Prior-aware Composition Inference for Spectral Topic Models

Moontae Lee[1] **David Bindel**[2] **David Mimno**[2]
[1]University of Illinois at Chicago, Microsoft Research at Redmond, [2]Cornell University

Abstract

Spectral algorithms operate on matrices or tensors of word co-occurrence to learn latent topics. These approaches remove the dependence on the original documents and produce substantial gains in efficiency with provable inference, but at a cost: the models can no longer infer any information about individual documents. Thresholded Linear Inverse is developed to learn document-specific topic compositions, but its linear characteristics limit the inference quality without considering any prior information on topic distributions. We propose two novel estimation methods that respect previously unclear prior structures of spectral topic models. Experiments on a variety of synthetic to real collections demonstrate that our Prior-Aware Dual Decomposition outperforms the baseline method, whereas our Prior-Aware Manifold Iteration performs even better on short realistic data.

1 Introduction

Mixed-membership models represent collections of discrete objects in terms of **topics** and **compositions** (Hofmann, 1999; Blei et al., 2003). The topics capture underlying themes, common genres, and hidden communities across the collections as distributions over the objects (Lee et al., 2015; Liu et al., 2009), whereas the compositions represent individual collections as distributions over topics, thus allowing users to retrieve documents, playlists, and network snapshots relevant to their topic queries (Blei and Lafferty, 2007; Steyvers and Griffiths, 2008; Hall et al., 2008; Talley et al., 2011; Goldstone and Underwood, 2014; Erlin, 2017). We use the standard terms: words, documents, and topics, but topic models are applicable to various data modalities.

Proceedings of the 23[rd]International Conference on Artificial Intelligence and Statistics (AISTATS) 2020, Palermo, Italy. PMLR: Volume 108.

In recent years, spectral topic models have emerged as alternatives to probabilistic counterparts due to their transparent inference and optimality guarantees (Arora et al., 2012, 2013; Bansal et al., 2014; Lee et al., 2015; Huang et al., 2016; Anandkumar et al., 2012a,b, 2013; Wang and Zhu, 2014). Because the input to these models is purely in terms of co-occurrence between word pairs or triples, users can limit their interaction with the training documents to a single trivially-parallelizable aggregation of individual co-occurrence statistics.

But the efficiency advantage of factoring out the documents is also a weakness: we lose the ability to say anything about the documents themselves. In practice, users of spectral topic models must go back and apply traditional inference on the original training data as if these were new, held-out documents. That is, given the learned topics and individual documents, they need to infer the posterior topic compositions, with the assumption of a sparse Dirichlet prior or a more complex logistic-normal prior (Blei and Lafferty, 2007) on the topic distributions. Estimating compositions with a sparse Dirichlet prior can be NP-hard even for trivial models (Sontag and Roy, 2011). Gibbs sampling for composition inference is asymptotically unbiased, but has no provable guarantees and may require many sampler steps (Yao et al., 2009). Variational inference methods may be faster than Gibbs sampling, but they often get trapped in local minima, learning inconsistent models as the number of topics varies (Yao et al., 2018).

Thresholded Linear Inverse (TLI) is the only available method that infers document-specific topic compositions for spectral topic models (Arora et al., 2016). TLI tries to compute an unbiased and small variance approximate inverse of the word-topic matrix, and then transforms the word distribution vector of a query document into a latent topic composition vector by a single matrix-vector product, which is later thresholded for sparser estimation. Unfortunately, finding the approximate inverse is computationally expensive and numerically unstable. Thresholding parameters are not intuitive and data-specific as well. Above all, linear characteristics of TLI neglects correlations between the topics encoded in the prior, working poorly on realistic

documents. Though users could further improve the inference quality by gradient updates with respect to MLE/MAP objectives, such post-processing is limited to specific parametric distributions at additional costs.

Topics are often strongly correlated in reality. Spectral topic models can learn topic correlations (Arabshahi and Anandkumar, 2017; Huang et al., 2016), and the second-order models based on word-word co-occurrence do not limit their prior correlations to a specific parametric family (Lee et al., 2015). By studying previously unclear spectral structures of the prior, we propose two novel composition inference methods that leverage the learned correlations as well as the learned topics. In **Prior-Aware Dual Decomposition (PADD)**, each sub-problem tries to find the best composition that maximally fits the word vector of individual documents in parallel, whereas the master-problem regularizes overall compositions to be aligned with the learned prior correlations like (Komodakis et al., 2011; Rush and Collins, 2012). In contrast, the **Prior-Aware Manifold Iteration (PAMI)** extracts an invariant factor of individual compositions from the topic correlations, and then keeps improving only the varying part with respect to the manifold structures of the underlying generative process by using recent advances in Manifold Alternating Direction Methods of Multipliers (MADMM) (Kovnatsky et al., 2016; Chen et al., 2018).

We evaluate our methods on various textual and non-textual real corpora, but also on semi-synthetic and semi-real corpora generated from uncorrelated and correlated topic models trained on the real data. Our experiments show that TLI works moderately only if the query document underlie a few synthetically separated topics with little correlations. In contrast, our PADD performs competitively to benchmark Gibbs sampling in almost every setting, while PAMI outperforms PADD especially on short realistic documents. Providing theoretical and empirical rationales, we also bridge the gap between spectral composition inference and the corresponding probabilistic posterior inference.

2 Spectral Topic Modeling

We begin this section with a formal introduction to spectral topic modeling. Consider a dataset of M documents consisting of tokens drawn from a vocabulary of N words. Topic models assume that K topics are used to generate this dataset, where each topic is a distribution over the words; we summarize the latent topics by the column-stochastic matrix $\boldsymbol{B} \in \mathbb{R}^{N \times K}$ where each column $\boldsymbol{b}_k \in \Delta^{N-1}$ represents the distribution of the topic k. For each document m, choose a topic composition $\boldsymbol{w}_m \in \Delta^{K-1}$ first from a certain prior $\mathfrak{f}$; we collect these hidden compositions into another column-stochastic matrix $\boldsymbol{W} \in \mathbb{R}^{K \times M}$. These models assume that each of the n_m tokens in the document m is then generated independently from the categorical distribution given by the word-probability vector $\boldsymbol{B}\boldsymbol{w}_m \in \mathbb{R}^N$.

Different models adopt different $\mathfrak{f}$ such as $\mathfrak{f} = \mathrm{Dir}(\boldsymbol{\alpha})$ for Latent Dirichlet Allocation (LDA) (Blei et al., 2003); $\mathfrak{f} = \mathcal{LN}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ or $\mathfrak{f} = \mathcal{PN}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ for Logistic/Probit-Normal Correlated Topic Models (CTMs) (Blei and Lafferty, 2007; Yu and Fokoue, 2014). Let $\boldsymbol{H} \in \mathbb{R}^{N \times M}$ be the word-document matrix where the m-th column vector $\boldsymbol{h}_m$ counts the occurrences of each word in the document m, and let $\widetilde{\boldsymbol{H}}$ be the column-normalized version of $\boldsymbol{H}$ that specifies the relative frequencies of each word rather than the raw counts. Then topic modeling aims to learn *latent topics* $\boldsymbol{B}$ and *hidden compositions* $\boldsymbol{W}$ given the *observed collections of words* $\boldsymbol{H}$. Equivalently, we seek a non-negative matrix factorization $\widetilde{\boldsymbol{H}} \approx \boldsymbol{B}\boldsymbol{W}$ but with a prior to make individual topic compositions $\{\boldsymbol{w}_m\}$ coherent within a corpus.

