\|_2 \leq \|\mathbf{e}_i^{\mathrm{T}} \boldsymbol{X}\|_2 \|\hat{\mathbf{V}} - \mathbf{V} \widetilde{\mathbf{O}}\| \|\mathbf{L}^{-1}\|$$

$$= \frac{\|\mathbf{e}_i^{\mathrm{T}} \boldsymbol{X}\|_2 \|\hat{\mathbf{V}} - \mathbf{V} \widetilde{\mathbf{O}}\|}{\lambda_K(\boldsymbol{\Pi})} \leq 5\sqrt{2} \kappa(\boldsymbol{\Pi}) \frac{\|\mathbf{e}_i^{\mathrm{T}} \boldsymbol{X}\|_2 \|\boldsymbol{X} - \boldsymbol{\Pi}\|}{\lambda_K^2(\boldsymbol{\Pi})}$$

$$G_3 = \|\mathbf{e}_i^{\mathrm{T}} (\boldsymbol{X} - \boldsymbol{\Pi}) \mathbf{V} \mathbf{L}^{-1} \mathbf{O}\|_2 \leq \|\mathbf{e}_i^{\mathrm{T}} (\boldsymbol{X} - \boldsymbol{\Pi})\|_2 \|\mathbf{V}\| \|\mathbf{L}^{-1}\| = \frac{\|\mathbf{e}_i^{\mathrm{T}} (\boldsymbol{X} - \boldsymbol{\Pi})\|_2}{\lambda_K(\boldsymbol{\Pi})}.$$

Combining these bounds proves the lemma.

B.2   Proof of Lemma 2

We first prove that matrix $\hat{\mathbf{H}}$ is non-degenerate. In this proof, we denote by $\mathbf{O}$ the orthogonal matrix, for which (15) holds, and by $\tilde{\mathbf{P}}$ the permutation matrix, for which the bound of Corollary 5 holds. Using Weyl's inequality [19, Theorem C.6] and Corollary 5 we obtain

$$\lambda_{\min}(\hat{\mathbf{H}}) \geq \lambda_{\min}(\tilde{\mathbf{P}} \mathbf{H} \mathbf{O}) - \|\hat{\mathbf{H}} - \tilde{\mathbf{P}} \mathbf{H} \mathbf{O}\|$$

$$\geq \lambda_{\min}(\mathbf{H}) - C_0 K^{1/2} \kappa(\boldsymbol{W}) \beta(\boldsymbol{X}, \boldsymbol{\Pi}).$$

Using this inequality, Assumption 2 with $\bar{C} \leq C_0^{-1}$, and equations (39), (40) we find

$$\lambda_{\min}(\hat{\mathbf{H}}) \geq \lambda_{\min}(\mathbf{H}) - \frac{1}{2\lambda_1(\boldsymbol{W})} = \frac{1}{2\lambda_1(\boldsymbol{W})}, \tag{26}$$

which proves that $\hat{\mathbf{H}}$ is invertible. Then, for the estimator $\hat{\boldsymbol{W}} = \hat{\mathbf{U}} \hat{\mathbf{H}}^{-1}$ and the permutation matrix $\mathbf{P} = \tilde{\mathbf{P}}^{-1}$ we have

$$\|\hat{\boldsymbol{W}} - \boldsymbol{W} \mathbf{P}\|_F = \|\hat{\mathbf{U}} \hat{\mathbf{H}}^{-1} - \mathbf{U} \mathbf{H}^{-1} \mathbf{P}\|_F$$

$$\leq \|\hat{\mathbf{U}} (\hat{\mathbf{H}}^{-1} - \mathbf{O}^{\mathrm{T}} \mathbf{H}^{-1} \mathbf{P})\|_F + \|(\hat{\mathbf{U}} - \mathbf{U} \mathbf{O}) [\mathbf{P}^{-1} \mathbf{H} \mathbf{O}]^{-1}\|_F$$

$$= I_1 + I_2.$$

We now bound separately I_1 and I_2. Due to (26) and (39) we have

$$\|\hat{\mathbf{H}}^{-1}\| = \frac{1}{\lambda_{\min}(\hat{\mathbf{H}})} \leq 2\lambda_1(\boldsymbol{W}), \quad \|\mathbf{H}^{-1}\| = \lambda_1(\boldsymbol{W}). \tag{27}$$

Using the fact that $\|\mathbf{A}^{-1} - \mathbf{B}^{-1}\|_F \leq \|\mathbf{A}^{-1}\| \|\mathbf{B}^{-1}\| \|\mathbf{A} - \mathbf{B}\|_F$ with $\mathbf{A} = \hat{\mathbf{H}}$, $\mathbf{B} = \mathbf{P}^{-1} \mathbf{H} \mathbf{O} = \tilde{\mathbf{P}} \mathbf{H} \mathbf{O}$, inequality (27) and Corollary 5 we find

$$I_1 = \|\hat{\mathbf{U}} (\hat{\mathbf{H}}^{-1} - \mathbf{O}^{\mathrm{T}} \mathbf{H}^{-1} \mathbf{P})\|_F \leq \|\hat{\mathbf{H}}^{-1} - \mathbf{O}^{\mathrm{T}} \mathbf{H}^{-1} \mathbf{P}\|_F$$

$$\leq \|\hat{\mathbf{H}}^{-1}\| \|\mathbf{O}^{\mathrm{T}}\mathbf{H}^{-1}\mathbf{P}\| \|\hat{\mathbf{H}} - \tilde{\mathbf{P}}\mathbf{H}\mathbf{O}\|_F$$
$$\leq 2C_0 K^{1/2}\lambda_1^2(\boldsymbol{W})\kappa(\boldsymbol{W})\beta(\boldsymbol{X}, \boldsymbol{\Pi}).$$

On the other hand,

$$I_2 = \|(\hat{\mathbf{U}} - \mathbf{U}\mathbf{O})[\mathbf{P}^{-1}\mathbf{H}\mathbf{O}]^{-1}\|_F \leq \|\hat{\mathbf{U}} - \mathbf{U}\mathbf{O}\|_F \|\mathbf{H}^{-1}\|$$
$$\leq \frac{5\sqrt{2K}\kappa(\boldsymbol{\Pi})\|\boldsymbol{X} - \boldsymbol{\Pi}\|}{\lambda_K(\boldsymbol{\Pi})}\|\mathbf{H}^{-1}\|,$$

where the last inequality follows from Corollary 4. Combining the above bounds we get

$$\|\hat{\boldsymbol{W}} - \boldsymbol{W}\mathbf{P}\|_F \leq CK^{1/2}\lambda_1(\boldsymbol{W})\left\{\lambda_1(\boldsymbol{W})\kappa(\boldsymbol{W})\beta(\boldsymbol{X}, \boldsymbol{\Pi}) + \frac{\kappa(\boldsymbol{\Pi})\|\boldsymbol{X} - \boldsymbol{\Pi}\|}{\lambda_K(\boldsymbol{\Pi})}\right\}.$$

B.3   Proof of Theorem 1

We apply Lemma 2 combined with the concentration inequalities of Lemma 4. First, we check that Assumption 2 holds with probability at least $1 - 2(n + p)^{-1}$. For $N \geq \log(n + p)$ we get from Lemma 4 that

$$\|\boldsymbol{X} - \boldsymbol{\Pi}\| \leq 4\sqrt{\frac{n\log(n + p)}{N}}$$

with probability at least $1 - 1/(n + p)$, and

$$\max_{1 \leq i \leq n} \|\mathbf{e}_i^{\mathrm{T}}(\boldsymbol{X} - \boldsymbol{\Pi})\|_2 \leq 5\sqrt{\frac{\log(n + p)}{N}}$$

with probability at least $1 - 1/(n+p)$. Notice also that $\max_i \|\mathbf{e}_i^{\mathrm{T}}\boldsymbol{X}\|_2 = \max_i \sqrt{\sum_{j=1}^p X_{ij}^2} \leq \max_i \sqrt{\sum_{j=1}^p X_{ij}} = 1$. Putting together the above remarks we deduce that, with probability at least $1 - 2(n + p)^{-1}$,

$$\beta(\boldsymbol{X}, \boldsymbol{\Pi}) \leq 5\left\{\kappa^2(\boldsymbol{\Pi})\sqrt{Kn} + \lambda_K(\boldsymbol{\Pi})\right\}\frac{\sqrt{\log(n + p)}}{\lambda_K^2(\boldsymbol{\Pi})\sqrt{N}}$$
$$\leq 10\kappa^2(\boldsymbol{\Pi})\sqrt{Kn}\frac{\sqrt{\log(n + p)}}{\lambda_K^2(\boldsymbol{\Pi})\sqrt{N}}$$

where we have used the inequality $\lambda_K(\boldsymbol{\Pi}) \leq \sqrt{n/K}$ proved in Lemma 7. Since $\lambda_K(\boldsymbol{\Pi})$ is chosen to satisfy (10) we get that, with probability at least $1 - 2(n + p)^{-1}$,

$$\beta(\boldsymbol{X}, \boldsymbol{\Pi}) \leq \frac{\bar{C}}{\lambda_1(\boldsymbol{W})\kappa(\boldsymbol{W})K\sqrt{K}}.$$

