Journal of Machine Learning Research 18 (2018) 1-61 Submitted 9/16; Revised 10/17; Published 4/18

The Search Problem in Mixture Models

Avik Ray AVIK@UTEXAS.EDU
Department of Electrical and Computer Engineering
University of Texas at Austin
Austin, TX 78701, USA

Joe Neeman NEEMAN@IAM.UNI-BONN.DE
Department of Mathematics
Rheinische Friedrich-Wilhelms-Universität Bonn
D-53115 Bonn, Germany

Sujay Sanghavi SANGHAVI@MAIL.UTEXAS.EDU
Department of Electrical and Computer Engineering
University of Texas at Austin
Austin, TX 78701, USA

Sanjay Shakkottai SHAKKOTT@AUSTIN.UTEXAS.EDU
Department of Electrical and Computer Engineering
University of Texas at Austin
Austin, TX 78701, USA

Editor: Animashree Anandkumar

Abstract

We consider the task of learning the parameters of a *single* component of a mixture model, for the case when we are given *side information* about that component; we call this the "search problem" in mixture models. We would like to solve this with computational and sample complexity lower than solving the overall original problem, where one learns parameters of all components.

Our main contributions are the development of a simple but general model for the notion of side information, and a corresponding simple matrix-based algorithm for solving the search problem in this general setting. We then specialize this model and algorithm to four common scenarios: Gaussian mixture models, LDA topic models, subspace clustering, and mixed linear regression. For each one of these we show that if (and only if) the side information is informative, we obtain parameter estimates with greater accuracy, and also improved computation complexity than existing moment based mixture model algorithms (e.g. tensor methods). We also illustrate several natural ways one can obtain such side information, for specific problem instances. Our experiments on real data sets (NY Times, Yelp, BSDS500) further demonstrate the practicality of our algorithms showing significant improvement in runtime and accuracy.

Keywords: mixture models, search, side information, semi-supervised, method of moments

1. Introduction

Mixture models denote the statistical setting where observed samples can come from one of several distinct underlying populations—each typically with its own probability distribution—

but are not labeled as separate in the data presented. They have been used to model a wide variety of phenomena, and have seen great success in practice, going back as far as Pearson (1894). In this paper we consider (what we call) the **search problem** in the mixture model setting: given some *special side information* about one of the mixture components, is it possible to efficiently learn the parameters of that component only? Given that there are known methods for learning the entire set of parameters of various mixture models, "efficient" here means more efficient (statistically and/or computationally) than existing methods for learning all the parameters.

As an example, we consider the "latent Dirichlet allocation" model for document generation. In this model, "underlying population" means the set of topics in a document, which determines the frequencies of different words in the document. "Side information" could be a word that is more common in the topic of interest than it is in any other topic: for example, the word "semi-supervised" might work if the topic of interest is machine learning.

Side information could also consist of a small number of labelled examples. We might have a small collection of documents about machine learning and also a much larger corpus that includes documents from many topics. Our methods will allow us to leverage the large, unlabelled corpus to obtain good estimates for word frequencies in machine learning articles—and these estimates will be much better than anything that could be learned from the small labelled sample.

Main contributions: We propose a general setting for side information in mixture models, and show how to solve the search problem by estimating certain matrices of moments. We prove error bounds on the resulting estimates; our rates have a sharp dependence on the sample size (although they are possibly not sharp in the other parameters).

We then specialize our approach to four popular families of mixture models: Gaussian mixture models with spherical covariances, latent Dirichlet allocation for topic models, mixed linear regression, and subspace clustering. We give concrete algorithms for these four families. Our results also include new moment derivations for mixed linear regression and subspace clustering models.

Finally, we simulate our algorithm on both real and synthetic data sets for the Gaussian mixture model, topic model, and subspace clustering applications. For synthetic data set we compare its performance to the tensor decomposition methods discussed by Anandkumar et al. (2014) in both GMM and LDA models, and k-means for subspace clustering. We show that our methods outperform the baseline when the side information is informative. We also demonstrate the practical applicability of our algorithms on three real data sets—the NY Times data set of news articles, Yelp data set of business reviews, and BSDS500 data set of images. In the first two text corpus, we show our algorithm recovers more coherent topics than topic modeling algorithm by Arora et al. (2013). In the BSDS500 data set, we demonstrate how our algorithm can be used for parallel image segmentation. In all three cases, our algorithm also exhibits significant computational gains over competing unsupervised and semi-supervised algorithms.

1.1 Related Work

There is a vast literature on mixture models; too much to even summarize here. We will therefore focus this section on two more closely related areas: method of moments estimators for mixture models, and learning with side information.

Mixture models and method of moments: A common method for learning mixture models is the EM algorithm of Dempster et al. (1977), which outputs a complete set of model parameters. However, EM may converge slowly (or not at all) [Redner and Walker 1984]; this weakness of EM has spurred a resurgence in method-of-moments estimators for mixture models. Although these methods go back to the pioneering work of Pearson (1894) on Gaussian mixture models, the last several years have seen important advances. Moitra and Valiant (2010), and Hardt and Price (2015) showed that Gaussian mixture models with two components can be learned in polynomial time. Hsu and Kakade (2013) considered mixtures of more Gaussians, but constrained to have spherical covariances. They gave a method based on third-order tensor decompositions, which was later generalized to other models in Anandkumar et al. (2014).

Learning with side information: As has been observed many times, often in practice one has access to a set of data that is somewhat richer than standard models of data in learning theory. The term *side information* is used as a catch-all for extra data that doesn't fit into pre-existing models; as such, the literature contains many incomparable models of side information.

Xing et al. (2002) and Yang et al. (2010) took unsupervised clustering as their starting point. For them, side information arrived as pairs of points that were known to belong to the same cluster; they showed how this extra information could substantially improve the performance of the k-means algorithm.

Kuusela and Ocone (2004) developed a framework for side information in the PAC learning model, in which extra samples with a particular dependence on the original samples could sometimes give a substantial benefit.

Many different types of metadata have been proposed for the *latent Dirichlet allocation* (LDA) model of document generation. Mcauliffe and Blei (2008) introduced the *supervised LDA* model, in which each document comes with an additional response variable from a generalized linear model. On the other hand Rosen-Zvi et al. (2004) proposed the *author-topic model*, in which the metadata (author names) affects the distribution of the documents themselves. From a more experimental point of view, Lu and Zhai (2008) used long, detailed product reviews as side information for categorizing short snippets and blog entries.

The notion of *semi-supervised learning* (see the book by Chapelle et al. (2006)) is also related to our framework of side information. In semi-supervised learning, the learner has access to a small number of labelled examples and a large number of unlabelled examples. This setting is useful for us too, although our general method does not strictly require data of this form.

2. Basic Idea and Algorithm

We now first briefly describe the basic mixture model setting, and then describe our method. These descriptions cover several popular specific examples for mixture models, and we detail the application to each of them in Section 3.

Setting: We are interested in the standard statistical setting of (parametric) mixture models: that is, samples are drawn i.i.d. from a distribution f given by

$$f(x) \;=\; \sum_{i=1}^{k} \alpha_i \, g(x;\mu_i).$$

Here g corresponds to a known parametric class of distributions, and k is the number of mixture components. The corresponding parameter vectors are $\mu_1, \ldots, \mu_k$, and their mixture weights / probabilities are $\alpha_1, \ldots, \alpha_k$. So, for example, in the case of the standard (spherical) Gaussian mixture model, $g(x;\mu_i)$ is the Gaussian pdf $\mathcal{N}(\mu_i, I)$. Thus each sample can be considered to be drawn by first selecting a mixture component μ_i with probability α_i, and then drawing the sample x according to $g(x;\mu_i)$. We assume all the μ_i's are *linearly independent*. This is a common assumption for learning mixture models using spectral methods.

Search problem: The standard parameter estimation problem is to find all the μ_i vectors given samples. In this paper we are interested in the search problem: we are given *side information* about one of the vectors—say μ_1, without loss of generality—and we would like to recover *only* μ_1. Of course, we would like to do this with sample and computational complexity lower than what would be required to estimate all parameter vectors (i.e., lower complexity than the standard case).

Side information: Our general procedure requires the following model for side information: we assume that we have access to a vector v such that the inner product with the parameter vector μ_1—the special one we are searching for—is higher than the inner product with any of the other μ_i; i.e. there exists $\delta > 0$ such that;

$$\langle \mu_1, v\rangle \geq \; (1+\delta)\langle \mu_i, v\rangle \quad \text{for all } i \neq 1$$

Section 3 shows how to obtain such side information in some specific models of interest: spherical Gaussian mixture models, mixed linear regression, subspace clustering and the LDA topic model.

We remark that it's also possible (and perhaps more intuitive in some situations) to ask for side information satisfying $|\langle \mu_1, v\rangle| \geq (1+\delta)|\langle \mu_i, v\rangle|$. However, our assumption above is slightly weaker, since for any v satisfying the latter assumption, either v or $-v$ satisfies the former assumption. Later, we show the above condition is sufficient for uniquely identifying the required parameter μ_1 (but it may not be necessary). We refer side information vector v as *informative* about μ_1 if it satisfies the above condition.

2.1 General Procedure

The main idea behind method of moments is to use samples to estimate certain moments of the distribution $f(x)$, using which we can recover the parameters of interest. For many mixture models (including the four common examples we detail), it is possible to easily and directly estimate using first and second order moments, given sufficient samples, the vector

$$m := \sum_{i=1}^{k} \alpha_i \mu_i. \tag{1}$$

and the matrix

$$A := \sum_{i=1}^{k} \alpha_i \mu_i \mu_i^T. \tag{2}$$

For example, in many models the estimate of vector m is simply the sample mean, and matrix A can be derived from the sample covariance matrix. The exact procedure for estimating m and A varies according to the particular parametric model g. The fact that m and A (and also higher-order tensors) can be estimated from samples is well known for many models, see Anandkumar et al. (2014) for a treatment of several different models, and for other pointers to the literature.

Typically, all mixture model components cannot be identified from just the first and second order moments (or m and A). It is often necessary to compute even higher order moment terms. In our search problem, given the side information, **we develop** procedures to estimate an alternative matrix B, using higher order moments, given by

$$B := \sum_{i=1}^{k} \alpha_i \langle \mu_i, v \rangle \mu_i \mu_i^T \tag{3}$$

Again, the exact procedure for estimating B from samples depends on the particular parametric model g.

For this section, we assume we are able to estimate A, B, m to within some accuracy. We will use the notation $\hat{A}, \hat{B}, \hat{m}$ to denote these finite sample estimates of A, B, m respectively, and n denotes the number of samples used to compute these estimates. With this in hand, we outline two general procedures for estimating μ_1 (i.e. the component that we are interested in). The first procedure is based on a whitening step, much like the one that is used in the spectral algorithms in Hsu and Kakade (2013); Anandkumar et al. (2012), and tensor decomposition methods of Anandkumar et al. (2014) (please see remarks in Section 3 for the differences for specific models). The second procedure uses a line search instead, and may be computationally favorable when k is large, because it avoids the need to invert a $k \times k$ matrix. Both Algorithms 1 and 2 take as input the estimates $\hat{A}, \hat{B}, \hat{m}$ (where $\hat{B}$ is constructed using side information vector v) and they output estimates of the first mixture component $\hat{\mu}_1$, and also the proportion of the first component $\hat{\alpha}_1$.

2.1.1 The Whitening Method

Our main result about Algorithm 1 is that if $\hat{A}$ and $\hat{B}$ are good estimates of A and B then Algorithm 1 outputs good estimates for μ_1 and α_1. In order to interpret Theorem 1 as an error rate, note that if all parameters but ϵ are fixed then the error is $O(\epsilon)$. Since standard concentration results yield $\epsilon = O(n^{-1/2})$, where n is the number of samples; our error rate in terms of n is also $O(n^{-1/2})$. This rate is sharp, since it is also the rate for estimating the mean of a single Gaussian vector (i.e. a GMM with only one component).