Joint Stochastic Matrix Factorization The matrix $\widetilde{\boldsymbol{H}}$ of word frequencies is sparse, noisy, and often inconveniently large. Let us consider instead the word co-occurrence matrix $\boldsymbol{C} \in \mathbb{R}^{N \times N}$, where $\boldsymbol{C}_{ij}$ indicates the joint probability of observing a pair of words (i, j). Then topic modeling corresponds to a second-order non-negative matrix factorization: $\boldsymbol{C} \approx \boldsymbol{B}\boldsymbol{A}\boldsymbol{B}^T$ where the column-stochastic matrix $\boldsymbol{B} \in \mathbb{R}^{N \times K}$ represents the topics as before and the joint-stochastic matrix $\boldsymbol{A} \in \mathbb{R}^{K \times K}$ represents the *topic correlations*. If the true compositions $\boldsymbol{W}^*$ that generate the data are known, we can define the true correlations by $\boldsymbol{A}^* := \frac{1}{M}\boldsymbol{W}^*\boldsymbol{W}^{*T}$ where $\boldsymbol{A}^*_{kl}$ is the joint probability for a pair of latent topics (k, l). By forming $\boldsymbol{C}$ as an unbiased estimator of the underlying generative process, we can identify $\boldsymbol{B}$ and $\boldsymbol{A}$ close to the true topics and their correlations.[1]

It is helpful to compare the matrix-based view of JSMF to the generative view of standard topic models. For each document m, the generative view (Figure 1) begins with the topic composition $\boldsymbol{w}_m$, focusing on how to produce streams of tokens. We keep choosing a topic z from $\boldsymbol{w}_m$ and then a word x from $\boldsymbol{b}_z$ for each of the n_m positions. The correlations between words that $\boldsymbol{w}_m \sim \mathfrak{f}$ induces are not explicitly modeled. In contrast, the matrix-based view (Figure 2) starts with *individual topic correlations* $\boldsymbol{A}_m$ for each document m. Then for each of the possible $n_m(n_m - 1)$ position pairs, *a pair of topics* (z_1, z_2) is selected first from $\boldsymbol{A}_m$, then *a pair of words* (x_1, x_2) is chosen according to the topics $(\boldsymbol{b}_{z_1}, \boldsymbol{b}_{z_2})$, respectively. The word co-occurrence matrix explicitly captures the resulting correlations induced by the *prior topic correlations* $\boldsymbol{A}$. This pair generation view has the following two important implications:

[1] As the number of documents M grows, $\boldsymbol{A}$ converges to the true $\boldsymbol{A}^*$ and the prior $\mathbb{E}_{\boldsymbol{w} \sim \mathfrak{f}}[\boldsymbol{w}\boldsymbol{w}^T]$ (Arora et al., 2012).

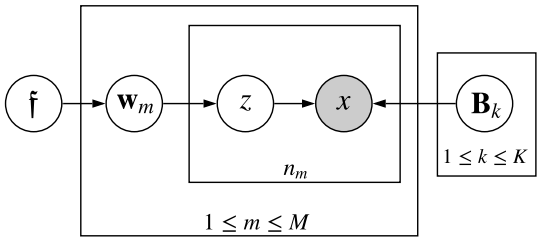

Figure 1: LDA/CTMs assert a topic composition $\boldsymbol{w}_m$ for each document m. $\mathfrak{f} = \mathrm{Dir}(\boldsymbol{\alpha})$, $\mathcal{LN}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, or $\mathcal{PN}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ provides a parametric prior for the entire corpus.

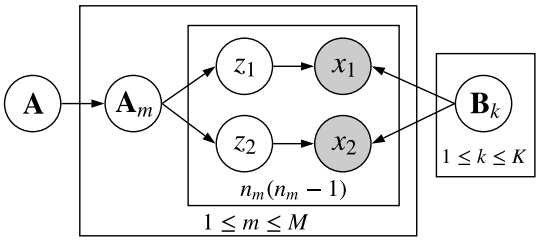

Figure 2: JSMF asserts a joint-stochastic matrix $\boldsymbol{A}_m$ to specify the correlations between two topics for each document m. $\boldsymbol{A}$ serves as a generic prior for the entire corpus.

- Topic correlation matrix $\boldsymbol{A}$ represents a flexible prior $\mathfrak{f}$ that does not specify any parametric family.
- $\boldsymbol{A}_m$ is a rank-1 joint-stochastic matrix $\boldsymbol{w}_m \boldsymbol{w}_m^T$ with $\boldsymbol{w}_m \sim \mathfrak{f}$, providing a fully generative process.

Recall that sharing the prior $\mathfrak{f}$ for $\{\boldsymbol{w}_m\} \sim \mathfrak{f}$ is the crux of modern topic modeling (Asuncion et al., 2009), and our flexible matrix prior $\boldsymbol{A}$ takes the role of $\mathfrak{f}$ for JSMF.

3 Document-specific Inference

Probabilistic topic models infer the latent topics $\boldsymbol{B}$ and the hidden compositions $\boldsymbol{W}$ together given the observations $\boldsymbol{H}$. In contrast, spectral topic models recover the topics $\boldsymbol{B}$ and their correlations $\boldsymbol{A}$ given the observations $\boldsymbol{C}$. Thus it is natural to formulate learning of the composition $\boldsymbol{W}$ as an estimation problem given the fixed $\boldsymbol{B}$, $\boldsymbol{A}$, and $\boldsymbol{H}$. Beside running likelihood-based inference on the original documents as if they are unseen data, Thresholded Linear Inverse (TLI) is the only algorithm we are aware of that has been designed for composition inference in this setting (Arora et al., 2016). We begin this section with a description of TLI, then introduce our new prior-aware algorithms.

Since each document m is generated by n_m multinomial choices, $\boldsymbol{h}_m \sim \mathrm{Mult}(n_m, \boldsymbol{B}\boldsymbol{w}_m)$, the frequency vector $\widetilde{\boldsymbol{h}}_m = \boldsymbol{h}_m \,/\, n_m$ satisfies the conditional expectation $\mathbb{E}_{\boldsymbol{w}_m}[\widetilde{\boldsymbol{h}}_m] = \boldsymbol{B}\boldsymbol{w}_m$. If $\boldsymbol{B}$ is full rank, there exist many left inverses $\boldsymbol{B}^\dagger$ (i.e., matrices such that $\boldsymbol{B}^\dagger \boldsymbol{B} = \boldsymbol{I}_K$); and for any left inverse, $\mathbb{E}_{\boldsymbol{w}_m}[\boldsymbol{B}^\dagger \widetilde{\boldsymbol{h}}_m] = \boldsymbol{w}_m$. However, not every left-inverse is equivalent; for example, large entries of $\boldsymbol{B}^\dagger$ increase the variance of the estimator. TLI chooses an *approximate* left inverse $\boldsymbol{B}^\dagger$ that controls the variance by minimizing its largest entry $\|\boldsymbol{B}^\dagger\|_\infty$ under the small bias constraint $\|\boldsymbol{B}^\dagger \boldsymbol{B} - \boldsymbol{I}_K\|_\infty \le \delta$.

Let $\lambda_\delta(\boldsymbol{B})$ denote the optimal value $\|\boldsymbol{B}^\dagger\|_\infty$ of the TLI; then one can bound the maximum violation $\|\boldsymbol{B}^\dagger \widetilde{\boldsymbol{h}}_m - \boldsymbol{w}_m\|_\infty$ by $\delta + 2\lambda_\delta(\boldsymbol{B})\sqrt{(\log K)/n_m}$ for an arbitrary prior $\mathfrak{f}$ from which $\boldsymbol{w}_m \sim \mathfrak{f}$ (Arora et al., 2016). Thus the TLI algorithm first computes the best approximate left-inverse $\boldsymbol{B}^\dagger$ of $\boldsymbol{B}$ for a given δ and makes an initial prediction $\boldsymbol{W} = \boldsymbol{B}^\dagger \widetilde{\boldsymbol{H}}$. Then for every column $\boldsymbol{w}_m$ of $\boldsymbol{W}$, it removes unlikely topics whose probability masses are smaller than a certain threshold: $\tau = \delta + 2\lambda_\delta(\boldsymbol{B})\sqrt{(\log K)/n_m}$. Despite the provable guarantees, TLI becomes inaccurate in the presence of correlated topics due to its linear estimation by a single multiplication. Typically topics in a corpus correlate each other unless its documents consist only of few topics. In addition, since the algorithm provides no guidance as to the optimal bias/variance trade-off, users must evaluate various $\boldsymbol{B}^\dagger$ with different δ parameters. However we observe that this optimization is numerically unstable, often yielding many NaN entries. Above all, TLI never uses the learned $\boldsymbol{A}$, the prior correlations between the topics in latent compositions.