Thus, on an event that has probability at least $1 - 2(n + p)^{-1}$, Assumption 2 is satisfied and we can apply Lemma 2. This yields that, with probability at least $1 - 2(n + p)^{-1}$,

$$\min_{\mathbf{P} \in \mathscr{P}}\|\hat{\boldsymbol{W}} - \boldsymbol{W}\mathbf{P}\|_F \leq CK^{1/2}\lambda_1(\boldsymbol{W})\left\{\lambda_1(\boldsymbol{W})\kappa(\boldsymbol{W})\beta(\boldsymbol{X}, \boldsymbol{\Pi}) + \frac{\kappa(\boldsymbol{\Pi})\|\boldsymbol{X} - \boldsymbol{\Pi}\|}{\lambda_K(\boldsymbol{\Pi})}\right\}$$

$$\leq C \frac{\lambda_1(\boldsymbol{W})\sqrt{nK\log(n+p)}}{\sqrt{N}\lambda_K(\boldsymbol{\Pi})} \left\{ \lambda_1(\boldsymbol{W})\kappa(\boldsymbol{W}) \frac{K^{1/2}\kappa^2(\boldsymbol{\Pi})}{\lambda_K(\boldsymbol{\Pi})} + \kappa(\boldsymbol{\Pi}) \right\}$$

$$\leq CK\sqrt{\frac{n\log(n+p)}{N}} \left(\frac{\lambda_1(\boldsymbol{W})}{\lambda_K(\boldsymbol{\Pi})} \right)^2 \kappa(\boldsymbol{W})\kappa^2(\boldsymbol{\Pi}),$$

where we have used the inequalities $\lambda_K(\boldsymbol{\Pi}) \leq \sqrt{n/K}$ and $\lambda_1(\boldsymbol{\Pi}) \leq \sqrt{K}\lambda_1(\boldsymbol{W})$ (see Lemma 7).

B.4 Proof of Theorem 2 and Corollary 3

We will use the following lemma.

Lemma 5. *Let $\boldsymbol{\Pi} \in \mathbb{R}^{n \times p}$ be a rank K matrix with smallest non-zero singular value $\lambda_K(\boldsymbol{\Pi})$, and $\boldsymbol{X} \in \mathbb{R}^{n \times p}$ be a matrix such that $\|\boldsymbol{X} - \boldsymbol{\Pi}\| \leq \tau$ for some $\tau > 0$. Let $\hat{K} = \max\{j : \lambda_j(\boldsymbol{X}) > \tau\}$. If $\lambda_K(\boldsymbol{\Pi}) > 2\tau$ then $\hat{K} = K$.*

Proof. By Weyl's inequality, we have $|\lambda_j(\boldsymbol{X}) - \lambda_j(\boldsymbol{\Pi})| \leq \tau$ for all j. Since $\lambda_j(\boldsymbol{\Pi}) = 0$ for $j \geq K+1$ we deduce that $\hat{K} \leq K$. On the other hand, $\hat{K} \geq K$. Indeed, condition $\lambda_K(\boldsymbol{\Pi}) > 2\tau$ implies that $\lambda_K(\boldsymbol{X}) \geq \lambda_K(\boldsymbol{\Pi}) - |\lambda_j(\boldsymbol{X}) - \lambda_j(\boldsymbol{\Pi})| > \tau$. $\square$

Theorem 2 is obtained by combining Theorem 1 with Lemma 5. Indeed, notice that the bound of Theorem 1 is proved on the event $\mathscr{A} := \left\{ \|\boldsymbol{X} - \boldsymbol{\Pi}\| \leq 4\sqrt{\frac{n\log(n+p)}{N}} \right\}$. Set $\tau = 4\sqrt{\frac{n\log(n+p)}{N}}$. It follows from Lemma 5 that if

$$\lambda_K(\boldsymbol{\Pi}) > 8\sqrt{\frac{n\log(n+p)}{N}} \tag{28}$$

then on the event $\mathscr{A}$ we have $\hat{K} = K$. But condition (28) is implied by (10) and (12). Therefore, the proof of Theorem 1 goes through verbatim if we replace K by $\hat{K}$. This yields Theorem 2. Corollary 3 is deduced from Theorem 2 in the same way as Corollary 1 was deduced from Theorem 1.

B.5 Proof of Theorem 3

We use the techniques of proving minimax lower bounds based on a reduction to the problem of testing multiple hypotheses [44, Chapter 2]. The hypotheses correspond to probability measures $\mathbb{P}_{\boldsymbol{\Pi}^{(j)}}$, where $\boldsymbol{\Pi}^{(j)} = \boldsymbol{W}^{(j)}\boldsymbol{A}$ with carefully chosen matrix $\boldsymbol{A}$ and matrices $\boldsymbol{W}^{(j)}$, $j = 0, 1, \ldots, T$. The construction of these matrices borrows some elements from the proofs of the lower bounds in [8, 22]. An additional subtlety is related to the fact that we need to grant Assumption 3 on the singular values. Without loss of generality we assume that n is a multiple of K and that K is even.

1. *Construction of the set of matrices $\boldsymbol{W}^{(j)}$.*
We first introduce the basic matrix $\boldsymbol{W}^{(0)}$ and then define matrices $\boldsymbol{W}^{(j)}$, $j = 1, \ldots, T$ as slightly perturbed versions of $\boldsymbol{W}^{(0)}$.

Let $\boldsymbol{D}_1$ be a $n \times K$ matrix composed of n/K blocks, each of which is the identity matrix $\boldsymbol{I}_K$ of size K:

$$\boldsymbol{D}_1^{\mathrm{T}} = \left[\begin{array}{c|c|c|c} \boldsymbol{I}_K & \boldsymbol{I}_K & \rule{1cm}{0.4pt} & \boldsymbol{I}_K \end{array} \right].$$

We have that $\boldsymbol{D}_1^{\mathrm{T}}\boldsymbol{D}_1 = (n/K)\boldsymbol{I}_K$ and $\sigma(\boldsymbol{D}_1) = \{\sqrt{n/K}, 0\}$, where $\sigma(\boldsymbol{D}_1)$ denotes the set of singular values of $\boldsymbol{D}_1$. Set

$$\gamma_1 = \frac{1}{4\,K}$$

and define the $n \times K$ matrix $\boldsymbol{D}_2$ by the relation

$$\boldsymbol{D}_2^{\mathrm{T}} = \gamma_1 \left[\begin{array}{c|c} \boldsymbol{0}_{K,K} & \boldsymbol{1}_{K,(n-K)} \end{array} \right]$$

where we denote by $\mathbf{1}_{n,p}$ (respectively, $\mathbf{0}_{n,p}$) the $n \times p$ matrix with all entries 1 (respectively, 0). Then, $\boldsymbol{D}_2^{\mathrm{T}} \boldsymbol{D}_2 = (n-K)\gamma_1^2 \mathbf{1}_{K,K}$ and $\sigma(\boldsymbol{D}_2) = \{\gamma_1\sqrt{K(n-K)}, 0\}$. We will further consider the matrix $\boldsymbol{D}_3 = \boldsymbol{D}_1 + \boldsymbol{D}_2$ given by the relation

$$\boldsymbol{D}_3^{\mathrm{T}} = \begin{bmatrix} 1 & \cdots & 0 & (1+\gamma_1) & \gamma_1 & \ldots & \gamma_1 & \ldots\ldots & (1+\gamma_1) & \gamma_1 & \ldots & \gamma_1 \\ \vdots & & \vdots & \gamma_1 & (1+\gamma_1) & \ldots & \gamma_1 & \ldots\ldots & \gamma_1 & (1+\gamma_1) & \ldots & \gamma_1 \\ \vdots & & \vdots & \vdots & & & \vdots & \ldots\ldots & \vdots & & & \vdots \\ 0 & \cdots & 1 & \gamma_1 & \gamma_1 & \ldots & (1+\gamma_1) & \ldots\ldots & \gamma_1 & \gamma_1 & \ldots & (1+\gamma_1) \end{bmatrix}$$
$$= \begin{bmatrix} \boldsymbol{I}_K & | & \boldsymbol{I}_K + \gamma_1\mathbf{1}_{K,K} & | & \ldots & | & \boldsymbol{I}_K + \gamma_1\mathbf{1}_{K,K}. \end{bmatrix}.$$

Applying Weyl's inequality [19, Theorem C.6], we get

$$\begin{cases} \lambda_1(\boldsymbol{D}_3) \leq \sqrt{n/K} + \frac{1}{4}\sqrt{(n-K)/K} \leq \frac{5}{4}\sqrt{n/K}, \\ \lambda_K(\boldsymbol{D}_3) \geq \sqrt{n/K} - \frac{1}{4}\sqrt{(n-K)/K} \geq \frac{3}{4}\sqrt{n/K}. \end{cases}$$

Finally, the basic matrix $\boldsymbol{W}^{(0)}$ is defined by the relation

$$(\boldsymbol{W}^{(0)})^{\mathrm{T}} = \boldsymbol{D}_3^{\mathrm{T}} - \begin{bmatrix} \mathbf{0}_{K,K} & | & K\gamma_1\boldsymbol{I}_K & | & \ldots & | & K\gamma_1\boldsymbol{I}_K \end{bmatrix}$$
$$= \begin{bmatrix} \boldsymbol{I}_K & | & (1-K\gamma_1)\boldsymbol{I}_K + \gamma_1\mathbf{1}_{K,K} & | & \ldots & | & (1-K\gamma_1)\boldsymbol{I}_K + \gamma_1\mathbf{1}_{K,K} \end{bmatrix}.$$

Clearly, $\boldsymbol{W}^{(0)}$ satisfies Assumption 1, all entries of $\boldsymbol{W}^{(0)}$ are non-negative and its rows sum up to 1. Applying Weyl's inequality to matrix $\boldsymbol{W}^{(0)}$ yields

$$\begin{cases} \lambda_1(\boldsymbol{W}^{(0)}) \leq \frac{5}{4}\sqrt{n/K} + K\gamma_1(\sqrt{n/K-1}) \leq \frac{3}{2}\sqrt{n/K}, \\ \lambda_K(\boldsymbol{W}^{(0)}) \geq \frac{3}{4}\sqrt{n/K} - K\gamma_1(\sqrt{n/K-1}) \geq \frac{1}{2}\sqrt{n/K}, \end{cases} \tag{29}$$

implying that $\kappa(\boldsymbol{W}^{(0)}) \leq 3$.