Theorem 1 *Suppose that $\mu_1, \dots, \mu_k$ are linearly independent, and that $\hat{A}$ is positive semi-definite. Also suppose that $\langle \mu_1, v \rangle \geq (1+\delta)\langle \mu_i, v \rangle$ for all $i \neq 1$. Assume that*

$$\max\{\|A - \hat{A}\|, \|B - \hat{B}\|, \|m - \hat{m}\|\} \leq \epsilon < \sigma_k(A)/4,$$

Algorithm 1 Extracting a mixture component from side information: the whitening method.

Input: $\hat{A}, \hat{B}, \hat{m}$
Output: $\hat{\mu}_1, \hat{\alpha}_1$

```
let {σ_j, v_j} be the singular values and singular vectors of Â, in non-increasing order
let V be the d × k matrix whose jth column is v_j
let D be the k × k diagonal matrix with D_jj = σ_j
let u be the largest eigenvector of D^{-1/2} V^T B̂ V D^{-1/2}
let w = V D^{1/2} u
let E be the span of {V D^{1/2} v : v ⊥ u}
write V V^T m̂ (uniquely) as aw + y, where y ∈ E
return w/a and a^2
```

and that the right hand side of (4) *is at most* α_1*. Then*

$$\|\mu_1 - \hat{\mu}_1\| \leq CR|\alpha_1^{-1/2} - \hat{\alpha}_1^{-1/2}| + C\frac{\sqrt{\sigma_1(A)}}{\sqrt{\alpha_1}}\eta \quad , \text{ and}$$
$$|\alpha_1 - \hat{\alpha}_1| \leq \frac{C\sqrt{\alpha_1}(\alpha_1 R + \eta)}{\sigma_k(A)}\left(\eta + R\frac{\epsilon}{\sigma_k(A)} + \epsilon\right) \tag{4}$$

where $\eta = \frac{\epsilon\sigma_1}{\delta\sigma_k^{5/2}}$, $R = \max_i \|\mu_i\|$, $\sigma_1(A) \geq \cdots \geq \sigma_k(A) > 0$ *are the non-zero singular values of* $A = \sum_i \alpha_i \mu_i \mu_i^T$*, and* C *is a universal constant.*

Our error bounds are somewhat complicated, and depend on many different parameters, so let us elaborate on them slightly. First of all, the dependence on $\sigma_1(A)$ and $\sigma_k(A)$ is of the order $\|\mu_1 - \hat{\mu}_1\| \lesssim \sigma_1(A)^{3/2}/\sigma_k(A)^{5/2}$, which is probably an artifact of the analysis, and not the true behavior of the algorithm. On the other hand, our dependence on ϵ is optimal: we have $|\alpha_1 - \hat{\alpha}_1| \lesssim \epsilon$ and $\|\mu_1 - \hat{\mu}_1\| \lesssim \epsilon$. Note also that our bound has no explicit dependence on k; this feature comes from the fact that our method is targeted at a single mixture component. By comparison, other methods typically give bounds in which the *averaged* per-mixture-component error does not depend on k. In terms of dependence on k, therefore, our bounds are better than previous bounds if there is only one component of interest.

Finally, let us remark on the assumption that the right hand side of (4) is at most α_1. This amounts to an assumption that ϵ is sufficiently small compared to all the other parameters. Without this assumption, the bound in (4) would not be very interesting, since $|\alpha_1 - \hat{\alpha}_1| \leq \alpha_1$ is too weak to give useful information about $\hat{\alpha}_1$ (it could even be zero).

We defer the actual analysis of Algorithm 1 to the appendix, but we will motivate the algorithm and give the basic idea of the proof by showing that if $\hat{A}, \hat{B}$, and $\hat{m}$ are equal to A, B and m respectively then Algorithm 1 outputs μ_1 and α_1 exactly.

Lemma 2 *Let* m*,* A*, and* B *be defined by in* (1)*,* (2)*, and* (3)*, where* $\mu_1, \ldots, \mu_k$ *are linearly independent. If* $\langle \mu_1, v\rangle > \langle \mu_i, v\rangle$ *for all* $i \neq 1$ *and we apply Algorithm 1 to* A*,* B*, and* m*, then it returns* μ_1 *and* α_1*.*

Proof Let V and D be as defined in Algorithm 1. Since A has rank k,

$$\sum_{i=1}^{k} \alpha_i D^{-1/2} V^T \mu_i \mu_i^T V D^{-1/2} = D^{-1/2} V^T A V D^{-1/2} = I_k.$$

Defining $u_i := \sqrt{\alpha_i} D^{-1/2} V^T \mu_i$, we have $\sum_i u_i u_i^T = I_k$, which implies that the u_i are orthonormal in $\mathbb{R}^k$. Now,

$$D^{-1/2} V^T B V D^{-1/2} = \sum_{i=1}^{k} \alpha_i \langle \mu_i, v \rangle D^{-1/2} V^T \mu_i \mu_i^T V D^{-1/2} = \sum_{i=1}^{k} \langle \mu_i, v \rangle u_i u_i^T.$$

Since $\langle \mu_1, v \rangle$ was assumed to be larger than all other $\langle \mu_i, v \rangle$, it follows that u_1 is the largest eigenvector of $D^{-1/2} V^T B V D^{-1/2}$. Now, if $w = V D^{1/2} u_1$ then $w = \sqrt{\alpha_1} \mu_1$.

Now, note that since the μ_i are linearly independent, there is a unique way to write $m = V V^T m = \sum_i \alpha_i \mu_1$ as $aw + y$, where y belongs to the span of $\{\mu_2, \dots, \mu_k\}$ (which is the same as the span of $\{V D^{1/2} u_i : i \geq 2\}$. Moreover, the unique choice of a that allows this representation must satisfy $aw = \alpha_1 \mu_1$, which implies that $a = \sqrt{\alpha_1}$. Therefore, $w/a = \mu_1$ and $a^2 = \alpha_1$. ■

The proof of Lemma 2 is crucial to understanding the algorithm, and also the broader message of this article: if we can get hold of two different normalizations of something, then we can learn something about it. In the proof of Lemma 2, this happens twice: first, we use the fact that A and B contain the same components (but with differing normalizations) to extract the span of a single component of interest. The differing normalization is crucial, because A by itself does not uniquely determine the set $\{\mu_1, \dots, \mu_k\}$, much less single out a specific component of interest.

In the second step of Lemma 2, we know $\sqrt{\alpha_1} \mu_1$, which is not enough to determine either α_1 or μ_1. However, we also have access to m, which involves a contribution of $\alpha_1 \mu_1$. Exploiting the difference between these two normalizations, we recover both α_1 and μ_1.

2.1.2 The Cancellation Method

Our second method avoids the matrix inversion in Algorithm 1, preferring a line search instead.

In the above Algorithm 2, we assume $\langle \mu_1, v \rangle > 0$. When this is not the case and B is a negative semi-definite matrix, we simply have to change the line search step to search for the smallest $\lambda < 0$ such that $\widehat{V} \widehat{V}^T (\widehat{A} - \lambda \widehat{B}) \widehat{V} \widehat{V}^T$ is PSD. Theorem 3 shows that with m, A, B estimated up to $O(\epsilon)$ error, the parameter estimation error in Algorithm 2 is also bounded as $O(\epsilon)$.

Theorem 3 *Suppose $\{\mu_1, \dots, \mu_k\}$ are linearly independent and v satisfies $\langle \mu_1, v \rangle \geq (1 + \delta) \langle \mu_i, v \rangle$ for all $i \neq 1$. Suppose that $\max\{\|\widehat{A} - A\|, \|\widehat{B} - B\|, \|\hat{m} - m\|\} < \epsilon$, and $\lambda_1 := 1/\langle \mu_1, v \rangle$. Then Algorithm 2 returns $\hat{\mu}_1, \hat{\alpha}_1$ with*

$$\begin{aligned} \|\hat{\mu}_1 - \mu_1\| &< \frac{C\epsilon}{\alpha_1^2 a_1^2} \left(\sigma_1(A) \left(1 + \frac{\alpha_1 a_1}{\sigma_{k-1}(Z_{\lambda_1})} \right) + \frac{\sigma_1(A) \eta_3 R}{\sigma_{k-1}(Z_{\lambda_1})} \right) \\ |\hat{\alpha}_1 - \alpha_1| &< \frac{C \sigma_1(A) \epsilon}{\alpha_1 a_1^3} \left(\eta_1 + \frac{\eta_2 R \eta_3}{\sigma_{k-1}(Z_{\lambda_1})} \right) \end{aligned}$$

Algorithm 2 Extracting a mixture component from side information: the cancellation method.

Input: $\hat{A}, \hat{B}, \hat{m}$
Output: $\hat{\mu}_1, \hat{\alpha}_1$

```
let $\widehat{V}$ be the $d \times k$ matrix of $k$ largest eigenvectors of $\hat{A}$;
search over $\lambda$ to find the largest $\lambda = \lambda^*$ such that $\widehat{V}\widehat{V}^T(\widehat{A} - \lambda\widehat{B})\widehat{V}\widehat{V}^T$ is PSD;
let $\widehat{Z}_{\lambda^*} = \widehat{A} - \lambda^*\widehat{B}$, and let $\{v_2, \dots, v_k\}$ be the top $k-1$ singular vectors of $\widehat{Z}_{\lambda^*}$
let $V_{1:(k-1)}$ be the $d \times (k-1)$ matrix with columns $\{v_2, \dots, v_k\}$
let $x_1 = \hat{m} - V_{1:(k-1)}V_{1:(k-1)}^T\hat{m}$
let $v_1 = x_1/\|x_1\|$
compute $c_i = v_1^T\widehat{A}v_i$ for $i = 1$ to $k$
let $a_i = c_i/\|x_1\|$ for $i = 1$ to $k$
return $\hat{\mu}_1 = \sum_{i=1}^k a_i v_i$ and $\hat{\alpha}_1 = c_1/a_1^2$
```

where $\eta_1 := \max\{\alpha_1 a_1(2a_1+1), 20\}$, $\eta_2 := \max\{\alpha_1 a_1^2, 10\}$, $\eta_3 = \max\{1, \lambda_1, \sigma_1(B)\}$, $R = \max\|\mu_i\|$, $a_1 = \|\mu_1 - \prod_{\mathcal{V}}\mu_1\|$, *where* $\mathcal{V} =$ *span*$\{\mu_2, \dots, \mu_k\}$, *and* C *is an universal constant.*

Again, we will defer the actual analysis to the appendix, and instead show that Algorithm 2 returns the exact answer when fed exact initial data. We will do this in two lemmas: Lemmas 4 and 5.