3.1 PADD: Prior-Aware Dual Decomposition

Many probabilistic algorithms, including Variational inference and Gibbs sampling, differ only in the amount of smoothing for updating $\boldsymbol{B}$ and $\boldsymbol{W}$ at each step (Asuncion et al., 2009). The choice of a proper prior $\mathfrak{f}$ and its hyper-parameters is critical to provide the inductive bias needed for successful inference particularly with short input documents (Wallach et al., 2009). Second-order spectral topic models learn the prior $\mathfrak{f}$ flexibly in terms of the topics correlations $\boldsymbol{A}$ as it closely approximates the prevalent topics and their correlations encoded in the true posterior $\boldsymbol{A}^*$.[2] Thus our Prior-Aware Dual Decomposition (PADD) tries to find the composition $\boldsymbol{W} = \{\boldsymbol{w}_m\}$ that best matches the overall topic correlations (given in $\boldsymbol{A}$) as well as the learned topics (given in $\boldsymbol{B}$) and individual word observations (given in $\boldsymbol{H}$) by solving the following optimization:

$$\min \quad \sum_{m=1}^{M} \|\boldsymbol{B}\boldsymbol{w}_m - \widetilde{\boldsymbol{h}}_m\|_2^2 \qquad (1)$$

$$\text{subject to} \quad \boldsymbol{w}_m \in \Delta^{K-1} \ \text{ and } \ \frac{1}{M}\sum_{m=1}^{M} \boldsymbol{w}_m \boldsymbol{w}_m^T = \boldsymbol{A}.$$

[2] If $\mathfrak{f} = \mathrm{Dir}(\boldsymbol{\alpha})$ is the right prior, we can estimate $\boldsymbol{\alpha}$ by matching its second-moment $\mathbb{E}_{\boldsymbol{w}\sim\mathfrak{f}}[\boldsymbol{w}\boldsymbol{w}^T]$ and $\boldsymbol{A}$. PADD might not find quality compositions if both M and n_m are small, because $\boldsymbol{A}$ could be far from $\boldsymbol{A}^*$ and the prior. However, this problem is universal in probabilistic algorithms.

Algorithm 1: PADD($\boldsymbol{H}, \boldsymbol{B}, \boldsymbol{A}, \tau$)

Input: Word-document matrix $\boldsymbol{H} \in \mathbb{R}^{N \times M}$
Word-topic matrix $\boldsymbol{B} \in \mathbb{R}^{N \times K}$
Topic-topic matrix $\boldsymbol{A} \in R^{K \times K}$
Output: Topic-document matrix $\boldsymbol{W} \in \mathbb{R}^{K \times M}$
begin
  $\widetilde{\boldsymbol{H}} \leftarrow$ column-normalize($\boldsymbol{H}$) (sparse matrix)
  $(\boldsymbol{B}^2, \boldsymbol{F}) \leftarrow (\boldsymbol{B}^T\boldsymbol{B}, \boldsymbol{B}^T\widetilde{\boldsymbol{H}})$
  $(\boldsymbol{\Lambda}^{(0)}, \boldsymbol{W}^0) \leftarrow (\boldsymbol{0}^{K\times K}, \breve{\boldsymbol{B}}\widetilde{\boldsymbol{H}})$
  repeat *with* $t = 0, 1, 2, ...$
    for *each* $m \in \{1, ..., M\}$ *in parallel* **do**
      $\bar{\boldsymbol{w}}_m \leftarrow$ CCP$(\boldsymbol{B}^2, (1/M)\boldsymbol{\Lambda}^{(t)}, \boldsymbol{F}_{*m}, \boldsymbol{W}^0_{*m})$
    end
    $\bar{\boldsymbol{\Lambda}} \leftarrow \boldsymbol{\Lambda}^{(t)} - \tau(\boldsymbol{A} - \frac{1}{M}\sum_{m=1}^{M}(\bar{\boldsymbol{w}}_m\bar{\boldsymbol{w}}_m^T))$
    $\boldsymbol{\Lambda}^{(t+1)} \leftarrow \max\{\boldsymbol{0}, (\bar{\boldsymbol{\Lambda}} + \bar{\boldsymbol{\Lambda}}^T)/2\}$
  until *until* $\bar{\boldsymbol{W}}$ *converges*
  $\boldsymbol{W} \leftarrow [\bar{\boldsymbol{w}}_1|...|\bar{\boldsymbol{w}}_M]$
end

Algorithm 2: PAMI($\boldsymbol{H}, \boldsymbol{B}, \boldsymbol{A}, \rho$)

Input/Output: Same as Algorithm 1
begin
  $\widetilde{\boldsymbol{H}} \leftarrow$ column-normalize($\boldsymbol{H}$) (sparse matrix)
  $\boldsymbol{R} \leftarrow$ Cholesky-factorize($M\boldsymbol{A}$)
  $\boldsymbol{F} \leftarrow \widetilde{\boldsymbol{H}}^T\boldsymbol{B}\boldsymbol{R}^T$
  $(\boldsymbol{\Lambda}^{(0)}, \boldsymbol{Q}^{(0)}) \leftarrow (\boldsymbol{0}^{K\times K}, \widetilde{\boldsymbol{H}}^T\breve{\boldsymbol{B}}^T\boldsymbol{R}^{-1})$
  repeat *with* $t = 0, 1, 2, ...$
    $\boldsymbol{V} \leftarrow \boldsymbol{Q}^{(t)} + \boldsymbol{\Lambda}^{(t)}$
    for *each* $m \in \{1, ..., M\}$ *in parallel* **do**
      $\bar{\boldsymbol{w}}_m \leftarrow$ SCLS$(\boldsymbol{R}^{-T}, \boldsymbol{V}^T_{m*})$
    end
    $\boldsymbol{P}^{(t+1)} \leftarrow [\bar{\boldsymbol{w}}_1|...|\bar{\boldsymbol{w}}_M]^T\boldsymbol{R}^{-1}$
    $(\bar{\boldsymbol{U}}_K, \bar{\boldsymbol{V}}_K) \leftarrow svd(\boldsymbol{F} + \rho\boldsymbol{P}^{(t+1)} - \rho\boldsymbol{\Lambda}^{(t)})$
    $\boldsymbol{Q}^{(t+1)} \leftarrow \bar{\boldsymbol{U}}_K\bar{\boldsymbol{V}}_K^T$
    $\boldsymbol{\Lambda}^{(t+1)} \leftarrow \boldsymbol{\Lambda}^{(t)} + (\boldsymbol{Q}^{(t+1)} - \boldsymbol{P}^{(t+1)})$
  until *until* $\boldsymbol{Q}$ *converges*
  $\boldsymbol{W} \leftarrow \boldsymbol{R}^T\boldsymbol{Q}^T$
end

Each solution composition $\boldsymbol{w}_m$ from (1) tries to fit the observed word-probability $\widetilde{\boldsymbol{h}}_m$ for each document m (i.e., *loss minimization*), while simultaneously matching the learned topic correlations $\boldsymbol{A}$ as a whole (i.e., *regularization*). Thus PADD shares the similar intuition with the prior-based Bayesian inference, but with more flexibility by not limiting the prior to a particular family. Users of PADD can also decide the balance between the loss and the regularizer more intuitively than controlling the bias parameter δ of TLI. However, it is not easy to solve (1) due to the non-linear coupling constraint $(1/M)\sum \boldsymbol{w}_m\boldsymbol{w}_m^T = \boldsymbol{A}$. We can construct a Lagrangian[3] by adding a symmetric matrix of dual variables $\boldsymbol{\Lambda} \in \mathbb{R}^{K\times K}$, Then $\mathcal{L}(\boldsymbol{w}_1, ..., \boldsymbol{w}_M, \boldsymbol{\Lambda})$ becomes

$$\sum_{m=1}^{M} \|\boldsymbol{B}\boldsymbol{w}_m - \widetilde{\boldsymbol{h}}_m\|_2^2 \;\; + \langle \boldsymbol{\Lambda}, \left(\frac{1}{M}\sum_{m=1}^{M}\boldsymbol{w}_m\boldsymbol{w}_m^T\right) - \boldsymbol{A}\rangle_F$$
$$= \sum_{m=1}^{M}\left\{\|\boldsymbol{B}\boldsymbol{w}_m - \widetilde{\boldsymbol{h}}_m\|_2^2 + \frac{1}{M}\langle \boldsymbol{\Lambda}, \boldsymbol{w}_m\boldsymbol{w}_m^T - \boldsymbol{A}\rangle_F\right\}. \quad (2)$$

Since minimizing the Lagrangian can be decomposed into M sub-problems given a dual $\boldsymbol{\Lambda}$, we use the dual decomposition (Komodakis et al., 2011; Rush and Collins, 2012). For a fixed $\boldsymbol{\Lambda}$, each sub-problem finds the currently optimal $\bar{\boldsymbol{w}}_m \in \Delta^{K-1}$ that minimizes the m-th term in (2); then the master-problem updates the dual matrix $\boldsymbol{\Lambda}$ based on its subgradient: $-\frac{1}{M}(\sum_{m=1}^{M}(\bar{\boldsymbol{w}}_m\bar{\boldsymbol{w}}_m^T - \boldsymbol{A})) \in \partial(\boldsymbol{\Lambda})$. To maintain the dual feasibility, we project the updated $\bar{\boldsymbol{\Lambda}}$ to the set of symmetric non-negative matrices before redistributing it back to the sub-problems as given in Algorithm 1.