Our next step is to define the matrices $\boldsymbol{W}^{(j)}$, $j = 1, \ldots, T$. Consider the set of binary sequences

$$M = \{0,1\}^{K(n-K)/2}.$$

Applying the Varshamov-Gilbert bound [44, Lemma 2.9] we get that there exist $w^{(j)} \in M$, $j = 1, \ldots, T$, such that:

$$\|w^{(i)} - w^{(j)}\|_1 = \|w^{(i)} - w^{(j)}\|_2^2 \geq \frac{K(n-K)}{16}, \text{ for any } 0 \leq i \neq j \leq T, \tag{30}$$

with $w^{(0)} = 0$ and

$$\log T \geq \frac{\log 2}{16} K(n-K). \tag{31}$$

We divide each $w^{(j)}$ into $(n-K)$ chunks as $w^{(j)} = (w_1^{(j)}, w_2^{(j)}, \ldots, w_{n-K}^{(j)})$ with $w_i^{(j)} \in \{0,1\}^{K/2}$. Next, for each $w_i^{(j)}$, we introduce its augmented counterpart defined as $\tilde{w}_i^{(j)} = (w_i^{(j)}, -w_i^{(j)}) \in \{-1,0,1\}^K$. In what follows, we set

$$\gamma = c_* \sqrt{\frac{N}{K(N-K)^2}},$$

where $c_* > 0$ is a small enough absolute constant. For $1 \leq j \leq T$, define the $(n - K) \times K$ matrix $\Omega^{(j)}$ and the $n \times K$ matrix $\Delta^{(j)}$ as follows:

$$\Omega^{(j)} = \gamma \begin{bmatrix} \tilde{w}_1^{(j)} \\ \tilde{w}_2^{(j)} \\ \vdots \\ \tilde{w}_{n-K}^{(j)} \end{bmatrix} \text{ and } \Delta^{(j)} = \left[\frac{\mathbf{0}_{K,K}}{\Omega^{(j)}} \right].$$

Note that all the entries of $(\Delta^{(j)})^{\mathrm{T}} \Delta^{(j)}$ are bounded in absolute value by $\gamma^2 (n - K)$, which yields

$$\|\Delta^{(j)}\| = \sqrt{\|(\Delta^{(j)})^{\mathrm{T}} \Delta^{(j)}\|} \leq \sqrt{\|(\Delta^{(j)})^{\mathrm{T}} \Delta^{(j)}\|_F} \leq \gamma \sqrt{(n - K)K}.$$

Thus, choosing c^* small enough and using the assumption that $N \geq 2K$ we obtain

$$\|\Delta^{(j)}\| \leq \frac{1}{4} \sqrt{n/K}.$$

Now, for $1 \leq j \leq T$, we define $\boldsymbol{W}^{(j)}$ as

$$\boldsymbol{W}^{(j)} = \boldsymbol{W}^{(0)} + \Delta^{(j)}. \tag{32}$$

It is easy to check that, for each $1 \leq j \leq T$, the rows of $\boldsymbol{W}^{(j)}$ are probability vectors if c^* is chosen small enough, and $\boldsymbol{W}^{(j)}$ satisfies Assumption 1. Moreover, using (29) and applying Weyl's inequality once again, we obtain

$$\begin{cases} \lambda_1(\boldsymbol{W}^{(j)}) \leq \frac{7}{4} \sqrt{n/K}, \\ \lambda_K(\boldsymbol{W}^{(j)}) \geq \frac{1}{4} \sqrt{n/K}, \end{cases} \tag{33}$$

so that $\kappa(\boldsymbol{W}^{(j)}) \leq 7$.

2. *Constructing matrix $\boldsymbol{A}$ and checking the fact that $\boldsymbol{\Pi}^{(j)} \in \mathscr{M}$, $j = 0, 1, \ldots, T$.*
Assume that p is a multiple of K (if it is not the case the definition of $\boldsymbol{A}$ should be modified by adding a block of zeros of the size of the residual). Define the following block matrix:

$$\boldsymbol{A}^0 = \{\underbrace{\mathbf{e}_1, 0_K, \ldots, 0_K}_{p/K}, \underbrace{\mathbf{e}_2, 0_K, \ldots, 0_K}_{p/K}, \ldots, \underbrace{\mathbf{e}_K, 0_K, \ldots, 0_K}_{p/K}\} \in \mathbb{R}^{K \times p},$$

where $(\mathbf{e}_1, \ldots, \mathbf{e}_K)$ is the canonical basis of $\mathbb{R}^K$ and $0_K \in \mathbb{R}^K$ is the vector with all entries 0. Define

$$\boldsymbol{A} := \frac{N - K}{N} \boldsymbol{A}^0 + \frac{K}{pN} \mathbf{1}_{K,p}.$$

All entries of $\boldsymbol{A}$ are non-negative and the rows of $\boldsymbol{A}$ sum up to 1. We have that $\sigma\left(\frac{N-K}{N} \boldsymbol{A}^0\right) = \left\{\frac{N-K}{N}, 0\right\}$ and $\sigma\left(\frac{K}{pN} \mathbf{1}_{K,p}\right) = \left\{\frac{K^{3/2}}{\sqrt{p}N}, 0\right\}$. Using the assumption that $K \leq p/4$ and Weyl's inequality we get

$$\begin{cases} \lambda_1(\boldsymbol{A}) \leq \frac{N-K}{N} + \frac{K^{3/2}}{\sqrt{p}N} \leq 1, \\ \lambda_K(\boldsymbol{A}) \geq \frac{N-K}{N} - \frac{K^{3/2}}{\sqrt{p}N} \geq 1/4, \end{cases} \tag{34}$$

which implies that $\kappa(\boldsymbol{A}) \leq 4$.

For $0 \leq j \leq T$, define $\boldsymbol{\Pi}^{(j)} = \boldsymbol{W}^{(j)} \boldsymbol{A}$. Using Lemma 8, (29), (33) and (34) we obtain

$$\lambda_K(\boldsymbol{\Pi}^{(j)}) = \lambda_K(\boldsymbol{W}^{(j)} \boldsymbol{A}) \geq \lambda_K(\boldsymbol{W}^{(j)}) \lambda_K(\boldsymbol{A}) \geq \frac{1}{16} \sqrt{n/K}. \tag{35}$$

It follows from (29), (33) and (35) that the first inequality in Assumption 3 is satisfied for $\boldsymbol{W} = \boldsymbol{W}^{(j)}$ and $\boldsymbol{\Pi} = \boldsymbol{\Pi}^{(j)} = \boldsymbol{W}^{(j)}\boldsymbol{A}$, $j = 0, 1, \ldots, T$. Next, using the first inequality in (35) and the fact that $\lambda_1(\boldsymbol{W}^{(j)}\boldsymbol{A}) \le \lambda_1(\boldsymbol{W}^{(j)})\lambda_1(\boldsymbol{A})$ yields

$$\kappa(\boldsymbol{W}^{(j)}\boldsymbol{A}) \le \kappa(\boldsymbol{W}^{(j)})\kappa(\boldsymbol{A}) \le C. \tag{36}$$

Thus, Assumption 3 is satisfied for $\boldsymbol{W} = \boldsymbol{W}^{(j)}$ and $\boldsymbol{\Pi} = \boldsymbol{\Pi}^{(j)} = \boldsymbol{W}^{(j)}\boldsymbol{A}$, $j = 0, 1, \ldots, T$. In conclusion, we have proved that $\boldsymbol{\Pi}^{(j)} \in \mathscr{M}$, $j = 0, 1, \ldots, T$.

To prove Theorem 3, we now use Theorem 2.5 in [44], according to which the lower bounds (13) and (14) hold if the following conditions are satisfied:

(a) $\mathrm{KL}(\mathbb{P}_{\boldsymbol{\Pi}^{(j)}}, \mathbb{P}_{\boldsymbol{\Pi}^{(0)}}) \le \frac{\log T}{16}$, for each $j = 1, \ldots, T$, where $\mathrm{KL}(\mathbb{P}, \mathbb{Q})$ denotes the Kullback-Leibler divergence between the probability measures $\mathbb{P}$ and $\mathbb{Q}$.