Lemma 4 *Let* $Z = \sum_{i=1}^k \gamma_i\mu_i\mu_i^T$ *where* $\{\mu_1, \dots, \mu_k\}$ *are linearly independent,* $\mu_i \in \mathbb{R}^d, \gamma_i \in \mathbb{R}$ *and* $d > k$*. If* $\gamma_1 < 0$ *and* $\gamma_i > 0$ *for all* $i \neq 1$ *then* Z *is not positive semi-definite.*

Proof Let Π denote the projection onto the orthogonal complement of span$\{\mu_2, \dots, \mu_k\}$. Let $x = \Pi\mu_1$, and note that $\langle x, \mu_1\rangle > 0$ but $\langle x, \mu_i\rangle = 0$ for all $i \neq 1$. Hence, $x^TZx = \gamma_1\langle x, \mu_1\rangle^2 < 0$ and so Z is not positive semi-definite. ■

Lemma 5 *Let* m*,* A*, and* B *be defined by in* (1), (2), *and* (3)*, where* $\mu_1, \dots, \mu_k$ *are linearly independent. If* $\langle\mu_1, v\rangle > \langle\mu_i, v\rangle$ *for all* $i \neq 1$ *and we apply Algorithm 2 to* A*,* B*, and* m*, then it returns* μ_1 *and* α_1*.*

Proof Define $w_i = \langle\mu_i, v\rangle$ and let $\gamma_i = \alpha_i(1 - \lambda w_i)$, so that

$$Z_\lambda = A - \lambda B = \sum_{i=1}^k \gamma_i\mu_i\mu_i^T.$$

Note that, in our case where $\widehat{A} = A$, and $\widehat{B} = B$, columns of $\widehat{V}$ simply form a common orthonormal bases of the row/column space of both matrices A, B. Therefore the matrix $\widehat{V}\widehat{V}^T(A - \lambda B)\widehat{V}\widehat{V}^T = A - \lambda B = Z_\lambda$. Now for $\lambda > \frac{1}{w_1}$, $\gamma_1 < 0$ and for all $\lambda \leq \frac{1}{w_1}$, $\gamma_i \geq 0$ for all i since $w_1 > w_i$, for every $i \neq 1$. By Lemma 4, $\lambda^* = \frac{1}{w_1}$ is the largest λ such that Z_λ is PSD; hence,

$$Z_{\lambda^*} = \sum_{i=2}^k \alpha_i(1 - \lambda^* w_i)\mu_i\mu_i^T.$$

From Lemma 26 in Appendix E.2 it follows that $k-1$ singular vectors $\{v_2, \ldots, v_k\}$ of Z_{λ^*} form a basis of the subspace $\mathcal{V} = \text{span}\{\mu_2, \ldots, \mu_k\}$. Let $\mathcal{V}_\perp$ be the perpendicular space of $\mathcal{V}$, and write $\Pi = I - V_{1:(k-1)}V_{1:(k-1)}^T$ for the orthogonal projection onto $\mathcal{V}_\perp$. Since $\Pi\mu_i = 0$ for $i \neq 1$, we have $x_1 = \Pi m = \alpha\Pi\mu_1$.

Now define $b_1, \ldots, b_k$ by $\mu_1 = \sum_{i=1}^k b_i v_i$. In order to prove that the algorithm returns μ_1 correctly, we need to show that $b_i = a_i := c_i/\|x_1\|$. Indeed,

$$c_i := v_1^T A v_i = \sum_{j=1}^k \alpha_j v_1^T \mu_j \mu_j^T v_i = \alpha_1 b_1 b_i,$$

since $v_1^T \mu_j = 0$ for $j \neq 1$. On the other hand, $\|x_1\| = \alpha\|\Pi\mu_1\| = \alpha b_1$, and so $b_i = a_i$, as claimed. Moreover, $\hat{\alpha}_1 = \frac{c_1}{a_1^2} = \alpha_1$, as claimed. ■

Optimization for λ^*: The first step of Algorithm 2 involves finding a smallest λ^* such that $\widehat{Z}'_{\lambda^*} = \widehat{V}\widehat{V}^T(\widehat{A} - \lambda^*\widehat{B})\widehat{V}\widehat{V}^T$ is PSD using line search. Although $\widehat{Z}'_\lambda$ is a $d \times d$ matrix, this step can be performed efficiently as follows. Instead of searching for λ directly for $\widehat{Z}'_\lambda$, we do this for a smaller $k \times k$ matrix $\widehat{V}^T\widehat{Z}'_\lambda\widehat{V} = \widehat{V}^T(\widehat{A} - \lambda^*\widehat{B})\widehat{V}$. This optimization step using line search can be performed in just $O(k^3 \log|\lambda^*|)$ time.

3. Specific Models

In this section we discuss how the search algorithms can be applied in four specific mixture models.

3.1 Gaussian Mixture Model with Spherical Covariance

The model: Besides the mixture parameters $\alpha_1, \ldots, \alpha_k$, the Gaussian mixture model (GMM) has mean parameters $\mu_1, \ldots, \mu_k \in \mathbb{R}^d$ and variance parameters $\sigma_1, \ldots, \sigma_k \in \mathbb{R}$. The conditional densities $g(\cdot; \mu_i, \sigma_i)$ are Gaussian, with mean μ_i and covariance $\sigma_i^2 I_d$. Explicitly,

$$g(x; \mu_i, \sigma_i) = \frac{1}{(2\pi\sigma_i^2)^{d/2}} e^{-\frac{\|x-\mu_i\|^2}{2\sigma_i^2}}.$$

Matrices A and B: We fix a vector $v \in \mathbb{R}^d$, with the assumption that $\langle v, \mu_1\rangle > \langle v, \mu_i\rangle$ for $i \neq 1$. Recall (from Section 2.1) that $m = \mathbb{E}[x] = \sum_i \alpha_i\mu_i$, $A = \sum_{i=1}^k \alpha_i\mu_i\mu_i^T$, and $B = \sum_{i=1}^k \alpha_i\langle\mu_i, v\rangle\mu_i\mu_i^T$. To compute these quantities, we first define σ^2 to be the $(k+1)$th-largest eigenvalue of the mixture covariance matrix $\mathbb{E}[(x-m)(x-m)^T]$, and let u be a corresponding eigenvector. Then let $\widetilde{m} = \mathbb{E}[x(u^T(x-m))^2]$. Then it follows from moment computations (see Hsu and Kakade (2013)) that:

$$\begin{aligned} A &= \mathbb{E}[xx^T] - \sigma^2 I_d \\ B &= \mathbb{E}[\langle x, v\rangle xx^T] - \widetilde{m}v^T - v\widetilde{m}^T - \langle\widetilde{m}, v\rangle I_d, \end{aligned}$$

Given the samples $\{\hat{x}_i\}$, we can now empirically evaluate these quantities (denoted by $\hat{m}, \hat{A}, \hat{B}$ respectively) by replacing expectations above by the corresponding sample averages; for instance we replace $\mathbb{E}[xx^T]$ by $\widehat{\mathbb{E}}[xx^T] \doteq (1/n)\sum_{j=1}^n \hat{x}_j\hat{x}_j^T$.

Examples of v: Assuming that $\|\mu_1\|^2 > \langle \mu_1, \mu_i \rangle$ for all $i \neq 1$—this will be true, for example, if $\|\mu_i\|$ are all the same—one can find a suitable vector v given a relatively small number of samples from the first mixture component. Specifically, if $\|\mu_1\|^2 \geq \langle \mu_1, \mu_i \rangle + \delta$ and $\|\mu_i\| \leq R$ for all $i \neq 1$ then standard Gaussian tail bounds imply the following: if $v := \ell^{-1} \sum_{j=1}^{\ell} x_j$ where $\ell = \Omega(R^2 \delta^{-2} \log k)$ and $x_1, \ldots, x_m$ are drawn independently from the distribution $g(\cdot; \mu_1, \sigma_1)$ then with high probability v satisfies $\langle v, \mu_1 \rangle > \langle v, \mu_i \rangle$ for all $i \neq 1$. Here, "high probability" means probability converging to 1 as the hidden constant in $\ell = \Omega(\cdot)$ grows. Note here that the number of tagged samples is nowhere near sufficient to estimate μ_1 by direct averaging; indeed to do so would require the number of samples to grow with the size of the underlying dimension.

Remarks: We note that spectral algorithms which uses the whitening procedure has been proposed before in the context of GMM e.g. Hsu and Kakade (2013). The primary difference between the algorithm in Hsu and Kakade (2013) and Algorithm 1 is that the former, in absence of side information, takes a projection of the third order moment tensor M_3 on a random unit vector to obtain the second matrix, where as our matrix B can be viewed as a projection of M_3 on the side information vector v. The main advantage of projecting onto v is that, when we have reliable side information, this will give a good singular value separation resulting in better empirical performance. The Cancellation algorithm however is distinctly different from both and has not been studied before.

3.2 Latent Dirichlet Allocation

The model: In the LDA model with k topics and a dictionary of size d, the parameters $\mu_1, \ldots, \mu_k \in \Delta_{d-1}$ are the probability distributions corresponding to each topic (Δ_{d-1} denotes the probability simplex $\{y \in \mathbb{R}^d : \sum_i y_i = 1, \min_i y_i \geq 0\}$). The LDA model introduced in Blei et al. (2003) differs slightly from the other models as the mixture distribution cannot be expressed exactly in the parametric form in Section 2. Instead we have a two level hierarchy as follows. Given $\bar{\alpha} = (\alpha_1, \ldots, \alpha_k)$, we first draw a topic distribution θ from the Dirichlet($\bar{\alpha}$) distribution. Given this $\theta = (\theta_1, \ldots, \theta_k)$ each word in the document is drawn i.i.d. from the distribution $\sum_{i=1}^{k} \theta_i \mu_i$. However still we can compute the vector m and the matrices A, B as shown below. Then with an appropriate v our algorithms can recover the topic distribution μ_1.

Matrices A and B: Let x_1 denote the random vector with $x_1(w) = 1$ if the first word is w, and 0 otherwise. Similarly define vectors x_2, x_3 corresponding to the second and third word respectively, and let $\alpha_0 = \sum_{i=1}^{k} \alpha_i$. Then, moment computations under the LDA distribution yields the following expressions for (m, A, B), defined in (1), (2), (3):

$$
\begin{aligned}
&m = \alpha_0 \mathbb{E}[x_1], \quad A = \alpha_0(\alpha_0 + 1)\mathbb{E}[x_1 x_2^T] - mm^T \\
&B = \frac{\alpha_0(\alpha_0+1)(\alpha_0+2)}{2} \mathbb{E}[\langle x_3, v \rangle x_1 x_2^T] - \frac{\alpha_0(\alpha_0+1)}{2} \big(\langle m, v \rangle \mathbb{E}[x_1 x_2^T] + \mathbb{E}[\langle x_3, v \rangle x_1 m^T] \\
&\quad + \mathbb{E}[\langle x_3, v \rangle m x_2^T] \big) + \langle m, v \rangle mm^T.
\end{aligned}
$$

With the given document samples, let $\hat{x}_i$ denote the normalized empirical word frequencies in the document i. Then, $\hat{m} = \frac{\alpha_0}{n} \sum_{i=1}^{n} \hat{x}_i$, and $\widehat{A}, \widehat{B}$ can be immediately estimated using the above expressions by replacing expectations with sample averages.

Using labeled words to find v: In order to recover the topic distribution μ_1 we now require a vector v which satisfies $\langle \mu_1, v\rangle > \langle \mu_i, v\rangle$ for $i \neq 1$. Now suppose we are given a *labeled word* ℓ such that its occurrence probability in topic 1 is the highest, i.e., $\mu_1(\ell) > \mu_i(\ell)$ for $i \neq 1$ (note that this does not mean ℓ is the most frequent word in topic 1, there may be words with higher occurrence probability in this topic). Then we can simply choose $v = e_\ell$ (the standard basis element with 1 in the ℓ-th coordinate). For most topics of practical interest it is possible to find such labeled words. For example the word "ball" can be a labeled word for topic sport, "party" is a labeled word for topic politics and so on. However, a labeled word is merely indicative of a topic and is not exclusive to a topic (e.g. the word "ball" can occur in other contexts as well). In this sense, the labelled word is quite different from the "anchor word" described in Arora et al. (2013). Note however that anchor words are also labeled words (but *not* vice-versa) since for an anchor word ℓ, $\mu_1(\ell) > 0$ and $\mu_i(\ell) = 0$ for $i \neq 1$.