Each sub-problem involves a non-convex quadratic programming as the quadratic coefficient $\boldsymbol{B}^T\boldsymbol{B} + (1/M)\boldsymbol{\Lambda}$ could be indefinite. By finding the smallest negative eigenvalue λ of $(1/M)\boldsymbol{\Lambda}$, we split $\boldsymbol{\Lambda}$ into the sum of the positive semidefinite $(1/M)\boldsymbol{\Lambda} - \lambda\boldsymbol{I}_K$ and the negative semidefinite $\lambda\boldsymbol{I}_K$, reformulating the objective function of each sub-problem as a difference of the two convex functions: $g(\boldsymbol{w}) = \boldsymbol{w}^T(\boldsymbol{B}^T\boldsymbol{B} + (1/M)\boldsymbol{\Lambda} - \lambda\boldsymbol{I}_K)\boldsymbol{w} - 2\widetilde{\boldsymbol{h}}_m^T\boldsymbol{B}\boldsymbol{w}$ and $h(\boldsymbol{w}) = -\lambda\boldsymbol{w}^T\boldsymbol{w}$ with the convex constraint $\boldsymbol{w} \in \Delta^{K-1}$. We adopt the Convex-Concave Programming (CCP) that quickly finds a quality minimizer by approximating $h(\boldsymbol{w})$ at each point $\boldsymbol{w}^{(t)}$ into an affine function $-\lambda\boldsymbol{w}^{(t)T}\boldsymbol{w}^{(t)} - 2\lambda\boldsymbol{w}^{(t)T}(\boldsymbol{w} - \boldsymbol{w}^{(t)T})$ (Yuille and Rangarajan, 2002; Shen et al., 2016). As we can easily serialize $\boldsymbol{H}$ into a stream of individual documents, PADD takes only $\mathcal{O}(K\max(N, M))$ space. By our design, CCP can effectively solve each sub-problem for every document m in parallel. The master-problem needs at most $\mathcal{O}(K^2)$ updates on the dual matrix $\boldsymbol{\Lambda}$.

Note that spectral topic models first learn the per-word topic distributions $\breve{\boldsymbol{B}} \in \mathbb{R}^{K\times N}$, then recovering the topics $\boldsymbol{B}$ from $\breve{\boldsymbol{B}}$ via applying the Bayes rule (Arora et al., 2013). Since $\breve{\boldsymbol{B}}_{ki}$ indicates the conditional probability of the topic k given the word i, we set the initializer as $\boldsymbol{W}^0 = \breve{\boldsymbol{B}}\widetilde{\boldsymbol{H}}$ whose $\boldsymbol{W}^0_{km} = \sum_i p(z{=}k|x{=}i)p(x{=}i; m) = p(z{=}k; m)$. Users can also warm-start with the current solution $\bar{\boldsymbol{W}}_{*m}$ to further accelerate the convergence.

[3] We do not explicitly introduce the Lagrange multipliers for the simplex constraint. Instead we later minimize the Lagrangian explicitly with this constraint $\boldsymbol{w}_m \in \Delta^{K-1}$. The operation $\langle\cdot,\cdot\rangle_F$ indicates the Frobenius inner product.

3.2 Prior-Aware Manifold Iteration (PAMI)

The Prior-Aware Manifold Iteration (PAMI) again tries to solve the optimization problem (1). Unlike PADD, however, the PAMI algorithm re-parametrizes the problem to explicitly enforce the constraint $(1/M)\boldsymbol{W}\boldsymbol{W}^T = \boldsymbol{A}$. Let $\boldsymbol{W}^T = \boldsymbol{Q}\boldsymbol{R}$ be the economy QR decomposition in which $\boldsymbol{R}$ has positive diagonal. Then $\boldsymbol{W}\boldsymbol{W}^T = \boldsymbol{R}^T(\boldsymbol{Q}^T\boldsymbol{Q})\boldsymbol{R} = \boldsymbol{R}^T\boldsymbol{R}$, and hence we automatically satisfy the constraint $\boldsymbol{W}\boldsymbol{W}^T = M\boldsymbol{A}$ by writing $\boldsymbol{W}^T = \boldsymbol{Q}\boldsymbol{R}$ where $\boldsymbol{R}$ is the Cholesky factor of $M\boldsymbol{A}$.

Define $J(\boldsymbol{W}) = \sum_m \|\boldsymbol{B}\boldsymbol{w}_m - \widetilde{\boldsymbol{h}}_m\|_2^2$. Say $\boldsymbol{e}_K$ means a K-dimensional vector of 1. Then $J(\boldsymbol{W}) = \operatorname{tr}(\boldsymbol{B}\boldsymbol{A}\boldsymbol{B}^T) - 2\operatorname{tr}(\boldsymbol{H}^T\boldsymbol{B}\boldsymbol{W}) + \operatorname{tr}(\boldsymbol{H}^T\boldsymbol{H})$ given $\sum_m \boldsymbol{w}_m\boldsymbol{w}_m^T = M\boldsymbol{A}$. Since the first and the third terms do not depend on the choice of $\boldsymbol{W}$, solving (1) is equivalent to maximizing $J(\boldsymbol{W}) = tr(\boldsymbol{H}^T\boldsymbol{B}\boldsymbol{W})$ subject to $\boldsymbol{W}\boldsymbol{W}^T = M\boldsymbol{A}$, $\boldsymbol{e}_K^T\boldsymbol{W} = \boldsymbol{e}_M^T$, and $\boldsymbol{W} \geq 0$. The decomposition $\boldsymbol{W}^T = \boldsymbol{Q}\boldsymbol{R}$ allows us to transform our optimization into the following simplex-constrained *Procrustes problem*:

$$\begin{aligned} \min \quad & g(\boldsymbol{Q}) = -\operatorname{tr}(\boldsymbol{Q}\boldsymbol{R}\boldsymbol{B}^T\boldsymbol{H}) \\ \text{subject to} \quad & \boldsymbol{Q}^T\boldsymbol{Q} = \boldsymbol{I}_K,\ \boldsymbol{Q}\boldsymbol{R}\boldsymbol{e}_K = \boldsymbol{e}_M,\ \boldsymbol{Q}\boldsymbol{R} \geq 0. \end{aligned} \tag{3}$$

Let $\mathcal{Q}$ denote the *Stiefel manifold* of $M \times K$ matrices with orthonormal columns, and let $\mathcal{P} = \mathcal{P}(\boldsymbol{R})$ be the (convex) set of $M \times K$ matrices that satisfy $\boldsymbol{P}\boldsymbol{R}\boldsymbol{e}_K = \boldsymbol{e}_M$ and $\boldsymbol{P}\boldsymbol{R} \geq 0$ for any $\boldsymbol{P} \in \mathcal{P}$. Let $f : \mathbb{R}^{M\times K} \to \mathbb{R}$ be an indicator (non-smooth but semi-continuous) function that is 0 if $\boldsymbol{P} \in \mathcal{P}$ or ∞ otherwise. Then we can formulate a consensus version of the problem (3):

$$\min_{\boldsymbol{P}\in\mathcal{P},\boldsymbol{Q}\in\mathcal{Q}} \ f(\boldsymbol{P}) + g(\boldsymbol{Q}) \ \text{ subject to } \ \boldsymbol{Q} - \boldsymbol{P} = 0. \tag{4}$$

Using the matrix dual variables $\boldsymbol{\Lambda}' \in \mathbb{R}^{M\times K}$, we construct the augmented Lagrangian with the penalty term ρ: $L_\rho(\boldsymbol{P}, \boldsymbol{Q}, \boldsymbol{\Lambda}') = f(\boldsymbol{P}) + g(\boldsymbol{Q}) + \langle \boldsymbol{\Lambda}', \boldsymbol{Q} - \boldsymbol{P}\rangle_F + \frac{\rho}{2}\|\boldsymbol{Q} - \boldsymbol{P}\|_F^2$. If rescaling $\boldsymbol{\Lambda} := \frac{1}{\rho}\boldsymbol{\Lambda}'$, then $L_\rho(\boldsymbol{P}, \boldsymbol{Q}, \boldsymbol{\Lambda}) = f(\boldsymbol{P}) + g(\boldsymbol{Q}) + \frac{\rho}{2}\|\boldsymbol{F} + \frac{1}{\rho}\boldsymbol{\Lambda}'\|_F^2 - \frac{1}{2\rho}\langle \boldsymbol{\Lambda}', \boldsymbol{\Lambda}'\rangle_F$. Based on the Manifold Alternating Direction Method of Multipliers (MADMM) (Kovnatsky et al., 2016; Chen et al., 2018), we can estimate the near-optimal primal $\boldsymbol{Q}^*, \boldsymbol{P}^*$, given our documents $\boldsymbol{H}$, the learned topics $\boldsymbol{B}$, and the invariant factor $\boldsymbol{R}$ from the learned correlations $\boldsymbol{A}$, by iterating the following updates until the convergence.