(b) For $0 \le j < \ell \le T$ we have $\min_{\mathbf{P}\in\mathscr{P}} \|\boldsymbol{W}^{(\ell)} - \boldsymbol{W}^{(j)}\mathbf{P}\|_F \ge c\sqrt{\frac{n}{N}}$ and $\min_{\mathbf{P}\in\mathscr{P}} \|\boldsymbol{W}^{(j)} - \boldsymbol{W}^{(\ell)}\mathbf{P}\|_1 \ge cn\sqrt{\frac{K}{N}}$. where $\mathscr{P}$ is the set of all permutation matrices and c is a positive constant.

(c) The maps $(\boldsymbol{M}_1, \boldsymbol{M}_2) \mapsto \min_{\mathbf{P}\in\mathscr{P}} \|\boldsymbol{M}_1 - \boldsymbol{M}_2\mathbf{P}\|_F$ and $(\boldsymbol{M}_1, \boldsymbol{M}_2) \mapsto \min_{\mathbf{P}\in\mathscr{P}} \|\boldsymbol{M}_1 - \boldsymbol{M}_2\mathbf{P}\|_1$ are semi-distances.

The rest of the proof is devoted to checking that these conditions (a) – (c) are indeed satisfied.

3. *Proof of (a).*

Our aim now is to derive an upper bound on the Kullback-Leibler divergence between $\mathbb{P}_{\boldsymbol{\Pi}^{(j)}}$ and $\mathbb{P}_{\boldsymbol{\Pi}^{(0)}}$, where

$$\boldsymbol{\Pi}^{(j)} = \boldsymbol{W}^{(j)}\boldsymbol{A} \quad \text{and} \quad \boldsymbol{\Pi}^{(0)} = \boldsymbol{W}^{(0)}\boldsymbol{A}.$$

To shorten the notation, we set

$$\alpha := \frac{N-K}{N} + \frac{K}{pN}, \qquad \beta := \frac{K}{pN}.$$

For any $1 \le i \le n$, $1 \le \ell \le p$, we have $\Pi_{i\ell}^{(0)} = \sum_{k=1}^{K} W_{ik}^{(0)} A_{k\ell}$. If $i \ge K+1$, for the entries in the ith row of matrix $\boldsymbol{\Pi}^{(0)}$ the following holds.

- For the columns ℓ such that $(\ell - 1)$ is a multiple of p/K:

  - $\Pi_{i\ell}^{(0)}$ takes once the value $\alpha + (K-1)\gamma_1(\beta - \alpha)$,
  - $\Pi_{i\ell}^{(0)}$ takes $K - 1$ times the value $\beta + \gamma_1(\alpha - \beta)$.

- For all other columns: $\Pi_{i\ell}^{(0)} \in \{\alpha, \beta\}$.

On the other hand, for any $1 \le j \le T$, by the definition of $\boldsymbol{W}^{(j)}$ in (32) we have

$$\boldsymbol{\Pi}^{(j)} = \boldsymbol{\Pi}^{(0)} + \Delta^{(j)}$$
$$= \boldsymbol{\Pi}^{(0)} + \begin{bmatrix} \mathbf{0}_{K,p} \\ \Omega^{(j)}\boldsymbol{A} \end{bmatrix}.$$

Therefore, for any $1 \le \ell \le p$, if $i \le K$, $\Pi_{i\ell}^{(j)} = \Pi_{i\ell}^{(0)}$, and if $i \ge K+1$, $\Pi_{i\ell}^{(j)} = \Pi_{i\ell}^{(0)} + \Delta_{i\ell}^{(j)}$, where

$$\Delta_{i\ell}^{(j)} = \gamma \left[\sum_{k=1}^{K/2} w_{i-K}^{(j)}(k)A_{k\ell} - \sum_{k=K/2+1}^{K} w_{i-K}^{(j)}(k - K/2)A_{k\ell} \right].$$

If $i \ge K+1$, for the entries in the ith row of matrix $\Delta^{(j)}$ the following holds.

- For the columns ℓ such that $(\ell - 1)$ is a multiple of p/K:

  - $\Delta_{i\ell}^{(j)}$ is $K/2$ times equal to $\gamma(\alpha - \beta)$,
  - $\Delta_{i\ell}^{(j)}$ is $K/2$ times equal to $-\gamma(\alpha - \beta)$.

- For all other ℓ: $\Delta_{i\ell}^{(j)} = 0$.

We are now ready to bound the Kullback-Leibler divergence between $\mathbb{P}_{\mathbf{\Pi}^{(j)}}$ and $\mathbb{P}_{\mathbf{\Pi}^{(0)}}$. Denote by $M_p(N, q)$ the multinomial distribution with parameters (N, q) where q is a probability vector in $\mathbb{R}^p$. We recall that the Kullback-Leibler divergence between two multinomial distributions $M_p(N, q_1)$ and $M_p(N, q_2)$ is equal to $N \sum_{\ell=1}^{p} q_{1\ell} \log(q_{1\ell}/q_{2\ell})$. Hence, we have

$$
\begin{aligned}
\mathrm{KL}\left(\mathbb{P}_{\mathbf{\Pi}^{(j)}}, \mathbb{P}_{\mathbf{\Pi}^{(0)}}\right) &= N \sum_{i=1}^{n} \sum_{\ell=1}^{p} \Pi_{i\ell}^{(j)} \log\left(\frac{\Pi_{i\ell}^{(j)}}{\Pi_{i\ell}^{(0)}}\right) \\
&= N \sum_{i=K+1}^{n} \sum_{\ell=1}^{p} \Pi_{i\ell}^{(j)} \log\left(\frac{\Pi_{i\ell}^{(j)}}{\Pi_{i\ell}^{(0)}}\right) \\
&= N \sum_{i=K+1}^{n} \sum_{\ell=1}^{p} \left(\Pi_{i\ell}^{(0)} + \Delta_{i\ell}^{(j)}\right) \log\left(1 + \frac{\Delta_{i\ell}^{(j)}}{\Pi_{i\ell}^{(0)}}\right) \\
&\leq N \sum_{i=K+1}^{n} \sum_{\ell=1}^{p} \left(\Delta_{i\ell}^{(j)} + \frac{(\Delta_{i\ell}^{(j)})^2}{\Pi_{i\ell}^{(0)}}\right).
\end{aligned}
$$

Note that, by construction, $\sum_{\ell=1}^{p} \Delta_{i\ell}^{(j)} = 0$. Therefore,

$$
\begin{aligned}
\mathrm{KL}\left(\mathbb{P}_{\mathbf{\Pi}^{(j)}}, \mathbb{P}_{\mathbf{\Pi}^{(0)}}\right) &\leq N \sum_{i=K+1}^{n} \sum_{\ell=1}^{p} \frac{(\Delta_{i\ell}^{(j)})^2}{\Pi_{i\ell}^{(0)}} \\
&= N \sum_{i=K+1}^{n} \sum_{(\ell-1)\text{ multiple of } p/K} \frac{(\Delta_{i\ell}^{(j)})^2}{\Pi_{i\ell}^{(0)}} \\
&\leq N \sum_{i=K+1}^{n} \sum_{(\ell-1)\text{ multiple of } p/K} \frac{\gamma^2(\alpha - \beta)^2}{\Pi_{i\ell}^{(0)}} \\
&\leq \frac{4c^*}{K} \sum_{i=K+1}^{n} \sum_{(\ell-1)\text{ multiple of } p/K} \frac{(N-K)^2}{N^2 \Pi_{i\ell}^{(0)}} \\
&\leq \frac{4c^*}{K} \sum_{i=K+1}^{n} \left[\frac{3N}{N-K} + \frac{2(K-1)KN}{N-K}\right] \\
&\leq \frac{c}{K}(n-K)\frac{K^2 N}{N-K} \\
&\leq cK(n-K) \\
&\leq \frac{\log T}{16},
\end{aligned}
$$

where we have used (31) and we have chosen c^* small enough, such that the constant c in the penultimate line does not exceed $(\log 2)/256$.

4. *Proof of (b).*

Note that for any $0 \leq j < \ell \leq T$ we have $\min_{\mathbf{P} \in \mathscr{P}} \|\boldsymbol{W}^{(\ell)} - \boldsymbol{W}^{(j)}\mathbf{P}\|_F = \|\boldsymbol{W}^{(\ell)} - \boldsymbol{W}^{(j)}\|_F$ since the first K rows are the same for matrices $W^{(\ell)}$ and $W^{(j)}$. Then,

$$
\begin{aligned}
\|\boldsymbol{W}^{(j)} - \boldsymbol{W}^{(\ell)}\|_F^2 = \|\Omega^{(j)} - \Omega^{(\ell)}\|_F^2 &= \sum_{i=1}^{n-K} \|\Omega_{i\cdot}^{(j)} - \Omega_{i\cdot}^{(\ell)}\|_2^2 \\
&= 2\gamma^2 \sum_{i=1}^{n-K} \|w_i^{(j)} - w_i^{(\ell)}\|_2^2 = 2\gamma^2 \|w^{(j)} - w^{(\ell)}\|_2^2 \\
&\geq \frac{\gamma^2}{8} K(n-K) \quad \text{(using (30))} \\
&= \frac{c_*^2}{8} \frac{N(n-K)}{(N-K)^2} \geq c\frac{n}{N} \quad \text{(since } K \leq n/2\text{)},
\end{aligned}
\tag{37}
$$

which proves (b) for the Frobenius norm. Quite analogously, for the ℓ_1-norm we get

$$
\begin{aligned}
\|\boldsymbol{W}^{(j)} - \boldsymbol{W}^{(\ell)}\|_1 = \|\Omega^{(j)} - \Omega^{(\ell)}\|_1 &= \sum_{i=1}^{n-K} \|\Omega_{i\cdot}^{(j)} - \Omega_{i\cdot}^{(\ell)}\|_1 \\
&\geq 2\gamma \|w^{(j)} - w^{(\ell)}\|_1 \\
&\geq c\,n\sqrt{\frac{K}{N}}.
\end{aligned}
$$