Using labeled documents to find v: If the different topics are not too similar, then we can estimate a suitable vector v from a small collection of documents that are mostly about the topic of interest. For example, if $\langle \mu_i, \mu_j\rangle \leq \eta\|\mu_i\|\|\mu_j\|$ for all $i \neq j$, and if we observe a total of m words from some collection of documents with $\theta_1 \geq (1+\delta)(1/2+\eta)$ then about $m = \Omega(\delta^{-2}\log k)$ words will suffice to find a suitable vector v.

Remarks: Similar to the case of GMM, a spectral algorithm using whitening procedure to estimate LDA components have been presented before in Anandkumar et al. (2012). Again the main difference with our Whitening algorithm being the fact that in Anandkumar et al. (2012) the second matrix is constructed by taking a random projection of the third order moment tensor $Triples$, and in Algorithm 1 this is constructed as a projection onto v. Empirically this results is a more stable algorithm due to guaranteed singular value separation. The Cancellation algorithm has not been previously studied in LDA model.

3.3 Mixed Regression

The model: In mixed linear regression the mixture samples generated are of the form $y = \langle x, \mu_i\rangle + \xi$, where $x \sim \mathcal{N}(0, I)$ and noise $\xi \sim \mathcal{N}(0, \sigma^2)$. As before, a sample is generated using the i-th linear component μ_i, with probability α_i. We have access to the observations (y, x) but the particular μ_i and ξ are unknown. Hence the conditional density $g(x, y; \mu_i, \sigma)$ is a multivariate Gaussian where $x \sim \mathcal{N}(0, I)$, $y \sim \mathcal{N}(0, \|\mu_i\|^2 + \sigma^2)$, and $\mathrm{Cov}(x, y) = \mu_i$.

Matrices A and B: To compute A and B, we consider the following moments (for more detailed derivations, see Appendix C):

$$
\begin{aligned}
M_{1,1} &= \mathbb{E}[yx] = \sum_{i=1}^{k} \alpha_i \mu_i \\
M_{2,2} &= \mathbb{E}[y^2 x x^T] = 2\sum_{i=1}^{k} \alpha_i \mu_i \mu_i^T + \sum_{i=1}^{k} \alpha_i (\sigma^2 + \|\mu_i\|^2) I \\
M_{3,1} &= \mathbb{E}[y^3 x] = 3\sum_{i=1}^{k} \alpha_i (\sigma^2 + \|\mu_i\|^2) \mu_i \\
M_{3,3} &= \mathbb{E}[y^3 \langle x, v\rangle x x^T] = 6\sum_{i=1}^{k} \alpha_i \langle \mu_i, v\rangle \mu_i \mu_i^T + \left(M_{3,1} v^T + v M_{3,1}^T + \langle M_{3,1}, v\rangle I\right)
\end{aligned}
$$

Let τ^2 be the smallest singular value of the matrix $M_{2,2}$. Then we can compute m, A, B as follows.

$$
\begin{aligned}
m &= M_{1,1}, \quad A = \frac{1}{2}(M_{2,2} - \tau^2 I) \\
B &= \frac{1}{6}(M_{3,3} - (M_{3,1} v^T + v M_{3,1}^T + \langle M_{3,1}, v\rangle I))
\end{aligned}
$$

As in the previous cases with finite samples the estimates $\hat{m}, \widehat{A}, \widehat{B}$ can be computed by taking their empirical expectations e.g., $\widehat{M}_{1,1} = \widehat{\mathbb{E}}[yx] = \frac{1}{n}\sum_{i=1}^{n} \hat{y}_i \hat{x}_i$ and so on, where $(\hat{y}_i, \hat{x}_i)$ denote the i-th sample.

Examples of v: Suppose we are given a few random labeled examples from the first component. Then assuming $\|\mu_1\|^2 > \langle \mu_1, \mu_i\rangle + \delta$, $\|\mu_i\|^2 \le R$, similar to the GMM case we can estimate a $v := \frac{1}{\ell}\sum_{j=1}^{\ell} \hat{y}_j \hat{x}_j$ using only $\ell = \Omega\left(R^4 \delta^{-2} \log k\right)$ labeled samples so that $\langle \mu_1, v\rangle > \langle \mu_i, v\rangle$ holds with high probability.

Remarks: Our construction of the second matrix B is a consequence of some new moment results for the mixed linear regression model. We present these detailed moment derivations in Appendix C.4. This also results in improved sample complexity bounds over previous moment based algorithms (discussed in Section 3.5).

3.4 Subspace Clustering

The model: Besides the mixture parameters $\alpha_1, \ldots, \alpha_k$, the subspace clustering model has parameters $U_1, \ldots, U_k \in \mathbb{R}^{d\times m}$ and $\sigma \in \mathbb{R}$, where the matrices $U_1, \ldots, U_k$ have orthonormal columns. The conditional distribution $g(\cdot; U_i)$ is a standard Gaussian variable supported on the column space of U_i, plus independent Gaussian noise. More precisely, we sample $y \sim \mathcal{N}(0, I_d)$ and set $x = U_i U_i^T y + \xi$, where $\xi \sim \mathcal{N}(0, \sigma^2 I_d)$ is independent of y.

Matrices A and B: The subspace clustering model does not quite fit into the basic method of Section 2; one motivation for presenting it is to show that the basic ideas in Section 2 are more flexible than they first appear. Suppose $v \in \mathbb{R}^d$ satisfies $\|U_1^T v\| > \|U_i^T v\|$

for all $i \neq 1$. We consider

$$\begin{aligned} A &:= \mathbb{E}[xx^T] - \sigma^2 I_d = \sum_{i=1}^k \alpha_i U_i U_i^T \\ B &:= \mathbb{E}[\langle x, v\rangle^2 xx^T] - \sigma^2 v^T A v I_d - \sigma^2 \|v\|^2 A - \sigma^4(\|v\|^2 I_d + vv^T) - 2\sigma^2(Avv^T + vv^T A) \\ &= \sum_{i=1}^k \alpha_i \|U_i^T v\|^2 U_i U_i^T + 2\sum_{i=1}^k \alpha_i U_i U_i^T vv^T U_i U_i^T \end{aligned}$$

and their empirical versions $\hat{A}$ and $\hat{B}$ (the computation giving the claimed formula for B is carried out in Appendix C). Now with these $\hat{A}$ and $\hat{B}$, we can recover the subspace U_1 using Algorithm 3. This algorithm uses the same principle behind the whitening method in Section 2.1.1, the key difference is that here we pick the top m eigenvectors of the whitened B matrix.

Algorithm 3 Subspace clustering algorithm

Input: $\hat{A}, \hat{B}$

Output: $\hat{U}$

1: let $\{\sigma_j, v_j\}$ be the singular values and singular vectors of $\hat{A}$, in non-increasing order
2: let V be the $d \times mk$ matrix whose jth column is v_j
3: let D be the $mk \times mk$ diagonal matrix with $D_{jj} = \sigma_j$
4: let $Y = [u_1, \ldots, u_m]$ be the matrix of m largest eigenvectors of $D^{-1/2} V^T \hat{B} V D^{-1/2}$
5: let $Z = V D^{1/2} Y$
6: let the columns of $\hat{U}$ be the m eigenvectors of the matrix ZZ^T

The following perturbation theorem guarantees that if the side information vector v is substantially more aligned with the subspace spanned by U_1 than it is with any other subspace, and the matrices A, B are estimated within ϵ accuracy, then Algorithm 3 can recover the required subspace with a small error.

Theorem 6 *Suppose that $\|\hat{A} - A\| \leq \epsilon$ and $\|\hat{B} - B\| \leq \epsilon$. Suppose that the side information vector v satisfies $\|U_i v\|^2 \leq (1/3 - \delta)\|U_1 v\|^2$. Then output $\hat{U}$ of Algorithm 3 satisfies*

$$\|\hat{U}\hat{U}^T - U_1 U_1^T\| \leq C\epsilon\alpha_1^{-1}\sigma_1(A)^2 \sigma_{mk}(A)^{-2}\delta^{-1}.$$

We prove Theorem 6 in Appendix F. Note that the conditions on v can be satisfied if the spaces U_i satisfy a certain affinity condition and we have a few labelled samples from U_1. Specifically, suppose that $\langle u, w\rangle < (\frac{1}{\sqrt{3}} - \eta)\|u\|\|w\|$ for every $u \in U_1$ and $w \in U_i$, $i \neq 1$. Then any $v \in U_1$ will satisfy the assumption of Theorem 6. Hence, a single labelled sample from U_1 (or several—depending on η—noisy samples) is enough to find a suitable v.

Remarks: To the best of our knowledge Algorithm 3 is the first moment based algorithm for the subspace clustering model. The detailed moment derivations are presented in Appendix C.5. Also our generative model allows samples to be noisy, hence they do not lie exactly on the subspace but close to it. Such a setting has not been considered in most subspace clustering literature.

3.5 Comparison

In this section we compare the theoretical performance of the Whitening and Cancellation algorithms with other algorithms. Both Whitening and Cancellation algorithms require estimating the quantities m, A, B by computing moments from the samples. Therefore the sample complexity primarily depends on how well these quantities concentrate. We compute the specific sample complexities for each model in Appendix G.

For Gaussian mixture model the sample complexity of our algorithm scales as $\tilde{\Omega}(d\epsilon^{-2}\log d)$ similar to moment based algorithm by Hsu and Kakade (2013) and tensor decomposition based algorithm by Anandkumar et al. (2014). In terms of runtime the Whitening algorithm is faster than the tensor decomposition based algorithm by Anandkumar et al. (2014). This can be viewed as follows. The first step in both the algorithms take $O(d^2k)$ time to compute the whitening matrix and in subsequent whitening steps. However, computing the largest eigenvector in Algorithm 1 takes only $O(k^2)$ time, faster than $O(k^5 \log k)$ time required for rank-k tensor power iteration (we also verify this in our experiments in Section 4).

In LDA topic model our algorithms have a sample complexity of $\tilde{\Omega}(\epsilon^{-2}\log d)$, again similar to tensor decomposition based algorithm by Anandkumar et al. (2014), and non-negative matrix factorization (NMF) based algorithm by Arora et al. (2013). The Whitening algorithm again is faster than tensor decomposition as argued for GMM case. The NMF based algorithm using optimization based RecoverKL/RecoverL2 procedures also has a runtime of $O(d^2k)$ similar to our algorithms (in Section 4 again we observe our algorithm to be faster in practice). The spectral topic modeling algorithm in Anandkumar et al. (2012) also has a computation complexity $O(d^2k)$ similar to our algorithms. However, its sample complexity has a high $\Omega(k^5)$ dependence on the number of components. This spectral algorithm also suffer from instability in practice due to the random projection step (as noted in Anandkumar et al. 2014).

In the case of mixed linear regression again our method has a sample complexity of $\tilde{\Omega}(d\epsilon^{-2}\log d)$ similar (upto log factors) to the convex optimization based approach by Chen et al. (2014), alternating minimization based approach by Yi et al. (2014), but better than tensor decomposition based method of Sedghi et al. (2016) which has a sample complexity of $\tilde{\Omega}(d^3\epsilon^{-2})$. However unlike the convex optimization and alternating minimization based techniques our method is also applicable when the number of components $k > 2$. As argued in GMM case the Whitening algorithm is again faster than the tensor algorithm by Sedghi et al. (2016).

Subspace clustering algorithms like greedy subspace clustering by Park et al. (2014), optimization based algorithms by Elhamifar and Vidal (2009), Soltanolkotabi and Candes (2012), requires the samples to exactly lie on a subspace. In contrast our moment based algorithm works even when the samples are noisy and perturbed from the actual subspace. Our subspace clustering algorithm also has a sample complexity of $\tilde{\Omega}(m\epsilon^{-2}\log d)$ which is similar (up to log factors) to greedy subspace clustering algorithm by Park et al. (2014).