$$\begin{aligned} \boldsymbol{P}^{(t+1)} &:= \operatorname*{arg\,min}_{\boldsymbol{P}\in\mathcal{P}} \ L_\rho(\boldsymbol{P}, \boldsymbol{Q}^{(t)}, \boldsymbol{\Lambda}^{(t)}) \\ &= \operatorname*{arg\,min}_{\boldsymbol{P}\in\mathcal{P}} \ f(\boldsymbol{P}) + \frac{\rho}{2}\|\boldsymbol{Q}^{(t)} - \boldsymbol{P} + \boldsymbol{\Lambda}^{(t)}\|_F^2 \\ \boldsymbol{Q}^{(t+1)} &:= \operatorname*{arg\,min}_{\boldsymbol{Q}\in\mathcal{Q}} \ L_\rho(\boldsymbol{P}^{(t+1)}, \boldsymbol{Q}, \boldsymbol{\Lambda}^{(t)}) \\ &= \operatorname*{arg\,min}_{\boldsymbol{Q}\in\mathcal{Q}} \ g(\boldsymbol{Q}) + \frac{\rho}{2}\|\boldsymbol{Q} - \boldsymbol{P}^{(t+1)} + \boldsymbol{\Lambda}^{(t)}\|_F^2 \\ \boldsymbol{\Lambda}^{(t+1)} &:= \ \boldsymbol{\Lambda}^{(t)} + \boldsymbol{Q}^{(t+1)} - \boldsymbol{P}^{(t+1)}. \end{aligned} \tag{5}$$

For scalable inference of (5), we need to efficiently find the minimizers $\boldsymbol{P}$ and $\boldsymbol{Q}$ at every iteration. Because f diverges to ∞ outside $\mathcal{P}$, the first sub-problem must choose $\boldsymbol{P}^{(t+1)}$ as the orthogonal projection $\Pi_{\mathcal{P}}(\boldsymbol{Q}^{(t)} + \boldsymbol{\Lambda}^{(t)})$. Note that $\boldsymbol{P} \in \mathcal{P}$ is identical to putting M independent constraints: $\boldsymbol{P}_{m*}\boldsymbol{R}\boldsymbol{e}_K = 1$ and $\boldsymbol{P}_{m*}\boldsymbol{R} \geq 0$ for each row $\boldsymbol{P}_{m*}$. If defining $\bar{\boldsymbol{w}}_m := \boldsymbol{R}^T\boldsymbol{P}_{*m}$ and $\boldsymbol{V} := \boldsymbol{Q}^{(t)} + \boldsymbol{\Lambda}^{(t)}$, finding $\Pi_{\mathcal{P}}(\boldsymbol{V})$ is equivalent to finding $\bar{\boldsymbol{w}}_m \in \mathbb{R}^K$ that minimizes $\|\boldsymbol{R}^{-T}\bar{\boldsymbol{w}}_m - \boldsymbol{V}_{m*}\|_2$ under the simplex constraint: $\boldsymbol{e}_K^T\bar{\boldsymbol{w}}_m = 1$ and $\bar{\boldsymbol{w}}_m \geq 0$. This Simplex-Constrained Least Square (SCLS) is a convex problem that can be efficiently solved with machine precision for each document m in parallel. For the second sub-problem, minimizing $g(\boldsymbol{Q}) + \frac{\rho}{2}\|\boldsymbol{Q} - \boldsymbol{P}^{(t+1)} + \boldsymbol{\Lambda}^{(t)}\|_F^2$ is equivalent to maximizing $\operatorname{tr}(\boldsymbol{Q}^T(\boldsymbol{H}^T\boldsymbol{B}\boldsymbol{R}^T + \rho\boldsymbol{P}^{(t+1)} - \rho\boldsymbol{\Lambda}^{(t)}))$ on the Stiefel manifold $\boldsymbol{Q} \in \mathcal{Q}$. One can easily get the Procrustes closed-form solution $\boldsymbol{Q}^{(t+1)} = \boldsymbol{U}_K\boldsymbol{V}_K^T$ by finding the K leading left-singular and right-singular vectors $\boldsymbol{U}_K$ and $\boldsymbol{V}_K$ from just one economic SVD of $(\boldsymbol{H}^T\boldsymbol{B}\boldsymbol{R}^T + \rho\boldsymbol{P}^{(t+1)} - \rho\boldsymbol{\Lambda}^{(t)}) \in \mathbb{R}^{K\times K}$.

4 Experimental Results

Evaluating the learned topic compositions $\boldsymbol{W}$ is not easy for real data because no ground-truth compositions $\boldsymbol{W}^*$ exist for quantitative comparison. Unlike reading the topics from the word-topic matrix $\boldsymbol{B}$, the topic-document matrix $\boldsymbol{W}$ does not support qualitative evaluations because the number of documents M is too large, and because topics in each document may not be obviously coherent or incoherent as words in each topic (Chang et al., 2009). Synthesizing documents from scratch is a popular option for theoreticians as we can manipulate the ground-truth $\boldsymbol{B}^*$ and $\boldsymbol{W}^*$, but the resulting documents would not be realistic enough to satisfy practitioners. Therefore we generate two corpora from two distinct processes but both using topics learned from real data.

The uncorrelated setting (*Semi-Synthetic, SS*) samples $\boldsymbol{W}^*$ from a LDA model with a Dirichlet prior. Given the training data $\boldsymbol{H}_0$ for each of the real corpora, we first run JSMF to learn K topics $\boldsymbol{B}_0$ and their correlations $\boldsymbol{A}_0$ with the AP-rectification.[4] Next we sample M columns of $\boldsymbol{W}^*$ from the $\text{Dir}(\boldsymbol{\alpha})$ with $\boldsymbol{\alpha} = (5/K)\boldsymbol{e}_K$, generating a corpus $\boldsymbol{H}^{SS}$ given $\boldsymbol{B}_0$ and $\boldsymbol{W}^*$. *SS* is far from realistic as the underlying topics in $\boldsymbol{H}^{SS}$ barely correlate with each other. But this design choice provides a fair comparison to the experiments of TLI in (Arora et al., 2016). On the other hand, the

[4]Spectral topic models are known unable to learn quality topics on real data that does not align with the model assumptions. Rectifying $\boldsymbol{C}$ in advance with the Alternating Projection is proven effective to learn competitive topics comparable to probabilistic inference (Lee et al., 2015).

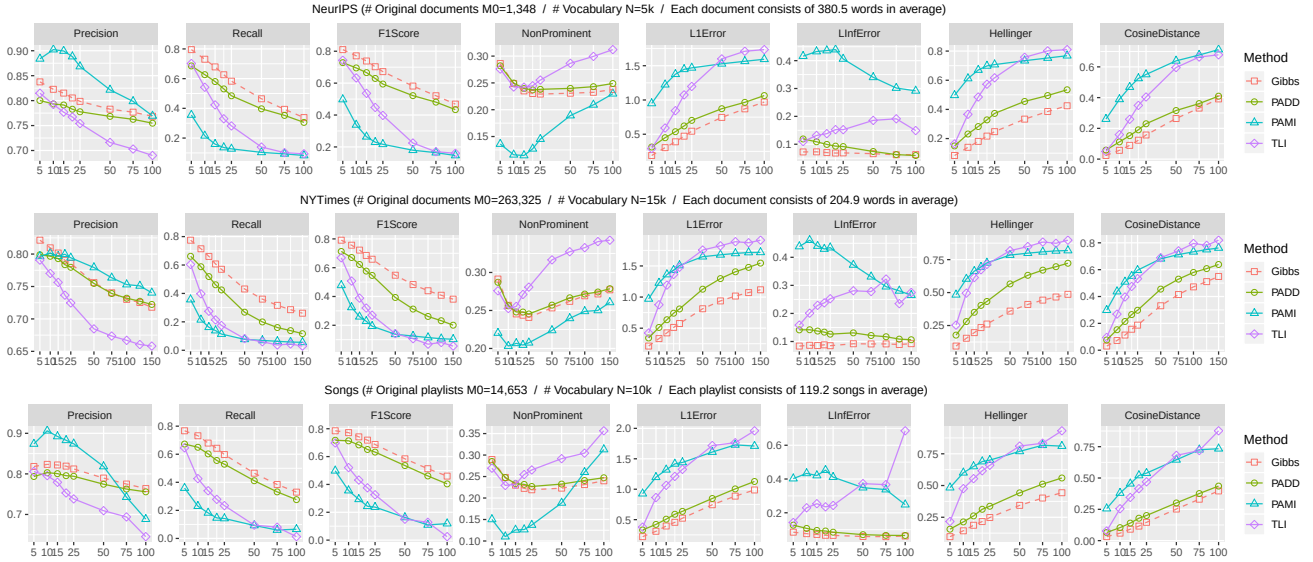

Figure 3: Semi-Synthetic (*SS*) corpus with highly sparse topics and little correlation. x-axis: # topics K. y-axis: higher numbers are better for the left three columns, lower numbers are better for the right five columns. PADD performs close to Gibbs across all settings, whereas PAMI performs poorly due to little correlation. TLI generally fails except on small K's.