5. *Proof of (c).*

We now prove that the map $(\boldsymbol{M}_1, \boldsymbol{M}_2) \mapsto \min_{\mathbf{P} \in \mathscr{P}} \|\boldsymbol{M}_1 - \boldsymbol{M}_2\mathbf{P}\|_F$ satisfies the triangle inequality. For any matrices $\boldsymbol{M}_1, \boldsymbol{M}_2, \boldsymbol{M}_3$, we have

$$
\begin{aligned}
\min_{\mathbf{P} \in \mathscr{P}} \|\boldsymbol{M}_1 - \boldsymbol{M}_2\mathbf{P}\|_F &= \min_{\mathbf{P},\mathbf{P}' \in \mathscr{P}} \|\boldsymbol{M}_1\mathbf{P}' - \boldsymbol{M}_2\mathbf{P}\|_F \\
&\leq \min_{\mathbf{P},\mathbf{P}' \in \mathscr{P}} \left(\|\boldsymbol{M}_1\mathbf{P}' - \boldsymbol{M}_3\|_F + \|\boldsymbol{M}_3 - \boldsymbol{M}_2\mathbf{P}\|_F \right) \\
&= \min_{\mathbf{P}' \in \mathscr{P}} \|\boldsymbol{M}_1\mathbf{P}' - \boldsymbol{M}_3\|_F + \min_{\mathbf{P} \in \mathscr{P}} \|\boldsymbol{M}_3 - \boldsymbol{M}_2\mathbf{P}\|_F \\
&= \min_{\mathbf{P}' \in \mathscr{P}} \|\boldsymbol{M}_1 - \boldsymbol{M}_3\mathbf{P}'\|_F + \min_{\mathbf{P} \in \mathscr{P}} \|\boldsymbol{M}_3 - \boldsymbol{M}_2\mathbf{P}\|_F.
\end{aligned}
$$

The same calculation holds with the ℓ_1-norm in place of the Frobenius norm. This completes the proof of Theorem 3.

C Auxiliary lemmas

Lemma 6. *Let Assumption 1 be satisfied. Then,*

$$
\lambda_K(\boldsymbol{W}) \geq 1.
\tag{38}
$$

If, in addition, $\lambda_K(\boldsymbol{\Pi}) > 0$ then the matrix $\mathbf{U}$ of left singular vectors of $\boldsymbol{\Pi}$ can be represented in the form (5), where $\mathbf{H}$ is a rank K matrix with singular values

$$
\lambda_1(\mathbf{H}) = \frac{1}{\lambda_K(\boldsymbol{W})}, \qquad \lambda_{\min}(\mathbf{H}) = \lambda_K(\mathbf{H}) = \frac{1}{\lambda_1(\boldsymbol{W})},
\tag{39}
$$

and the condition number satisfying

$$
\kappa(\mathbf{H}) = \kappa(\boldsymbol{W}).
\tag{40}
$$

Proof. Let $J^* \subseteq \{1, \ldots, n\}$ be the set of K row indices of $\boldsymbol{W}$ corresponding to anchor documents. By Assumption 1 we have $\boldsymbol{W}_{J^*} = \boldsymbol{I}_K$. Hence,

$$\lambda_K(\boldsymbol{W}) = \min_{\|a\|_2 = 1} \|\boldsymbol{W}a\|_2 \geq \min_{\|a\|_2 = 1} \|\boldsymbol{W}_{J^*} a\|_2 = 1,$$

which proves (38). Next, if $\lambda_K(\boldsymbol{\Pi}) > 0$ then matrix $\mathbf{L}$ is positive definite and we define $\mathbf{H} := \boldsymbol{A}\mathbf{V}\mathbf{L}^{-1}$. In view of (4) we have $\boldsymbol{W}\mathbf{H} = \boldsymbol{\Pi}\mathbf{V}\mathbf{L}^{-1} = \mathbf{U}$, which yields (5). We now prove that $\mathbf{H}$ is non-degenerate. Indeed, (38) implies that matrix $\boldsymbol{W}^{\mathrm{T}}\boldsymbol{W} \in \mathbb{R}^{K \times K}$ is positive definite, so that $\mathbf{H} = (\boldsymbol{W}^{\mathrm{T}}\boldsymbol{W})^{-1}\boldsymbol{W}^{\mathrm{T}}\mathbf{U}$. Then for the minimal singular value $\lambda_{\min}(\mathbf{H})$ of matrix $\mathbf{H}$ we have

$$\lambda_{\min}(\mathbf{H}) = \min_{\|a\|_2 = 1} \|(\boldsymbol{W}^{\mathrm{T}}\boldsymbol{W})^{-1}\boldsymbol{W}^{\mathrm{T}}\mathbf{U}a\|_2$$

$$\geq \min_{x \in \mathbb{R}^n : \|x\|_2 = 1} \|(\boldsymbol{W}^{\mathrm{T}}\boldsymbol{W})^{-1}\boldsymbol{W}^{\mathrm{T}}x\|_2 = \frac{1}{\lambda_1(\boldsymbol{W})} > 0.$$

Thus, $\mathbf{H}$ is non-degenerate and we can write $\boldsymbol{W} = \mathbf{U}\mathbf{H}^{-1}$ implying (39). Equality (40) is an immediate consequence of (39). $\qquad\square$

Lemma 7. *Let $\boldsymbol{W}, \boldsymbol{A}$ and $\boldsymbol{\Pi} = \boldsymbol{W}\boldsymbol{A}$ be matrices with non-negative entries satisfying (2). Then the singular values of matrices $\boldsymbol{W}$ and $\boldsymbol{\Pi}$ satisfy the inequalities*

$$\lambda_K(\boldsymbol{\Pi}) \leq \sqrt{n/K}, \tag{41}$$

$$\sqrt{n/K} \leq \lambda_1(\boldsymbol{W}) \leq \sqrt{n}, \tag{42}$$

$$\lambda_1(\boldsymbol{\Pi}) \leq \sqrt{K}\lambda_1(\boldsymbol{W}). \tag{43}$$

Proof. Inequality (41) follows from the fact that

$$K\lambda_K^2(\boldsymbol{\Pi}) \leq \lambda_1^2(\boldsymbol{\Pi}) + \cdots + \lambda_K^2(\boldsymbol{\Pi}) = \|\boldsymbol{\Pi}\|_F^2 \leq n.$$

Next, using (2) we obtain

$$\lambda_1(\boldsymbol{W}) \leq \|\boldsymbol{W}\|_F = \sqrt{\sum_{i=1}^{n}\sum_{k=1}^{K} W_{ik}^2} \leq \sqrt{\sum_{i=1}^{n}\sum_{k=1}^{K} W_{ik}} = \sqrt{n}. \tag{44}$$

On the other hand, for $a = (1/\sqrt{K}, \ldots, 1/\sqrt{K})^{\mathrm{T}} \in \mathbb{R}^K$ we have

$$\lambda_1(\boldsymbol{W}) \geq \|\boldsymbol{W}a\|_2 = \sqrt{\sum_{i=1}^{n} \frac{1}{K}\left(\sum_{k=1}^{K} W_{ik}\right)^2} = \sqrt{\frac{n}{K}}.$$

Quite similarly to (44), using (2) we get $\|\boldsymbol{A}\| = \lambda_1(\boldsymbol{A}) \leq \|\boldsymbol{A}\|_F \leq \sqrt{K}$, which implies (43):

$$\|\boldsymbol{\Pi}\| = \|\boldsymbol{W}\boldsymbol{A}\| \leq \|\boldsymbol{W}\|\|\boldsymbol{A}\| \leq \sqrt{K}\lambda_1(\boldsymbol{W}).$$

$\qquad\square$

Lemma 8. *Let $K \leq \min(n, p)$. For any two matrices $\boldsymbol{W} \in \mathbb{R}^{n \times K}$ and $\boldsymbol{A} \in \mathbb{R}^{K \times p}$ we have*

$$\lambda_K(\boldsymbol{W}\boldsymbol{A}) \geq \lambda_K(\boldsymbol{W})\lambda_K(\boldsymbol{A}). \tag{45}$$