We note that it is possible to use approximation methods like randomized svd to further speed up the Whitening, Cancellation and tensor decomposition based algorithms by Anandkumar et al. (2014), however this will result in decreased accuracy in both algorithms. We refer to Huang et al. (2015) for such stochastic optimization, and parallelization techniques used to speed up the tensor algorithms.

In a setting where side information is provided on each of the k components, observe that we can run the Whitening algorithm independently for each of the k components, possibly in parallel. Hence we can recover all k components, without loosing the runtime advantage of the Whitening algorithm. We demonstrate this application on real data set in Section 4.2. In terms of the overall computation time, it can be shown that running the Whitening algorithm for all k components is still faster than the tensor decomposition based algorithm by Anandkumar et al. (2014), when $k = \Omega(n^{\frac{1}{3}} d^{\frac{1}{3}})$.

4. Experiments

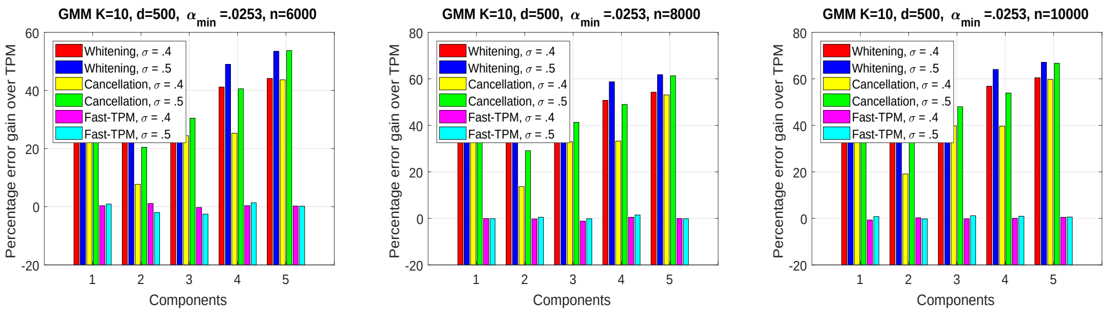
In this section we present the empirical performance of our Whitening, Cancellation, and Subspace clustering algorithms. We consider three of the settings: the Gaussian Mixture Model (GMM), and Latent Dirichlet Allocation (LDA), and Subspace clustering, and validate our algorithms on both real and synthetic data sets.

4.1 Synthetic Data Set

First we compare the sample complexity and runtime of our algorithms with the robust tensor decomposition algorithm by Anandkumar et al. (2014), which is based on tensor power iteration, for learning mixture models (we refer to this as the TPM algorithm). Our second baseline algorithm is a faster heuristic of TPM where we start the tensor power iterations initialized with side information vector v, and recover just the first component. We refer this as the Fast-TPM algorithm. For the Cancellation algorithm we compute the optimum λ for cancellation using two different techniques as follows. First, let $\widehat{Z}'_\lambda = V^T \widehat{Z}_\lambda V$, where V is the matrix of top k singular vectors of $\widehat{A}$. In the first method, we perform a line search over positive λ to find the minimum λ such that $\sigma_k(\widehat{Z}'_\lambda)$ falls below certain threshold. This method works well in GMM case. In a second method we minimize the convex function $\|\widehat{Z}'_\lambda\|_* + \lambda$, subject to $\lambda \geq 0$. This method performs better in the case of LDA. Note that for the Cancellation algorithm after estimating λ, instead of using m and A to find μ_1 we can follow the same steps using $m' = Av$ and B to recover μ_1. Theoretically it has the same performance, however empirically we observe this to work slightly better and we use this version for our experiments. We implement all algorithms for our synthetic data experiments using MATLAB.

Performance metric: We compute the estimation error of parameter μ_1 as $\mathcal{E} = \|\hat{\mu}_1 - \mu_1\|$. In our figures we plot the quantity "percentage relative error gain" which is defined as $G = 100(\mathcal{E}_T - \mathcal{E}_A)/\mathcal{E}_T$, where $\mathcal{E}_T$ is the TPM error and $\mathcal{E}_A$ is the error for Whitening / Cancellation / Fast-TPM algorithm. Note that a positive error gain implies that the TPM error is greater than that of the competing algorithm. In the subspace clustering model we plot similar percentage relative error gain over the baseline k-means algorithm.

Gaussian mixture model: We generate synthetic data sets for GMM with different k, d, α_i, σ, and v. Figure 1 shows the percentage relative error gains of the Whitening, Cancellation, and Fast-TPM algorithms over the TPM algorithm in a GMM with various values of k, d, α_i, σ, and n. The μ_i were generated randomly over the sphere of norm $r = 10$. We define $\alpha_{min} := \min_i \alpha_i$. The side information vector v was chosen as follows. Let $\{v_1, \ldots, v_k\}$ be a orthonormal basis of $\text{span}\{\mu_1, \ldots, \mu_k\}$, such that $\{v_2, \ldots, v_k\} \in \text{span}\{\mu_2, \ldots, \mu_k\}$. Then we



(a) (b) (c)

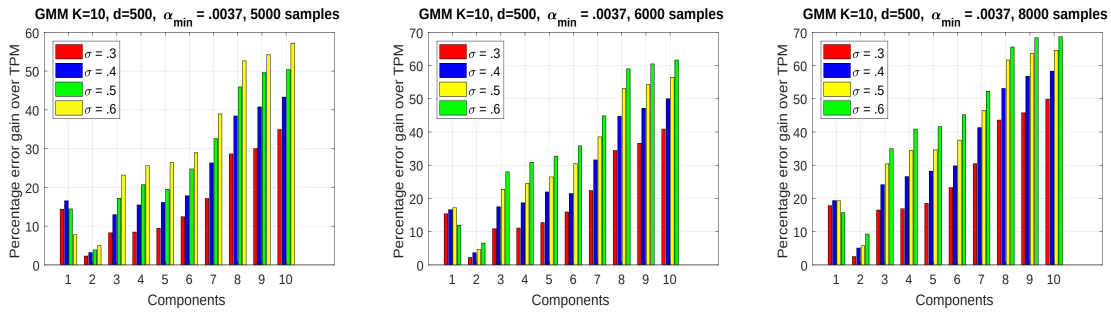
Figure 1: Figure showing the percentage relative error gain by the Whitening, Cancellation, and Fast-TPM algorithm over the TPM algorithm for 5 components of increasing size, in a GMM with $k = 10, d = 500, \sigma \in \{.4, .5\}$, and three different sample complexities (a) $n = 6000$ (b) $n = 8000$ (c) $n = 10000$. Our algorithms shows increasingly better gain over TPM and Fast-TPM as α_i, σ and n increase.

choose $v = \sqrt{\gamma} v_1 + \sqrt{(1-\gamma)/(k-1)} \sum_{i=2}^{k} v_i$ for some $\gamma \in (0, 1)$ such that the condition $\langle \mu_1, v\rangle > \langle \mu_i, v\rangle$ is satisfied. We observe that in all the cases, our algorithms have lower error (positive error gain) than both the tensor algorithms. Moreover, our methods' advantage increases with increasing proportion α_i, increasing sample size n, and increasing variance σ. We also observe that the Fast-TPM algorithm has the same error performance as TPM (error gain close to zero).

Figure 2 gives an example where the Whitening algorithm can successfully recover even rare components. Here we consider a GMM with $k = 10, d = 500$ with the rarest component having probability $\alpha_{min} = .0037$. Again we observe positive relative error gains over TPM algorithm for increasing number of samples n.

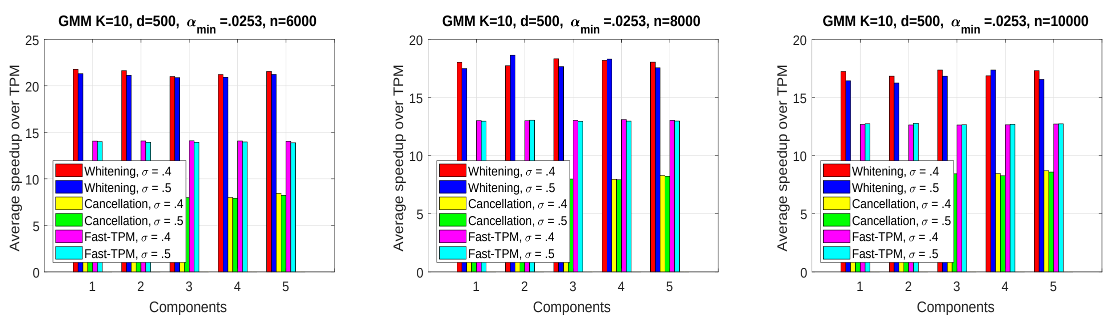
In Figure 3 we plot the speedup of the algorithms over TPM, and observe that the Whitening and Cancellation algorithms are much faster (high speedup) than the TPM algorithm. We also observe that the Fast-TPM algorithm is faster than TPM and Cancellation algorithms, but slower than Whitening algorithm. Note that, while it is also possible to speed up the basic TPM algorithm compared here using techniques such as randomized svd and stochastic tensor gradient descent [Huang et al. 2015], such approximate methods will reduce the overall accuracy. Moreover the randomized svd techniques can also be applied to the search algorithms presented in this paper, to obtain further speedups.

Topic Modeling: We generate a synthetic LDA document corpus according to the model in Blei et al. (2003). The lengths of the documents are generated using a Poission(L) distribution where L is the mean document length. In Figure 4 we plot the percentage relative error gain of the Whitening, Cancellation, and Fast-TPM algorithms over the TPM algorithm. Our side information was a labeled word w satisfying $\mu_1(w) > \mu_i(w)$ for $i \neq 1$. Again we observe positive error gains over the TPM algorithm. Although the Fast-TPM algorithm sometimes perform better than TPM for more frequent topics, the Whitening



(a) (b) (c)

Figure 2: Figure showing the percentage relative error gain of the Whitening algorithm over the TPM algorithm in presence of rare components ($\alpha_{min} = .0037$), for a GMM with $k = 10, d = 500, \sigma \in \{.3, .4, .5, .6\}$, and number of samples (a) $n = 5000$ (b) $n = 6000$ (c) $n = 8000$. The Whitening algorithm recovers even the rarest component with increasing error gain over TPM as the number of samples increase.



(a) (b) (c)

Figure 3: Figure showing the average speedup of Whitening, Cancellation, and Fast-TPM algorithms over TPM, for 5 components of increasing size, in a GMM with $k = 10, d = 500, \sigma \in \{.4, .5\}$, and three different sample complexities (a) $n = 6000$ (b) $n = 8000$ (c) $n = 10000$. The Whitening algorithm is the fastest.

algorithm still outperforms it. Note that the performance varies across topics since the probability of the labeled word is different for each topic.