correlated setting (*Semi-Real, SR*) samples $\boldsymbol{W}^*$ from a CTM model with the logistic-normal prior. Given the real training data $\boldsymbol{H}_0$, we first run CTM-Gibbs (Chen et al., 2013) to learn K topics $\boldsymbol{B}_0$ and the prior parameters $(\boldsymbol{\mu}_0, \boldsymbol{\Sigma}_0)$. Rather than artificially controlling the prior hyper-parameters like $\boldsymbol{H}^{SS}$, we synthesize a corpus $\boldsymbol{H}^{SR}$ given $\boldsymbol{B}_0$ and $\boldsymbol{W}^*$ by sampling $\boldsymbol{W}^*$ from the learned prior $\mathcal{LN}(\boldsymbol{\mu}_0, \boldsymbol{\Sigma}_0)$. Thus *SR* involves non-trivial correlation between topics, closely simulating the characteristics of real-world document collections.

We also prepare real corpora (*Fully-Real, FR*) $\boldsymbol{H}^{FR}$, which is the 10% held-out of the original data that has never been used in training (i.e., $\boldsymbol{H}_0 \cap \boldsymbol{H}^{FR} = \emptyset$). The training and held-out data $\boldsymbol{H}_0 \cup \boldsymbol{H}^{FR}$ come from the standard sources: NeurIPS papers and NYTimes articles in the UCI repository. We also prepare Yelp reviews from the academic dataset used in (Lee and Mimno, 2014). These reviews are short, being difficult to understand alone without a proper context. Beyond these textual datasets, each playlist in our Songs dataset consists of a sequence of songs played in real music stations (Chen et al., 2012). Different from the burstiness of words in real texts, individual playlists are less likely to repeat the same song multiples times, creating another challenge. We curate the vocabulary identically to the previous work that adopts these datasets Lee et al. (2015). The precise statistics of individual datasets are available in the result figures.

For thorough validation, we evaluate both information retrieval performance and metric similarities. Given the learned composition $\boldsymbol{W}$, we extract the prominent topics for each document by selecting the most contributing topics first until their cumulative mass gets close to 0.8 (Yao et al., 2009). Then we measure the precision, recall, and F1-score against the prominent topics of the truth compositions $\boldsymbol{W}^*$. Non-Prominent captures how much probability mass $\boldsymbol{W}$ puts on the non-prominent topics of $\boldsymbol{W}^*$. For metric similarities, we report ℓ_1-error and ℓ_∞-error for fair comparison to (Arora et al., 2016) as well as Hellinger and cosine distances similar to (Blei and Lafferty, 2007).

Note that Gibbs in our synthetic experiments is not a baseline that the two prior-aware algorithms try to outperform, but the strongest benchmark designed to guide the maximal performance. We run the collapsed Gibbs sampling[5] with the fixed ground-truth topics: $\boldsymbol{B}_0$, the true prior distributions: Dir or $\mathcal{LN}$, and their ground-truth hyper-parameters: $(5/K)\boldsymbol{e}_K$ or $(\boldsymbol{\mu}_0, \boldsymbol{\Sigma}_0)$ that are used to generate the datasets, then updating only the topic compositions $\boldsymbol{W}$. In contrast, we run TLI provided with $\boldsymbol{B}_0$ and PADD/PAMI provided with $\boldsymbol{B}_0$ and $\boldsymbol{A}_0$ to leverage the topic correlations. Thus the spectral algorithms access neither the specific prior distributions nor their ground-truth hyper-parameters.

In the uncorrelated setting (*SS*) in Figure 3, PADD performs close to Gibbs in all dataset and model, whereas PAMI works poorly as the number of topics K grows. This is because the topics in *SS* barely have correlations as the concentration parameter $(5/K)\boldsymbol{e}_K$ gets close to $\mathbf{0}$. Thus PAMI cannot extract any useful invariant information from the topic correlations. Note that even Gibbs sampling shows relatively high ℓ_1-error especially for the models with large K. $\mathrm{Dir}((5/K)\boldsymbol{e}_K)$ generates

[5]We discard the initial 200 burn-in samples, then further run Gibbs 1,000 iterations for collecting quality samples more close to the posterior compositions.

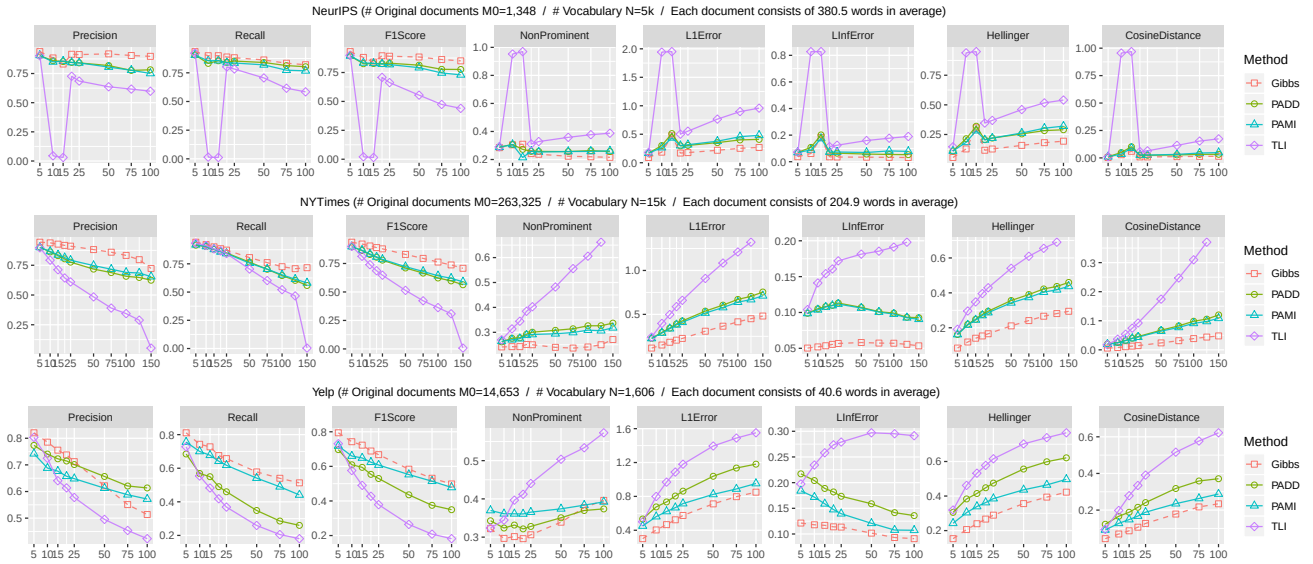

Figure 4: Semi-Real (*SR*) corpus with realistic topic correlations. x-axis: # topics K. y-axis: higher numbers are better for the left three columns, lower numbers are better for the right five columns. PADD/PAMI both perform closely to Gibbs on NeurIPS and NYTimes, but PAMI outperforms PADD on the small short dataset, Yelp. TLI fails except on tiny K's.

extremely sparse compositions, so any variability in other topics causes catastrophic errors even with Gibbs sampling with the ground-truth hyper-parameters. In contrast, the situation is quite different in the correlated setting (*SR*) in Figure 4. Both PADD and PAMI are now comparable to Gibbs on NeurIPS and NYTimes, but only PAMI is close to Gibbs on Yelp dataset. This result agrees with the algorithm of PAMI that relies more aggressively on the prior information given in the topic correlations. We also verify that our experiments on *SS* and *SR* are not sensitive to the different numbers (1k, 5k, 10k, 50k, 100k) of synthesized documents.