29

Proof. We consider only the case $\lambda_K(\boldsymbol{A}) > 0$ since otherwise (45) is trivial. By Courant-Fischer min-max formula (see, e.g., [19, Theorem C.3]) we have

$$\lambda_K(\boldsymbol{W}\boldsymbol{A}) = \max_{S:\dim(S)=K} \min_{y\in S\setminus\{0\}} \frac{\|\boldsymbol{W}\boldsymbol{A}y\|_2}{\|y\|_2},$$

where the maximum is taken over all linear spans S of K vectors in $\mathbb{R}^p$. Since $\lambda_K(\boldsymbol{A}) > 0$ and $\boldsymbol{A}y \in \mathbb{R}^K$ we can write

$$
\begin{aligned}
\lambda_K(\boldsymbol{W}\boldsymbol{A}) &= \max_{S:\dim(S)=K} \min_{y\in S\setminus\{0\}} \frac{\|\boldsymbol{W}\boldsymbol{A}y\|_2}{\|\boldsymbol{A}y\|_2} \frac{\|\boldsymbol{A}y\|_2}{\|y\|_2} \\
&\geq \min_{x\in\mathbb{R}^K\setminus\{0\}} \frac{\|\boldsymbol{W}x\|_2}{\|x\|_2} \max_{S:\dim(S)=K} \min_{y\in S\setminus\{0\}} \frac{\|\boldsymbol{A}y\|_2}{\|y\|_2} \\
&= \lambda_K(\boldsymbol{W})\lambda_K(\boldsymbol{A}).
\end{aligned}
$$

$\square$

Lemma 9. *Let $\boldsymbol{A}^0$ be a matrix with the following block structure:*

$$\boldsymbol{A}^0 = \left[\alpha_1\mathbf{e}_1,\ldots,\alpha_K\mathbf{e}_K,\underbrace{0_K,\ldots,0_K}_{p-K}\right] \in \mathbb{R}^{K\times p},$$

where $(\mathbf{e}_1,\ldots,\mathbf{e}_K)$ is the canonical basis of $\mathbb{R}^K$, $\alpha_i \in (0,1)$ and $0_K \in \mathbb{R}^K$ is the vector with all entries 0. Let

$$\boldsymbol{A} = \mathbf{P}_1(\boldsymbol{A}^0 + \boldsymbol{A}^1)\mathbf{P}_2 \quad and \quad \boldsymbol{\Pi} = \boldsymbol{W}\boldsymbol{A}$$

where $\mathbf{P}_1, \mathbf{P}_2$ are permutation matrices, $\|\boldsymbol{A}^1\| \leq \beta$, and $\boldsymbol{W} \in \mathbb{R}^{n\times K}$. If $\min_{1\leq i\leq K}\alpha_i - \beta \geq C$ then $\lambda_K(\boldsymbol{\Pi}) \geq C\lambda_K(\boldsymbol{W})$.

Proof. Matrix $\boldsymbol{A}^0$ has K top non-zero singular values $\alpha_1,\ldots,\alpha_K$. Using Weyl's inequality (see, e.g., [19, Theorem C.6]) we get

$$\lambda_K(\boldsymbol{A}) = \lambda_K(\boldsymbol{A}^0 + \boldsymbol{A}^1) \geq \min_{1\leq i\leq K}\alpha_i - \beta \geq C.$$

Combining this inequality with (45) yields the result. $\square$

References

[1] Anima Anandkumar, Dean P Foster, Daniel J Hsu, Sham M Kakade, and Yi-Kai Liu. A spectral algorithm for latent Dirichlet allocation. In Advances in Neural Information Processing Systems, pages 917–925, 2012.

[2] Animashree Anandkumar, Rong Ge, Daniel Hsu, Sham M Kakade, and Matus Telgarsky. Tensor decompositions for learning latent variable models. Journal of Machine Learning Research, 15:2773–2832, 2014.

[3] Mario Cesar Ugulino Araujo, Teresa Cristina Bezerra Saldanha, Roberto Kawakami Harrop Galvao, Takashi Yoneyama, Henrique Caldas Chame, and Valeria Visani. The successive projections algorithm for variable selection in spectroscopic multicomponent analysis. Chemometrics and Intelligent Laboratory Systems, 57(2):65 – 73, 2001.

[4] Sanjeev Arora, Rong Ge, Yonatan Halpern, David Mimno, Ankur Moitra, David Sontag, Yichen Wu, and Michael Zhu. A practical algorithm for topic modeling with provable guarantees. In International Conference on Machine Learning, pages 280–288, 2013.

[5] Sanjeev Arora, Rong Ge, and Ankur Moitra. Learning topic models–going beyond svd. In 2012 IEEE 53rd Annual Symposium on Foundations of Computer Science, pages 1–10. IEEE, 2012.

[6] Trapit Bansal, Chiranjib Bhattacharyya, and Ravindran Kannan. A provable svd-based algorithm for learning topics in dominant admixture corpus. In Advances in Neural Information Processing Systems, pages 1997–2005, 2014.

[7] Manuele Bicego, Pietro Lovato, Alessandro Perina, Marianna Fasoli, Massimo Delledonne, Mario Pezzotti, Annalisa Polverari, and Vittorio Murino. Investigating topic models' capabilities in expression microarray data classification. IEEE/ACM transactions on computational biology and bioinformatics, 9(6):1831–1836, 2012.

[8] Xin Bing, Florentina Bunea, and Marten Wegkamp. A fast algorithm with minimax optimal guarantees for topic models with an unknown number of topics. Bernoulli 26(3): 1765-1796, 2020.

[9] David M Blei and John D Lafferty. Dynamic topic models. In Proceedings of the 23rd international conference on Machine learning, pages 113–120, 2006.

[10] David M Blei and John D Lafferty. A correlated topic model of science. The Annals of Applied Statistics, 1(1):17–35, 2007.

[11] David M Blei, Andrew Y Ng, and Michael I Jordan. Latent Dirichlet allocation. Journal of machine Learning research, 3(Jan):993–1022, 2003.

[12] Jen-Tzung Chien and Chuang-Hua Chueh. Dirichlet class language models for speech recognition. IEEE Transactions on Audio, Speech, and Language Processing, 19(3):482–495, 2010.

[13] Andrzej Cichocki, Rafal Zdunek, Anh Huy Phan, and Shun ichi Amari. Nonnegative Matrix and Tensor Factorizations - Applications to Exploratory Multi-way Data Analysis and Blind Source Separation. Wiley, 2009.

[14] Stephan A Curiskis, Barry Drake, Thomas R Osborn, and Paul J Kennedy. An evaluation of document clustering and topic modelling in two online social networks: Twitter and reddit. Information Processing & Management, 57(2):102034, 2020.

[15] Weicong Ding, Mohammad Hossein Rohban, Prakash Ishwar, and Venkatesh Saligrama. Topic discovery through data dependent and random projections. In Sanjoy Dasgupta and David McAllester, editors, Proceedings of the 30th International Conference on Machine Learning, volume 28(3) of Proceedings of Machine Learning Research, pages 1202–1210, Atlanta, Georgia, USA, 17–19 Jun 2013. PMLR.

[16] David Donoho and Victoria Stodden. When does non-negative matrix factorization give a correct decomposition into parts? In S. Thrun, L. Saul, and B. Schölkopf, editors, Advances in Neural Information Processing Systems, volume 16. MIT Press, 2004.

[17] N. Gillis and S. A. Vavasis. Fast and robust recursive algorithmsfor separable nonnegative matrix factorization. IEEE Transactions on Pattern Analysis and Machine Intelligence, 36(4):698–714, April 2014.

[18] Nicolas Gillis and Stephen A. Vavasis. Semidefinite programming based preconditioning for more robust near-separable nonnegative matrix factorization. SIAM Journal on Optimization, 25(1):677–698, 2015.

[19] Christophe Giraud. Introduction to high-dimensional statistics. Chapman and Hall, 2015.

[20] Donna Harman. Overview of the first trec conference. In Proceedings of the 16th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR '93, page 36–47, New York, NY, USA, 1993. Association for Computing Machinery.

[21] Thomas Hofmann. Probabilistic latent semantic indexing. In Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval, pages 50–57, 1999.

[22] Zheng Tracy Ke and Minzhe Wang. A new svd approach to optimal topic estimation. arXiv preprint arXiv:1704.07016, 2017.

[23] John D Lafferty and David M Blei. Correlated topic models. In Advances in neural information processing systems, pages 147–154, 2006.

[24] Daniel D. Lee and H. Sebastian Seung. Learning the parts of objects by nonnegative matrix factorization. Nature, 401:788–791, 1999.

[25] Jing Lei and Alessandro Rinaldo. Consistency of spectral clustering in stochastic block models. Ann. Statist., 43(1):215–237, 02 2015.

[26] Li-Jia Li, Chong Wang, Yongwhan Lim, David M Blei, and Li Fei-Fei. Building and using a semantivisual image hierarchy. In 2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, pages 3336–3343. IEEE, 2010.

[27] Wei Li and Andrew McCallum. Pachinko allocation: Dag-structured mixture models of topic correlations. In Proceedings of the 23rd international conference on Machine learning, pages 577–584, 2006.

[28] Xueyu Mao, Purnamrita Sarkar, and Deepayan Chakrabarti. Overlapping clustering models, and one (class) svm to bind them all. In Advances in Neural Information Processing Systems, pages 2126–2136, 2018.

[29] Xueyu Mao, Purnamrita Sarkar, and Deepayan Chakrabarti. Estimating mixed memberships with sharp eigenvector deviations. Journal of the American Statistical Association, 2020.

[30] Andrew McCallum, Andrés Corrada-Emmanuel, and Xuerui Wang. The author-recipient-topic model for topic and role discovery in social networks: Experiments with enron and academic email. Computer Science Department Faculty Publication Series, page 44, 2005.

[31] Tomohiko Mizutani. Ellipsoidal rounding for nonnegative matrix factorization under noisy separability. Journal of Machine Learning Research, 15:1011–1039, 2014.