Subspace Clustering: We generate synthetic data for the subspace clustering model described in section 3.4 using parameters $d = 500, k = 5, m = 10$, and $\alpha_i \in [.1, .3]$. First we generate $k = 5$ random subspaces with orthonormal basis $\{U_i\}_{i=1}^k$, each of dimension $m = 10$. Then we generate random points on these subspaces, and add white Gaussian perturbations with $\sigma \in \{.1, .2\}$. We choose the side information vector v similar to the sensitivity experiment in GMM, and ensuring $\|U_1^T v\| > \|U_i^T v\|$, for $i \neq 1$. Note that due to the added Gaussian noise, our samples do not lie exactly on the subspaces $\{U_i\}_{i=1}^k$, but close

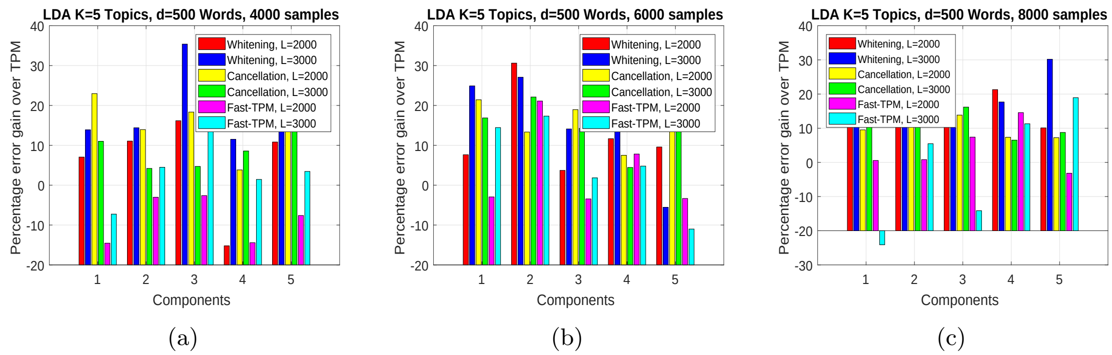


Figure 4: Figure showing the percentage relative error gain in each component of the Whitening, Cancellation, and Fast-TPM algorithms over the TPM algorithm in an LDA model with $k = 5, d = 500$, mean document length $L \in \{2000, 3000\}$, and number of documents (a) $n = 4000$ (b) $n = 6000$ (c) $n = 8000$. The Whitening algorithm show an improvement over TPM and Fast-TPM with increasing samples.

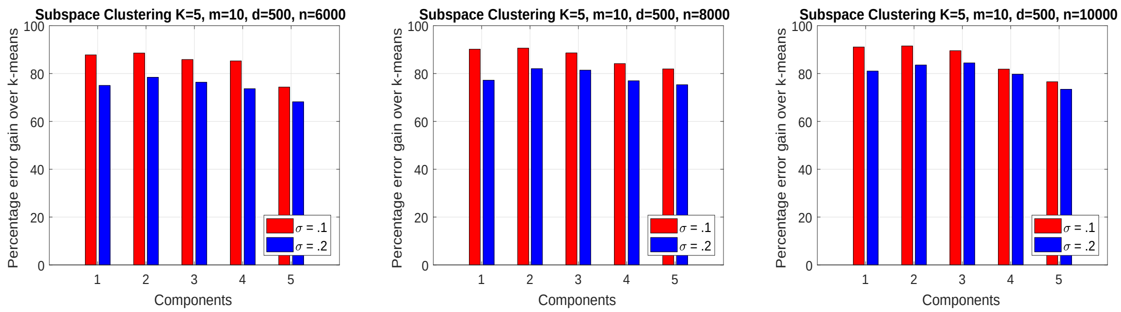
to it. Traditional subspace clustering algorithms, which assume points to lie exactly on the subspace, may not perform well. The TPM algorithm is also not well suited for this model since (a) the required moment tensor will be of 4^{th} order resulting in high computation cost (b) even if mk basis of the tensor are recovered, finding the target subspace will involve a further combinatorial search of $\binom{mk}{m}$ subspaces and finding the one having the strongest projection of v. Therefore we choose the k-means algorithm as our baseline for this model and compare with Algorithm 3. First we compute k clusters using k-means, then we find an m dimensional basis for each cluster using svd, finally we choose the target subspace as the one having the largest projection of v. If $\widehat{U}_1$ is the estimated orthonormal basis for the target subspace U_1, we compute the error as $\mathcal{E} = \|\widehat{U}_1 \widehat{U}_1^T - U_1 U_1^T\| / \|U_1 U_1^T\|$.

Figure 5 shows that Algorithm 3 has a much better error performance over k-means. In the speedup plots in Figure 6 we also observe that our subspace search algorithm is over $4X$ times faster than k-means.

4.2 Real Data Sets

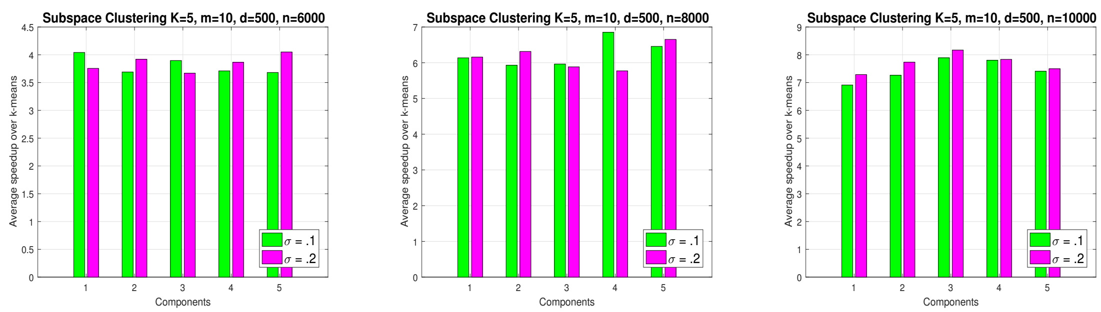
Topic Modeling: In this section we compare the performance of Whitening algorithm with a recent non-negative matrix factorization based topic modeling algorithm by Arora et al. (2013) (we refer this as NMF algorithm), and also the semi-supervised version of this NMF algorithm (we refer to this as SS-NMF). We test on two real large data sets; (a) New York Times news article data set [UCI 2008] (300,000 articles) (b) Yelp data set of business reviews [Yelp 2014] (335,022 reviews). We run both algorithms for $k = 100$ topics. For this experiment we do not consider the TPM algorithm by Anandkumar et al. (2014) since its runtime with $k = 100$ topics becomes extremely large on these data sets.[1]

1. To be more precise, with just $k = 10$ topics, the tensor algorithm takes 908 seconds in NY Times data set, compared to just 188 seconds for the Whitening algorithm (using MATLAB).



(a) (b) (c)

Figure 5: Figure showing the percentage relative error gain by our subspace search algorithm (Algorithm 3) over k-means for 5 components of increasing size, in a subspace clustering model with $k = 5, m = 10, d = 500, \sigma \in \{.1, .2\}$, and three different sample complexities (a) $n = 6000$ (b) $n = 8000$ (c) $n = 10000$. Our algorithm shows much better error performance than k-means.



(a) (b) (c)

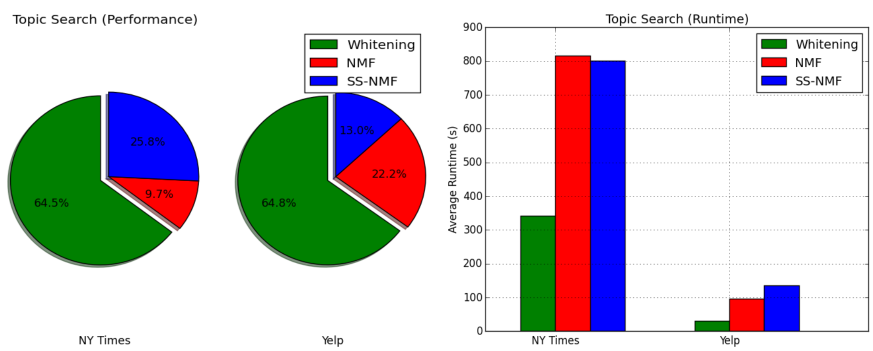
Figure 6: Figure showing the average speedup of our subspace search algorithm (Algorithm 3) over k-means, for 5 components of increasing size, in a subspace clustering model with $k = 5, m = 10, d = 500, \sigma \in \{.1, .2\}$, and three different sample complexities (a) $n = 6000$ (b) $n = 8000$ (c) $n = 10000$. Our subspace clustering algorithm shows high speedup over k-means.

In contrast, the NMF algorithm is known to be faster, and produce topics of comparable quality to more popular variational inference based algorithms [Blei et al. 2003]. The side information for this experiment are chosen as follows. First from the set of topics produced by NMF algorithm we choose a subset of interpretable topics, then we choose labeled words representative of these topics. We test with a set of 62 labeled words for NY Times data set and 54 labeled words for Yelp data set. Note that given labeled word w_l the whitening algorithm produces one topic distribution μ_1, but the NMF algorithm finds k topics. Therefore for NMF algorithm the target topic i is the one which has the highest probability of the labeled word i.e., $\mu_i(w_l)$. For the semi-supervised NMF we first compute the weighted word-word co-occurrence matrix Q_w where we re-weigh each document by the normalized frequency of the labeled word w_l. Then we apply the NMF algorithm [Arora et al. 2013] on this weighted matrix Q_w. All three algorithms were implemented in Python.

Performance metric: We compare the quality of the topics returned by Whitening, NMF, and SS-NMF algorithms using the pointwise mutual information (PMI) score, known to be a good metric for topic coherence [Newman et al. 2010; Röder et al. 2015]. However in order to also capture the relevance of the estimated topic to the labeled word we compute PMI score for topic i as,

$$PMI(\text{topic i}) = \frac{1}{20} \sum_{w \in \mathcal{T}_{20}^i} \log \frac{p(w_l, w)}{p(w_l)p(w)}$$

where w_l is the labeled word, $\mathcal{T}_{20}^i$ is the set of top 20 words in the i-th topic. The probabilities $p(w_l, w), p(w), p(w_l)$ are computed over a larger data set of English Wikipedia articles to reduce noise [Newman et al. 2011]. For whitening algorithm we choose $\alpha_0 = .01$. Note that other supervised topic modeling algorithms e.g. supervised LDA by Mcauliffe and Blei (2008), labeled LDA by Ramage et al. (2009) require a much stronger notion of side-information than just labeled words, hence we could not compare with them.



(a) (b)

Figure 7: Figure comparing the performance of Whitening, NMF [Arora et al. 2013], and semi-supervised NMF (SS-NMF) algorithms on NY Times and Yelp data sets. (a) Topics estimated by Whitening algorithm have the best PMI score in 40 out of 62 labeled words for NY Times data set, and 35 out of 54 labeled words in Yelp data set. (b) Whitening shows more than 2X speedup over competing algorithm in both data sets.

In Figure 7 (a) we plot the percentage of labeled words for which each algorithm has the best PMI score. Observe that for most labeled words (40 out of 62 labeled words for NY Times data set, and 35 out of 54 labeled words in Yelp data set) the Whitening algorithm estimates topic with better PMI score over NMF and SS-NMF algorithms. The Whitening algorithm is also more than twice as fast as NMF and SS-NMF[2] as shown in Figure 7 (b).

2. For large corpus the NMF algorithm runs much faster than Gibbs sampling and variational inference based algorithms [Arora et al. 2013].

A complete list of topics and PMI scores returned by the algorithms for every labeled word is presented in Tables 2, 3 of Appendix B. Notice that the Whitening algorithm often estimates more coherent topics which are more relevant to the given labeled word than topics produced by the NMF/SS-NMF algorithm. For example in NY Times data set with the labeled word *student* the Whitening algorithm returns top five words in the topic as *student, school, teacher, percent, program*; however those returned by NMF algorithm are *test, school, student, ignore, export*; and those by SS-NMF algorithm are *student, university, shooting, shot, rampage.*

Parallel image segmentation: One method to perform image segmentation is to use GMM clustering. In this experiment we demonstrate how GMM search algorithm can be used to parallelize image segmentation in vision applications. For this we consider the BSDS500 data set introduced in Arbelaez et al. (2011) and choose a subset of 70 images having less than 4 segments in the ground truth. Note that this data set has up to six ground truth segmentation by human users for each image. We randomly choose one pixel from each segment in ground truth as side-information v. We compare our Whitening algorithm with the seeded k-means clustering [Basu et al. 2002] where the centers are initialized by these side-information pixels (we refer to this as s-Kmeans). The Whitening algorithm uses one pixel from the i-th cluster to compute μ_i, in parallel for every i, and then it assigns each pixel to its closest μ_i. The segmentation quality is compared using normalized mutual information (NMI) metric [Manning et al. 2008]. To avoid local minimum in s-Kmeans we consider the maximum NMI over 5 initializations of side-information for each ground truth, and then we compute average NMI over all ground truths for an image.