In the real setting (*FR*) in Figure 5, we only report F1Score and Hellinger as they represent the overall behaviors. PADD/PAMI lose some precision comparing to *SS* and *SR* settings, but this is because we pretend that Gibbs provides the ground-truth compositions $\boldsymbol{W}^*$.[6] Overall, PADD and PAMI perform comparably in NYTimes and Yelp, whereas they become separated from $K=50$ in Songs. It motivates us to measure other metrics for deeper understanding. LossDiff and RegularizerDiff evaluate $\|\boldsymbol{B}\boldsymbol{W}^* - \widetilde{\boldsymbol{H}}\|_F^2 - \|\boldsymbol{B}\boldsymbol{W} - \widetilde{\boldsymbol{H}}\|_F^2$ and $\|(1/M)\boldsymbol{W}^*\boldsymbol{W}^{*T} - \boldsymbol{A}_0\|_F - \|(1/M)\boldsymbol{W}\boldsymbol{W}^T - \boldsymbol{A}_0\|_F$ that are the differences in the two parts of our objective (1) against the true composition $\boldsymbol{W}^*$. It shows PAMI is more afraid of deviating from the learned correlations $\boldsymbol{A}$, whereas PADD is more afraid of poorly fitting the given observations $\boldsymbol{H}$. Note also that PADD/PAMI sometimes perform better than Gibbs in terms of TotalDiff, the difference in our objective value (1).

[6]We run Dirichlet Gibbs with $\boldsymbol{B}_0$ and the best hyper-parameters fitted by moment-matching with $\boldsymbol{A}_0$, which are no longer ground-truth but the best accessible information.

What is going on? Probabilistic topic models (especially Bayesian) try to infer both latent topics $\boldsymbol{B}$ and hidden compositions $\boldsymbol{W}$ that approximately maximize the marginal likelihood of the observed documents $\boldsymbol{H}$:

$$\prod_{m=1}^{M} \int_{\boldsymbol{w}_m} p(\boldsymbol{w}_m|\mathfrak{f}) \prod_{i=1}^{N} (\boldsymbol{B}\boldsymbol{w}_m)_i^{\boldsymbol{h}_{mi}} d\boldsymbol{w}_m. \tag{6}$$

They consider all possible compositions $\boldsymbol{w}_m \in \Delta^{K-1}$ under the prior $\mathfrak{f}$. However, if the learned topics $\boldsymbol{B}$ and correlations $\boldsymbol{A}$ (which takes the role of $\mathfrak{f}$) are provided, the MAP estimator $\boldsymbol{w}_m$ that maximizes $p(\boldsymbol{w}_m; \boldsymbol{A}) \prod_{i=1}^{N} (\boldsymbol{B}\boldsymbol{w}_m)_i^{\boldsymbol{h}_{mi}}$ is a reasonable pointwise choice.[7] Recall that the MLE of $\boldsymbol{w}_m$ — that maximizes the likelihood of the multinomially chosen words $\boldsymbol{h}_m$ — is the composition that makes the word-probability parameters $\boldsymbol{B}\boldsymbol{w}_m$ equal to the empirical frequencies $\widetilde{\boldsymbol{h}}_m$. The loss function of our objective (1) tries to find the best $\boldsymbol{w}_m$ that makes $\boldsymbol{B}\boldsymbol{w}_m \approx \widetilde{\boldsymbol{h}}_m$, thereby maximizing the likelihood term $\prod_{i=1}^{N} (\boldsymbol{B}\boldsymbol{w}_m)_i^{\boldsymbol{h}_{mi}}$. LogLikeliDiff measures an average difference of the log-likelihood $\sum_i \boldsymbol{h}_{mi} \log(\boldsymbol{B}\boldsymbol{w}_m)_i - \sum_i \boldsymbol{h}_{mi} \log(\boldsymbol{B}\boldsymbol{w}^*_m)_i$ against the true composition $\boldsymbol{w}^*$. The similar trends between LogLikeliDiff and LossDiff support this observation.[8]

When we try to regularize individual compositions $\boldsymbol{w}_m$ based on the learned topic correlations $\boldsymbol{A}$, our spectral algorithms adopt the Frobenius norm to measure the deviation, whereas a specific prior $\mathfrak{f}$ regulates $\boldsymbol{w}_m$ in probabilistic inference. As we are not aware of any method

[7]It is proven that Bayesian posterior is concentrated in the ϵ-ball of the best pointwise estimator with high probability (Arora et al., 2016).

[8]We keep denoting m-th column vector of $\boldsymbol{H}$ by $\boldsymbol{h}_m$. Thus $\boldsymbol{h}_{mi}$ means (i,m)-entry $\boldsymbol{H}_{im}$ of $\boldsymbol{H}$.

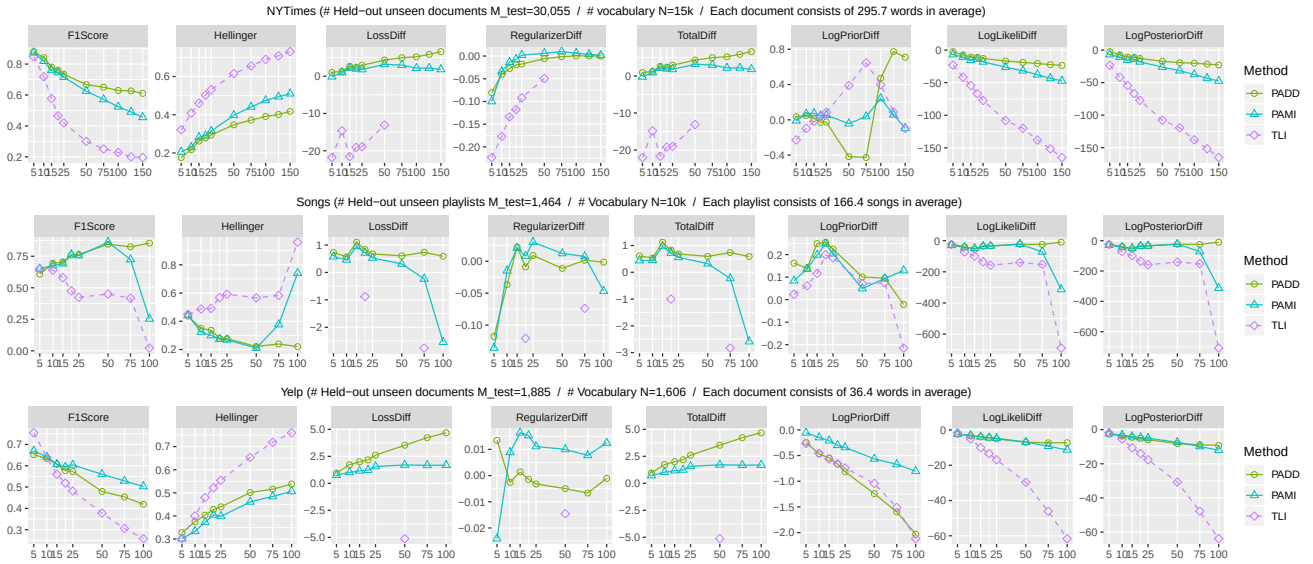

Figure 5: Fully-Real (FR) corpus consisting of real documents held-out from the training data. x-axis: # topics K. y-axis: higher numbers are better for every column except the second column. Pretending Gibbs provides the ground-truth. PADD and PAMI split on Songs from $K \geq 50$, showing difference in their underlying behaviors. TLI fails completely.

that samples a rank-1 correlation matrix $\boldsymbol{A}_m = \boldsymbol{w}_m \boldsymbol{w}_m^T$ directly from $\boldsymbol{A}$, we introduce a generic metric: LogPriorDiff that measures the average difference in log priors $\log(1 - HD(\boldsymbol{w}_m \boldsymbol{w}_m^T \| \boldsymbol{A}_0)) - \log(1 - HD(\boldsymbol{w}_m^* \boldsymbol{w}_m^{*T} \| \boldsymbol{A}_0))$ against the true composition $\boldsymbol{w}_m^*$. This metric is induced by Hellinger Distance under the assumption: $p(\boldsymbol{w}_m | \boldsymbol{A}) \propto 1 - HD(\boldsymbol{w}_m \boldsymbol{w}_m^T \| A)$.[9] The disagreement between RegularizerDiff and LogPriorDiff implies that we can further improve PADD/PAMI by finding a better weighting scheme rather than measuring the prior deviation uniformly for all entry by the Frobenius norm.