[32] Tomohiko Mizutani. Robustness analysis of preconditioned successive projection algorithm for general form of separable NMF problem. Linear Algebra and its Applications, 497:1 – 22, 2016.

[33] Biagio Palese and A Usai. The relative importance of service quality dimensions in e-commerce experiences. International Journal of Information Management, 40:132–140, 2018.

[34] Maxim Panov, Konstantin Slavnov, and Roman Ushakov. Consistent estimation of mixed memberships with successive projections. In International Conference on Complex Networks and their Applications, pages 53–64. Springer, 2017.

[35] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. Journal of Machine Learning Research, 12:2825–2830, 2011.

[36] V. Perrone, P. A. Jenkins, D. Spano, and Y. W. Teh. Poisson random fields for dynamic feature models. Journal of Machine Learning Research (JMLR), December 2017.

[37] Ian Porteous, David Newman, Alexander Ihler, Arthur Asuncion, Padhraic Smyth, and Max Welling. Fast collapsed gibbs sampling for latent Dirichlet allocation. In Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining, pages 569–577, 2008.

[38] Jonathan K Pritchard, Matthew Stephens, and Peter Donnelly. Inference of population structure using multilocus genotype data. Genetics, 155(2):945–959, 2000.

[39] Daniel Ramage, David Hall, Ramesh Nallapati, and Christopher D Manning. Labeled lda: A supervised topic model for credit attribution in multi-labeled corpora. In Proceedings of the 2009 conference on empirical methods in natural language processing, pages 248–256, 2009.

[40] Ben Recht, Christopher Re, Joel Tropp, and Victor Bittorf. Factoring nonnegative matrices with linear programs. In F. Pereira, C. J. C. Burges, L. Bottou, and K. Q. Weinberger, editors, Advances in Neural Information Processing Systems, volume 25. Curran Associates, Inc., 2012.

[41] Julia Silge and David Robinson. Text mining with r: A tidy approach, 2020.

[42] Yee W Teh, Michael I Jordan, Matthew J Beal, and David M Blei. Sharing clusters among related groups: Hierarchical Dirichlet processes. In Advances in neural information processing systems, pages 1385–1392, 2005.

[43] Joel A. Tropp. An introduction to matrix concentration inequalities. Foundations and Trends in Machine Learning, 8(1–2):1–230, 2015.

[44] Alexandre B Tsybakov. Introduction to nonparametric estimation. Springer, 2008.

[45] Hanna M Wallach. Topic modeling: beyond bag-of-words. In Proceedings of the 23rd international conference on Machine learning, pages 977–984, 2006.

[46] Hui Yuan, Wei Xu, Qian Li, and Raymond Lau. Topic sentiment mining for sales performance prediction in e-commerce. Annals of Operations Research, 270(1-2):553–576, 2018.

[47] Ke Zhai, Jordan Boyd-Graber, Nima Asadi, and Mohamad L Alkhouja. Mr. lda: A flexible large scale topic modeling package using variational inference in mapreduce. In Proceedings of the 21st international conference on World Wide Web, pages 879–888, 2012.

[48] Qiqi Zhu, Yanfei Zhong, Liangpei Zhang, and Deren Li. Scene classification based on the fully sparse semantic topic model. IEEE Transactions on Geoscience and Remote Sensing, 55(10):5525–5538, 2017.

Appendix A Additional Experiments: Estimation of topic-word matrix

In this section, we investigate the SPOC estimator of topic-word matrix A using the sequence of experiments on synthetic data similar to those of Section 6. Figures 6-9 below present the results of simulations with different values of parameters n, p, N and the number of topics K. The generation of matrices W and A was performed in the same way as in Section 6. For each value on the x-axes of the figures, we present the averaged result over 10 simulations.

Our objective was to assess the effect of each of parameters n, p, N, K on the Frobenius error between A and the estimator $\hat{A}$ derived from SPOC algorithm via (9). For comparison, we provide the same simulation
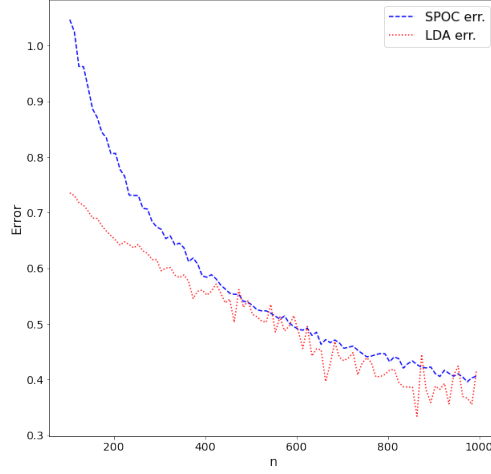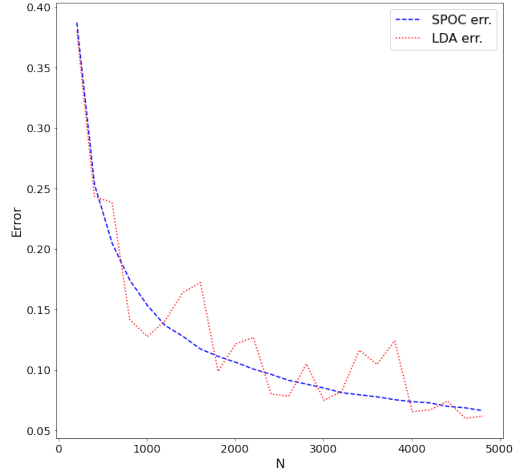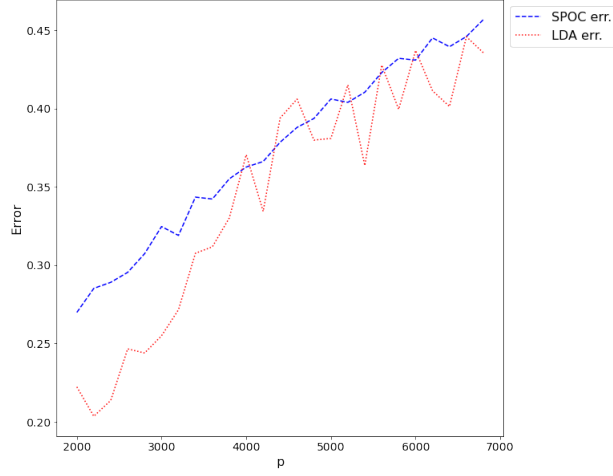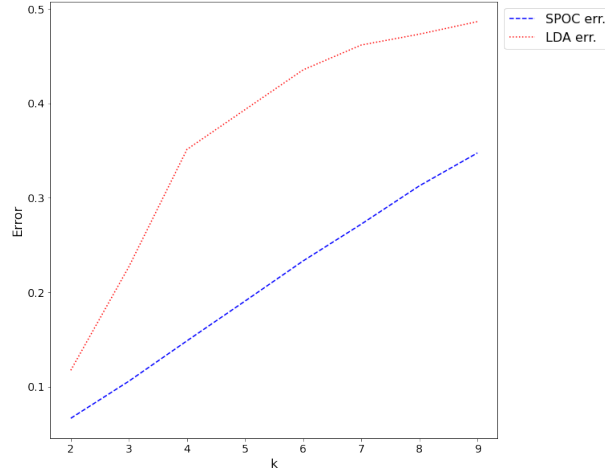
Figure 6: The n-dependency of the Frobenius error of $\hat{\boldsymbol{A}}$ using SPOC and LDA algorithms. Total number of words $p = 5000$, number of sampled words in each document $N = 200$.

study for the LDA estimator of $\boldsymbol{A}$. The experiments show that the SPOC estimator is very competitive with LDA while being more stable. One may also notice that the Frobenius error of both SPOC and LDA estimators has a behavior similar to the optimal rates for estimation of matrix $\boldsymbol{A}$ under ℓ_1-error derived in [8, 22]. Indeed, the estimation error is decreasing as function of n and N and it is increasing as function of p and K.

Figure 7: The N-dependency of the Frobenius error of $\hat{\boldsymbol{A}}$ using SPOC and LDA algorithms. Total number of words $p = 5000$, number of documents $n = 1000$.

34

Figure 8: The p-dependency of the Frobenius error of $\hat{\boldsymbol{A}}$ using SPOC and LDA algorithms. Total number of words $n = 1000$, number of sampled words in each document $N = 200$.

Figure 9: The K-dependency of the Frobenius error of $\hat{\boldsymbol{A}}$ using SPOC and LDA algorithms. Total number of words $p = 5000$, number of sampled words in each document $N = 5000$, number of documents $n = 1000$.

Appendix B Additional Experiments: Empirical study of singular values of word-document and topic-document matrices

Most conditions and assumptions used throughout the paper are satisfied for fairly general choices of parameters. However, Assumption 3 enforces certain bounds on matrices W and Π which might seem restrictive. The goal of this section is to experimentally show that singular values and quotients appearing in Assumption 3 admit reasonably small upper bounds.