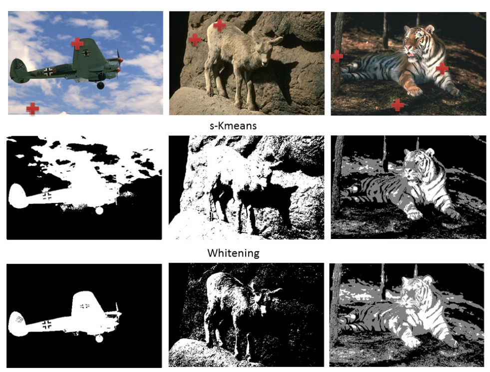


Figure 8: Figure comparing the performance of image segmentation by Whitening (row 3) and s-Kmeans (row 2) algorithms, with images selected from the BSDS500 data set. The side-information pixels are shown in red plus in the original image (row 1). In the segmented images (rows $2, 3$) the segments are shown in different shades. Observe that the Whitening algorithm often isolates the foreground segment better than s-Kmeans.

We summarize our result in Table 1. Observe that the Whitening algorithm has a slightly better NMI performance over s-Kmeans in the BSDS test data set and similar performance

Data set	N	N_W	N_K	T_W (s)	T_K (s)	$\overline{NMI}_W$	$\overline{NMI}_K$
BSDS test	30	17	13	6.7	81.5	0.17	0.13
BSDS train	25	12	13	8.2	89.8	0.15	0.15
BSDS val	15	8	7	10.6	117.2	0.11	0.09

Table 1: Table comparing the performance of Whitening and s-Kmeans algorithm on BSDS data set. N is the total number of images, N_W is the number of images where segmentation produced by Whitening has a better NMI than s-Kmeans, and N_K is the number of images where segmentation of s-Kmeans has a better NMI. T_W is the median runtime of Whitening algorithm and T_K is the median runtime of s-Kmeans. $\overline{NMI}_W$ and $\overline{NMI}_K$ are the median NMI scores for the Whitening and s-Kmeans algorithms respectively. Whitening runs much faster than s-Kmeans.

in BSDS train and BSDS val data sets. However the Whitening algorithm runs an order of magnitude faster than s-Kmeans.

5. Conclusion and Discussion

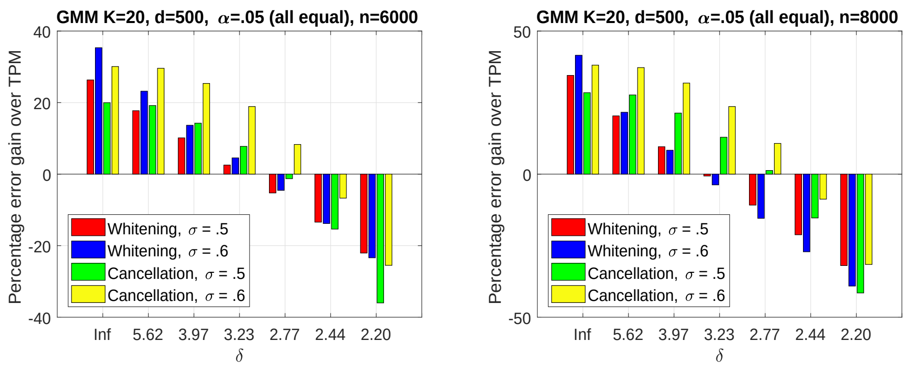
In this paper we developed a new, simple and flexible framework for incorporating side information into mixture model learning. The underlying motivation was to provide a principled way to take into account extra input (e.g. generated by human data analysts etc.). Even for cases where this input is very limited compared to the size/dimensionality of the data, we show meaningful statistical and computational performance improvement over baseline unsupervised and semi-supervised methods. More generally, developing methods which work with very limited human input is a promising research endeavor, in our opinion.

Acknowledgments

We would like to acknowledge support from NSF grants CNS-1320175, 0954059, ARO grants W911NF-15-1-0227, W911NF-14-1-0387, W911NF-16-1-0377, and the US DoT supported D- STOP Tier 1 University Transportation Center. The authors also acknowledge the Texas Advanced Computing Center [TACC 2018] at The University of Texas at Austin for providing HPC resources that have contributed to the research results reported within this paper.

Appendix A. More Experiments for Gaussian Mixture Models

In Figure 9 we show the sensitivity of the Whitening and Cancellation algorithms in GMM with $k = 20, d = 500$, all equal probability components, and two different values of σ and n. Observe that the percentage error gain of the algorithms decreases with decreasing values of $\delta = \min_{i \neq 1} \frac{\langle \mu_1, v \rangle}{\langle \mu_i, v \rangle}$, as we would expect, and it eventually becomes negative when the performance become worse than TPM algorithm. Also here the Cancellation algorithm shows lesser sensitivity, hence better performance compared to the Whitening algorithm.



(a) (b)

Figure 9: Sensitivity plots showing how the percentage relative error gain of the Whitening and Cancellation algorithms over the TPM algorithm decrease with decreasing values of the parameter $\delta = \min_{i \neq 1} \frac{\langle \mu_1, v \rangle}{\langle \mu_i, v \rangle}$, in GMM with $k = 20, d = 500$, all equal probability components, for different values of variance $\sigma \in \{.5, .6\}$, and two different sample complexities (a) $n = 6000$ (b) $n = 8000$.

Appendix B. Complete Results on New York Times and Yelp Data Set

In this section we provide more detailed result of our experiments on NY Times and Yelp data sets. In Tables 2, 3 we show for every labeled word, the top five words in the topics computed by Whtening, NMF, and SS-NMF algorithms along with their corresponding PMI scores.

Table 2: Results of topic search by Whitening and NMF algorithms on NYtimes data set of $300,000$ news articles using $K = 100$ topics and 62 labeled words.

NY Times data set							
Label word	Algo	topword-1	topword-2	topword-3	topword-4	topword-5	PMI
passenger	Whitening	flight	security	passenger	airport	hour	0.1424
	NMF	security	government	official	percent	bill	0.0499
	SSNMF	passenger	plane	flight	fire	crash	0.1711
coach	Whitening	coach	season	job	team	head	0.2637

Label word	Algo	topword-1	topword-2	topword-3	topword-4	topword-5	PMI
	NMF	team	coach	season	player	jet	0.1740
	SSNMF	coach	arrived	assistant	defenseman	ended	0.1756
art	Whitening	information	question	today	eastern	daily	0.0255
	NMF	art	show	dessert	book	home	0.0769
	SSNMF	art	artist	show	painting	museum	0.1250
campaign	Whitening	campaign	al gore	money	political	republican	0.1530
	NMF	al gore	campaign	george bush	president	bush	0.1608
	SSNMF	nra	florida	article	senator	presidential	0.0926
energy	Whitening	corp	meeting	list	dividend	partial	0.0815
	NMF	corp	meeting	list	group	dividend	0.0570
	SSNMF	partial	energy	dividend	meeting	corp	0.0254
tax	Whitening	tax	cut	taxes	percent	income	0.2126
	NMF	graf	president	bush	mail	information	0.0722
	SSNMF	tax	income	cut	taxes	site	0.2279
chef	Whitening	cup	minutes	food	article	add	0.0227
	NMF	buy	panelist	flavor	thought	product	0.0130
	SSNMF	tobacco	chef	restaurant	pastry	article	0.1495
oil	Whitening	oil	cup	minutes	prices	companies	0.1460
	NMF	oil	million	prices	percent	market	0.0928
	SSNMF	oil	company	listing	largest	brazil	0.0902
court	Whitening	court	case	law	decision	lawyer	0.2288
	NMF	official	court	case	attack	government	0.1285
	SSNMF	chicago	court	decision	ruling	justices	0.1834
election	Whitening	election	ballot	vote	voter	florida	0.2132
	NMF	election	ballot	al gore	bush	vote	0.2155
	SSNMF	gained	election	article	presidential	independence	0.1702
lawyer	Whitening	case	court	lawyer	death	trial	0.1830
	NMF	official	court	case	attack	government	0.1017
	SSNMF	lawyer	rat	legal	client	jokes	0.1314
anthrax	Whitening	mail	official	anthrax	attack	worker	0.0600
	NMF	anthrax	official	mail	worker	letter	0.0156
	SSNMF	anthrax	poverty	cb	show	return	-0.0776
golf	Whitening	tiger wood	shot	round	player	tour	0.1288
	NMF	tiger wood	shot	round	player	play	0.1356
	SSNMF	misstated	master	tee	hit	golf	0.1356
bacteria	Whitening	mail	anthrax	official	test	found	-0.0763
	NMF	anthrax	official	mail	worker	letter	-0.1097
	SSNMF	mas	bacteria	con	una	anos	-0.2420
film	Whitening	film	movie	director	character	actor	0.1906
	NMF	article	misstated	new york	company	million	0.0288
	SSNMF	kiss	film	actress	article	role	0.1295
tourist	Whitening	million	www	percent	building	night	0.0481
	NMF	team	tour	lance armstrong	won	race	-0.0405
	SSNMF	tourist	million	visitor	official	campaign	0.0995
horse	Whitening	race	won	win	run	track	0.1129
	NMF	race	won	horse	win	kentucky derby	0.1338
	SSNMF	horse	truck	road	official	killed	0.0433
republican	Whitening	campaign	george bush	bush	election	republican	0.2449
	NMF	al gore	campaign	george bush	president	bush	0.1868
	SSNMF	republican	democrat	democratic	house	parties	0.1053
computer	Whitening	computer	system	microsoft	program	software	0.1904

Label word	Algo	topword-1	topword-2	topword-3	topword-4	topword-5	PMI
	NMF	company	computer	microsoft	system	companies	0.1533
	SSNMF	computer	chip	mail	program	buy	0.1903
palestinian	Whitening	palestinian	israel	israeli	yasser arafat	peace	0.2189
	NMF	palestinian	israel	official	israeli	yasser arafat	0.1950
	SSNMF	palestinian	reformer	reform	authority	arab	0.1519
movie	Whitening	film	movie	director	character	actor	0.1492
	NMF	film	show	actor	movie	thought	0.0901
	SSNMF	red sox	movie	interview	seattle	host	0.0388
tennis	Whitening	player	play	won	game	women	0.1054
	NMF	game	play	player	point	andre agassi	0.1187
	SSNMF	motif	tennis	season	pros	image	0.1480
fight	Whitening	won	night	fight	win	sport	0.0566
	NMF	fight	mike tyson	lennox lewis	million	round	0.1181
	SSNMF	fight	pound	fighter	beat	boxing	0.1254
music	Whitening	music	song	record	album	band	0.2298
	NMF	music	company	million	companies	napster	0.0812
	SSNMF	music	mp3	customer	digital	online	0.0150
tablespoon	Whitening	cup	minutes	add	oil	tablespoon	0.0608
	NMF	cup	minutes	add	tablespoon	water	0.0431
	SSNMF	coffee	bean	tablespoon	cup	ground	-0.0765
nuclear	Whitening	bush	US	official	system	administration	0.1223
	NMF	official	bush	government	US	nuclear	0.1356
	SSNMF	ibm	nuclear	computer	research	fastest	-0.0253
racing	Whitening	race	car	driver	team	season	0.1443
	NMF	car	race	driver	team	season	0.1319
	SSNMF	sport	file	los angeles	racing	notebook	-0.0640
war	Whitening	military	taliban	war	afghanistan	us	0.0916
	NMF	taliban	official	afghanistan	government	us	0.0796
	SSNMF	russian	war	chechnya	army	veteran	0.1296
quarterback	Whitening	yard	season	game	play	team	0.2389
	NMF	game	team	play	yard	season	0.1773
	SSNMF	effort	quarterback	ucla	heroic	alabama	0.1472
stock	Whitening	stock	market	percent	company	fund	0.1585
	NMF	percent	stock	market	company	companies	0.1338
	SSNMF	stock	market	price	shares	investment	0.0507
ball	Whitening	game	run	yard	play	hit	0.1782
	NMF	run	game	inning	hit	season	0.1361
	SSNMF	ball	hit	run	inning	home	0.1708
patient	Whitening	patient	doctor	care	health	drug	0.2532
	NMF	official	virus	percent	new york	found	0.1003
	SSNMF	patient	study	doctor	article	brain	0.1334
champion	Whitening	won	win	round	shot	tiger wood	0.1029
	NMF	fight	mike tyson	lennox lewis	million	round	0.0955
	SSNMF	olympic	champion	final	meet	medalist	0.1177
business	Whitening	business	company	question	information	companies	0.0887
	NMF	information	eastern	commentary	daily	business	0.0311
	SSNMF	publication	business	send	released	businesses	0.0996
government	Whitening	government	official	country	federal	political	0.1524
	NMF	graf	president	bush	mail	information	0.0767
	SSNMF	program	government	computer	local	newspaper	0.0784
season	Whitening	season	team	game	games	play	0.1799
	NMF	team	game	season	play	games	0.1406