Although we are optimizing for accuracy rather than speed, PADD and PAMI are both competitive in their time complexities. Running PADD and PAMI on our largest NYTimes dataset with $K = 50$ takes 6804.5 and 7117.9 seconds in Matlab without using the per-document parallelism. TLI takes 6780.3 seconds to learn the inverse of $\boldsymbol{B}$ with the 8-core parallelism in Python codes provided by the original authors (Arora et al., 2016). Running Gibbs with 8-cores in parallel takes 3,420 seconds in the optimized Java Mallet. It is worth mentioning that TLI often produces NaN entries due to the numerical instability of the matrix inversion. AP-rectification (Lee et al., 2015) vastly improves this problem, but many data points for TLI in Figure 5 are still missing.[10] AP-rectification also helps PAMI ensure the positive semidefiniteness of $\boldsymbol{A}$. Throughout the experiment, we run 50 iterations of PADD with $\tau = 0.05$ and 50 iterations of PAMI with $\rho = 0.01$.

[9] Hellinger Distance is normalized to $[0, 1]$, allowing us to come up with a generic unnormalized potential.

[10] We try our best to ignore NaN entries, but diverging to $\pm\infty$ drops the data points in Loss/Regularizer panels.

5 Conclusion

Fast and accurate composition inference for new documents is a vital component of a topic-based workflow, especially for spectral algorithms that do not by themselves produce topic compositions, even for training documents. Putting a prior on these compositions is the crux for learning coherent topics, but its identity has been unclear for spectral topic models. We show that topic correlations serve as a flexible parameter-free prior, and we design two novel algorithms that take the advantage of these correlations to infer quality compositions. Our Prior-Aware Dual Decomposition (PADD) performs close to the benchmark Gibbs sampling across nearly all settings, whereas Prior-Aware Manifold Iteration (PAMI) shows its excellence in learning from short real documents by leveraging the topic correlations more aggressively than PADD. Both algorithms fit the workflow of modern distributed systems, being easily scalable on streaming document collections.

With the robust composition inference that is aware of topic correlations latent in the data, we can now fill out the necessary tools to make spectral topic models a full competitor to likelihood-based methods. Our prior-aware algorithms require neither a specific prior distribution nor its proper hyper-parameters, which are often difficult to know a priori. Users in our algorithms can better control the trade-off between loss-minimization and regularization, gaining additional flexibility. Although the benefits of our composition inference are mostly relevant in second-order spectral models, they are widely applicable in any setting that involves inferring mixture proportions or, more broadly, finding MAP estimates with the matrix-induced priors.

Acknowledgements

This work was completed when the first author was visiting Microsoft Research at Redmond. We thank both the department of Information and Decision Sciences in UIC Business School and the Deep Learning group in Microsoft Research. They provided useful resources and thoughtful advice.

References

A. Anandkumar, D. P. Foster, D. Hsu, S. Kakade, and Y. Liu. A spectral algorithm for latent Dirichlet allocation. In *NIPS*, 2012a.

A. Anandkumar, S. M. Kakade, D. P. Foster, Y.-K. Liu, and D. Hsu. Two svds suffice: Spectral decompositions for probabilistic topic modeling and latent dirichlet allocation. Technical report, 2012b.

A. Anandkumar, D. J. Hsu, M. Janzamin, and S. Kakade. When are overcomplete topic models identifiable? uniqueness of tensor tucker decompositions with structured sparsity. *CoRR*, 2013.

F. Arabshahi and A. Anandkumar. Spectral methods for correlated topic models. *AISTATS*, 2017.

S. Arora, R. Ge, and A. Moitra. Learning topic models – going beyond SVD. In *FOCS*, 2012.

S. Arora, R. Ge, Y. Halpern, D. Mimno, A. Moitra, D. Sontag, Y. Wu, and M. Zhu. A practical algorithm for topic modeling with provable guarantees. In *ICML*, 2013.

S. Arora, R. Ge, F. Koehler, T. Ma, and A. Moitra. Provable algorithms for inference in topic models. In *ICML*, pages 2859–2867, 2016.

A. Asuncion, M. Welling, P. Smyth, and Y. W. Teh. On smoothing and inference for topic models. In *In Proceedings of the 25th Conference on Uncertainty in Artificial Intelligence*, 2009.

T. Bansal, C. Bhattacharyya, and R. Kannan. A provable svd-based algorithm for learning topics in dominant admixture corpus. In *Advances in Neural Information Processing Systems 27*. 2014.

D. Blei and J. Lafferty. A correlated topic model of science. *Annals of Applied Statistics*, pages 17–35, 2007.

D. Blei, A. Ng, and M. Jordan. Latent Dirichlet allocation. *Journal of Machine Learning Research*, pages 993–1022, 2003. Preliminary version in *NIPS* 2001.

J. Chang, J. Boyd-Graber, C. Wang, S. Gerrish, and D. M. Blei. Reading tea leaves: How humans interpret topic models. In *NIPS*, 2009.

J. Chen, J. Zhu, Z. Wang, X. Zheng, and B. Zhang. Scalable inference for logistic-normal topic models. In *NIPS*, pages 2445–2453, 2013.

S. Chen, J. L. Moore, D. Turnbull, and T. Joachims. Playlist prediction via metric embedding. In *Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 714–722. ACM, 2012.

S. Chen, S. Ma, A. M.-C. So, and T. Zhang. Proximal gradient method for manifold optimization. *arXiv preprint arXiv:1811.00980*, 2018.

M. Erlin. Topic modeling, epistemology, and the english and german novel. *Cultural Analytics*, May 2017.

A. Goldstone and T. Underwood. The quiet transformations of literary studies: What thirteen thousand scholars could tell us. *New Literary History*, 45(3), Summer 2014.

D. Hall, D. Jurafsky, and C. D. Manning. Studying the history of ideas using topic models. In *EMNLP*, pages 363–371, 2008.

T. Hofmann. Probabilistic latent semantic analysis. In *UAI*, pages 289–296, 1999.

K. Huang, X. Fu, and N. D. Sidiropoulos. Anchor-free correlated topic modeling: Identifiability and algorithm. In *NIPS*, 2016.

N. Komodakis, N. Paragios, and G. Tziritas. Mrf energy minimization and beyond via dual decomposition. *IEEE transactions on pattern analysis and machine intelligence*, 33(3):531–552, 2011.

A. Kovnatsky, K. Glashoff, and M. M. Bronstein. Madmm: a generic algorithm for non-smooth optimization on manifolds. In *European Conference on Computer Vision*, pages 680–696. Springer, 2016.

M. Lee and D. Mimno. Low-dimensional embeddings for interpretable anchor-based topic inference. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1319–1328. Association for Computational Linguistics, 2014.

M. Lee, D. Bindel, and D. Mimno. Robust spectral inference for joint stochastic matrix factorization. In *NIPS*, 2015.

Y. Liu, A. Niculescu-Mizil, and W. Gryc. Topic-link lda: Joint models of topic and author community. In *ICML*, 2009.

A. M. Rush and M. Collins. A tutorial on dual decomposition and lagrangian relaxation for inference in natural language processing. *J. Artif. Intell. Res.(JAIR)*, 45:305–362, 2012.

X. Shen, S. Diamond, Y. Gu, and S. Boyd. Disciplined convex-concave programming. In *2016 IEEE 55th Conference on Decision and Control (CDC)*, pages 1009–1014. IEEE, 2016.

D. Sontag and D. Roy. Complexity of inference in latent Dirichlet allocation. In *NIPS*, pages 1008–1016, 2011.

M. Steyvers and T. L. Griffiths. Rational analysis as a link between human memory and information retrieval. *The probabilistic mind: Prospects for Bayesian cognitive science*, pages 329–349, 2008.

E. M. Talley, D. Newman, D. Mimno, B. W. H. II, H. M. Wallach, G. A. P. C. Burns, M. Leenders, and A. McCallum. Database of nih grants using machine-learned categories and graphical clustering. *Nature Methods*, 8(7):443–444, June 2011.

H. M. Wallach, D. M. Mimno, and A. McCallum. Rethinking lda: Why priors matter. In *NIPS*. 2009.

Y. Wang and J. Zhu. Spectral methods for supervised topic models. In *NIPS*, 2014.

L. Yao, D. Mimno, and A. McCallum. Efficient methods for topic model inference on streaming document collections. In *KDD*, 2009.

Y. Yao, A. Vehtari, D. Simpson, and A. Gelman. Yes, but did it work?: Evaluating variational inference. *Proceedings of the 28th International Conference on Machine Learning*, 2018.

X. Yu and E. Fokoue. Probit normal correlated topic model. In *Open Journal of Statistics*, pages 879–888, 2014.

A. L. Yuille and A. Rangarajan. The concave-convex procedure (cccp). In *Advances in neural information processing systems*, pages 1033–1040, 2002.