We consider matrices W, Π and A generated in the following way. In most experiments we take $K = 3$ and the matrix W has the following structure: K rows of W are canonical basis vectors, each of the remaining $n - K$ rows is generated independently using the Dirichlet distribution with parameter $\alpha = (0.1, 0.15, 0.2)$. In the experiments where K must vary, we define W in a different way. Namely, for the $n - K$ rows that are not canonical basis vectors, each element W_{kj} is generated from the uniform distribution on $[0, 1]$ and then each row of the matrix is normalized to have $\sum_{k=1}^{K} W_{ik} = 1$. For the matrix A, we take K columns proportional to canonical basis vectors with coefficients equal to random variables $U_k, k = 1, \dots, K$ uniformly distributed on $[0, 1]$. The elements A_{kj} of matrix A in the remaining $p - K$ columns are obtained by generating numbers from the uniform distribution on $[0, 1]$ and then normalizing each row of the matrix to have $\sum_{j=K+1}^{p} A_{kj} = 1 - U_k, k = 1, \dots, K$. The resulting matrix A has normalized rows such that $\sum_{j=1}^{p} A_{kj} = 1$.

We essentially use the same parameters as in the experiments reported in Section 6. The dependencies of the condition numbers $\kappa(\Pi)$ and $\kappa(W)$ on parameters n, p and K are presented on Figures 10 and 11. All the condition numbers have small to moderate values for a quite wide range of parameters n and p, while the dependence on K is stronger. Additionally, we study the ratio $\lambda_1(W)/\lambda_K(\Pi)$ also appearing in Assumption 3. As presented on Figure 12 it shows the tendencies similar to the condition numbers.

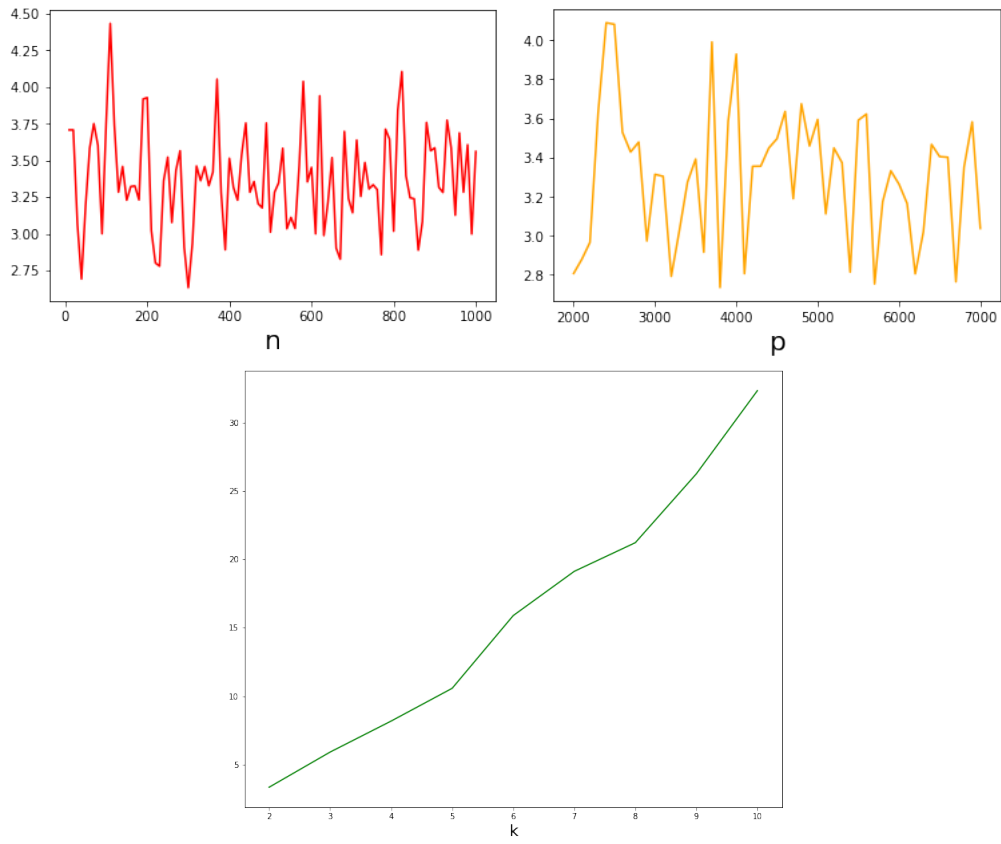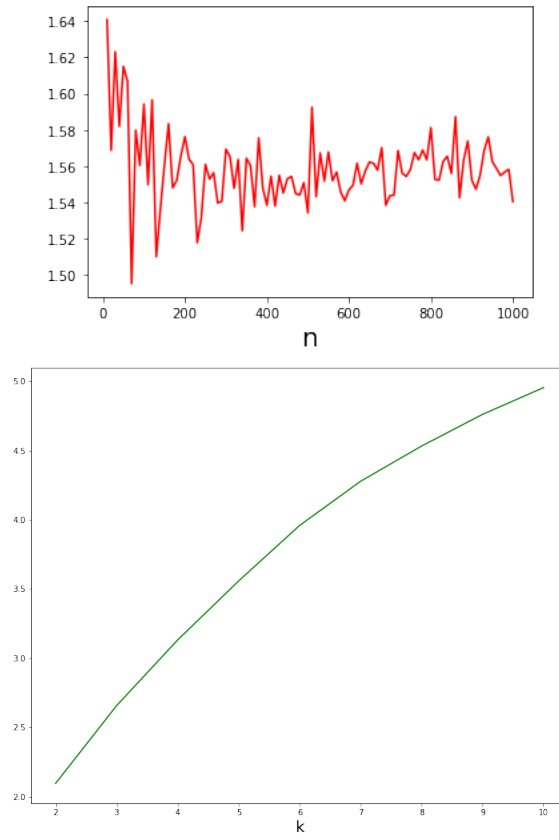Figure 10: The dependency of $\kappa(\mathbf{\Pi})$ on parameters n, p and K.

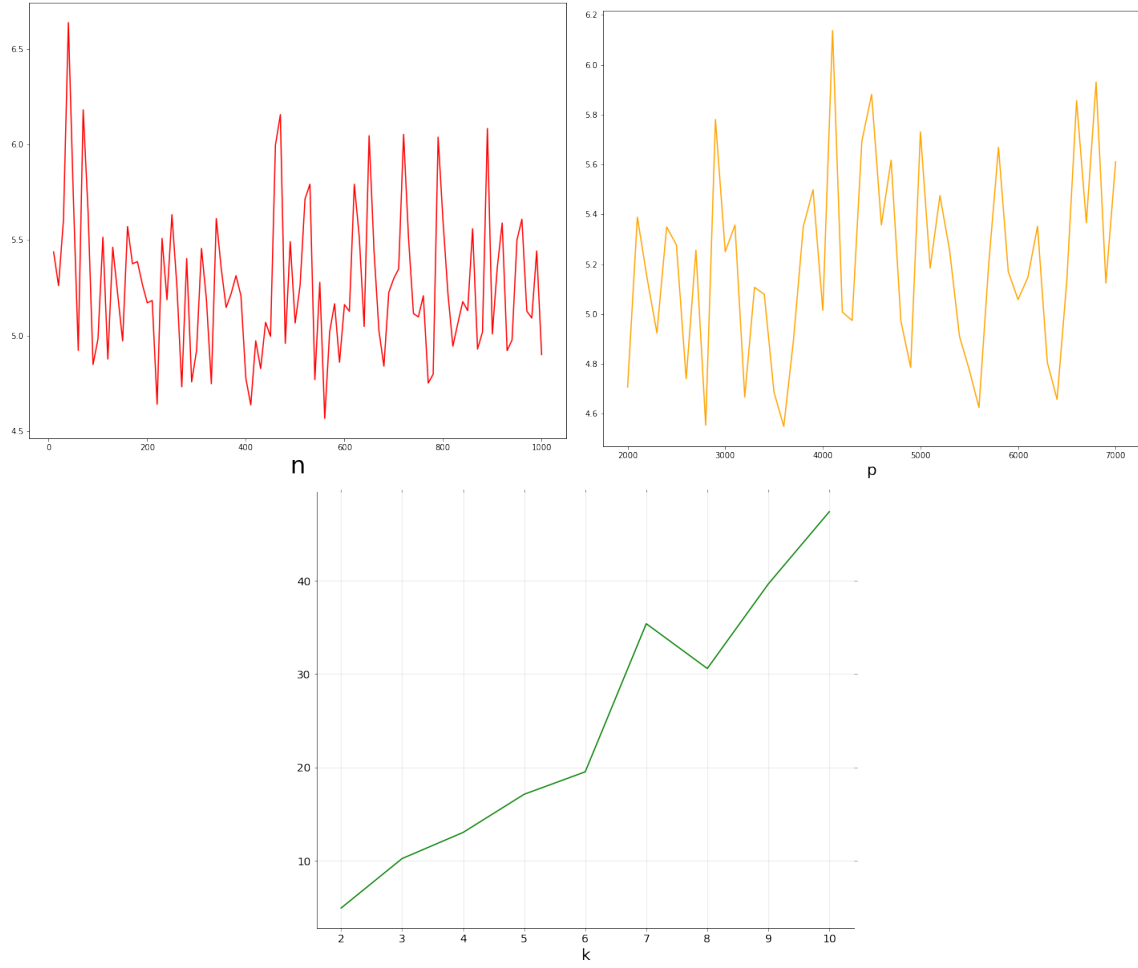Figure 11: The dependency of $\kappa(\boldsymbol{W})$ on parameters n and K.

Figure 12: The dependency of $\lambda_1(\boldsymbol{W})/\lambda_K(\boldsymbol{\Pi})$ on parameters n, p and K.