Label word	Algo	topword-1	topword-2	topword-3	topword-4	topword-5	PMI
	SSNMF	season	cotton	fact	simple	variety	0.0626
prison	Whitening	death	case	lawyer	court	trial	0.1333
	NMF	advise	spot	earlier	held	today	-0.0340
	SSNMF	prison	inmates	security	population	bed	0.1472
internet	Whitening	file	spot	internet	read	output	0.0359
	NMF	file	spot	new york	sport	los angeles	0.0228
	SSNMF	wonderful	mail	al gore	george bush	message	0.0766
rain	Whitening	air	part	high	wind	rain	0.1963
	NMF	air	wind	shower	rain	storm	0.1939
	SSNMF	chicago sun times	nominated	rain	east	thought	0.0179
game	Whitening	game	team	play	games	season	0.2000
	NMF	team	game	season	play	games	0.1722
	SSNMF	covering	game	tonight	coverage	celebration	0.0531
voter	Whitening	election	ballot	vote	percent	voter	0.2068
	NMF	election	ballot	al gore	bush	vote	0.1870
	SSNMF	voter	poll	percent	primary	election	0.2067
baseball	Whitening	player	team	season	game	sport	0.1691
	NMF	team	chicago white sox	mariner	season	player	0.1803
	SSNMF	velocity	baseball	air	shot	test	0.0629
student	Whitening	student	school	teacher	percent	program	0.2077
	NMF	test	school	student	ignore	export	0.0729
	SSNMF	student	university	shooting	shot	rampage	0.1396
president	Whitening	president	vice	white house	george bush	executive	0.2116
	NMF	graf	president	bush	mail	information	0.0758
	SSNMF	hedge	president	television	broadway	produced	0.0226
afghan	Whitening	taliban	afghanistan	military	us	war	0.1684
	NMF	taliban	official	afghanistan	government	us	0.1413
	SSNMF	afghan	afghanistan	blanket	friend	country	0.0577
medal	Whitening	team	games	won	women	american	0.1822
	NMF	team	tour	lance armstrong	won	race	0.0348
	SSNMF	endit	medal	honor	winner	newspaper	0.0786
teacher	Whitening	school	student	teacher	high	program	0.1566
	NMF	test	school	student	ignore	export	0.0388
	SSNMF	teacher	program	pay	school	teaching	0.1499
television	Whitening	show	home	network	television	night	0.1721
	NMF	los angeles daily new	spot	newspaper	new york	show	0.1456
	SSNMF	clinton	home	television	survived	tonight	-0.0090
democratic	Whitening	al gore	campaign	election	political	republican	0.1837
	NMF	al gore	campaign	george bush	president	bush	0.1677
	SSNMF	environmental	democratic	national committee	nominee	fund	0.0813
onion	Whitening	cup	minutes	add	oil	tablespoon	0.1039
	NMF	cup	minutes	add	tablespoon	water	0.1072
	SSNMF	flavor	panelist	ounces	buy	onion	0.1188
campus	Whitening	student	school	college	teacher	program	0.1314
	NMF	game	season	team	play	coach	-0.0595
	SSNMF	campus	operation	aol	building	center	0.0645
car	Whitening	car	driver	race	racing	seat	0.2047
	NMF	car	race	driver	team	season	0.1222

Label word	Algo	topword-1	topword-2	topword-3	topword-4	topword-5	PMI
	SSNMF	car	team	race	driver	winston cup	0.1516
industry	Whitening	companies	percent	company	business	industry	0.1430
	NMF	music	company	million	companies	napster	0.0821
	SSNMF	xxx	show	trade	software	entertainment	0.1161
planet	Whitening	film	today	system	movie	team	-0.0054
	NMF	wire	inadvertently	kill	mandatory	today	-0.0750
	SSNMF	captor	planet	film	kill	astronomer	0.0949
credit	Whitening	bill	money	member	system	number	0.1257
	NMF	bill	tax	bush	member	percent	0.0287
	SSNMF	donation	card	credit	account	voted	0.1382
race	Whitening	race	car	driver	won	win	0.1917
	NMF	car	race	driver	team	season	0.1814
	SSNMF	amazing	race	show	tonight	sit	0.0502
wine	Whitening	cup	minutes	food	add	oil	0.0499
	NMF	wine	wines	percent	company	million	0.0748
	SSNMF	wine	wines	bottle	bottles	age	0.1082
prosecutor	Whitening	case	death	lawyer	court	trial	0.1952
	NMF	official	court	case	attack	government	0.1363
	SSNMF	prosecutor	lawyer	attorney	incorrectly	general	0.1406
team	Whitening	team	season	game	player	play	0.1654
	NMF	team	game	season	play	games	0.1558
	SSNMF	team	qualify	olympic	article	member	0.1530
economy	Whitening	percent	market	economy	stock	cut	0.1528
	NMF	percent	stock	market	company	companies	0.1048
	SSNMF	percent	economy	quarter	rate	recession	0.1452
wind	Whitening	air	high	part	wind	rain	0.1909
	NMF	air	wind	shower	rain	storm	0.1895
	SSNMF	wash	wind	school	winter	white	0.1902
software	Whitening	microsoft	computer	system	company	software	0.1981
	NMF	company	computer	microsoft	system	companies	0.1911
	SSNMF	xxx	software	industry	show	trade	0.1222

Table 3: Results of topic search by Whitening and NMF algorithms on Yelp data set of $335,022$ reviews of businesses using $K = 100$ topics and 54 labeled words.

Yelp data set							
Label word	Algo	topword-1	topword-2	topword-3	topword-4	topword-5	PMI
cheese	Whitening	cheese	pizza	time	sandwich	back	0.1842
	NMF	bagel	coffee	bagels	cheese	sandwich	0.1666
	SSNMF	bartender	cheese	tasty	made	server	0.0555
salon	Whitening	hair	salon	nails	nail	back	0.0678
	NMF	hair	absolute	cut	beautiful	salon	-0.0192
	SSNMF	salon	manicure	back	nail	clean	0.0375
mexican	Whitening	mexican	burrito	tacos	salsa	cheese	0.0506
	NMF	mexican	fresh	burrito	tacos	time	0.0389
	SSNMF	exit	mexican	bland	restaurants	world	-0.0720
chinese	Whitening	chicken	chinese	rice	hot	fast	0.0978
	NMF	chicken	chinese	fast	rice	time	0.0717
	SSNMF	chinese	area	type	lot	east	0.0455

Label word	Algo	topword-1	topword-2	topword-3	topword-4	topword-5	PMI
tea	Whitening	coffee	find	things	tea	starbucks	0.1079
	NMF	find	store	things	tea	oil	0.0470
	SSNMF	tea	coffee	starbucks	safeway	ice	0.1787
sushi	Whitening	sushi	roll	happy	rolls	fish	0.0330
	NMF	cooks	fun	hash	browns	reasonable	-0.0441
	SSNMF	2nd	sushi	time	location	amazing	-0.1112
nail	Whitening	nails	nail	pedicure	salon	time	0.1385
	NMF	nails	nail	pedicure	time	salon	0.1316
	SSNMF	nail	nails	grandma	cut	make	0.0658
wash	Whitening	car	wash	clean	time	job	0.0617
	NMF	car	wash	back	time	job	0.0583
	SSNMF	car	wash	feels	clean	time	0.0290
insurance	Whitening	years	business	office	recommend	family	0.0856
	NMF	office	work	walk	time	insurance	0.0189
	SSNMF	insurance	years	business	steve	saved	0.0459
cream	Whitening	ice	cream	chocolate	cold	wait	0.1739
	NMF	ice	cream	school	cone	kids	0.1111
	SSNMF	cream	ice	wait	stone	cold	0.1494
hair	Whitening	hair	beautiful	absolute	years	salon	0.0749
	NMF	hair	absolute	cut	beautiful	salon	0.0507
	SSNMF	beautiful	hair	years	cut	time	0.0532
yoga	Whitening	classes	class	yoga	studio	gym	0.0928
	NMF	yoga	classes	class	studio	time	0.0816
	SSNMF	yoga	practice	dave	feel	amazing	0.0391
tire	Whitening	tire	tires	oil	car	discount	0.0739
	NMF	tire	car	tires	back	time	0.0634
	SSNMF	tire	tires	car	discount	time	0.0274
vietnamese	Whitening	time	chicken	thai	rice	chinese	-0.0442
	NMF	pho	chicken	rice	sauce	back	0.0825
	SSNMF	vietnamese	cake	chinese	back	fresh	-0.0105
donuts	Whitening	donuts	fresh	coffee	donut	chocolate	-0.0349
	NMF	donuts	coffee	donut	store	location	-0.0040
	SSNMF	donuts	donut	chocolate	time	selection	-0.1298
crust	Whitening	pizza	crust	wings	sauce	cheese	0.0068
	NMF	pizza	crust	wings	time	cheese	-0.0503
	SSNMF	min	pizza	crust	hut	pretty	-0.1131
ice	Whitening	ice	cream	cold	chocolate	flavors	0.1234
	NMF	ice	cream	school	cone	kids	0.0718
	SSNMF	ice	cream	wait	stone	cold	0.1312
pharmacy	Whitening	store	location	big	feel	kids	0.0075
	NMF	store	time	location	pharmacy	helpful	0.0049
	SSNMF	pharmacy	customer	clean	safeway	rude	-0.0127
beer	Whitening	bar	time	beer	wings	drinks	0.0900
	NMF	pizza	brick	pretty	bar	box	-0.0190
	SSNMF	beers	beer	operated	hand	locally	0.0817
bike	Whitening	bike	shop	guys	tires	back	0.0053
	NMF	bike	shop	back	bikes	time	0.0525
	SSNMF	bike	time	gun	pretty	store	-0.0293
yogurt	Whitening	yogurt	flavors	toppings	frozen	chocolate	0.0659
	NMF	yogurt	flavors	toppings	frozen	chocolate	0.0420
	SSNMF	yogurt	flavors	back	ice	shop	-0.1370
korean	Whitening	sushi	chinese	time	fresh	rice	-0.0311
	NMF	magazine	market	farmer	farmers	boston	-0.0702

Label word	Algo	topword-1	topword-2	topword-3	topword-4	topword-5	PMI
	SSNMF	korean	chicken	pretty	fried	spicy	0.0376
	Whitening	pizza	crust	wings	time	cheese	0.1491
pizza	NMF	pizza	brick	pretty	bar	box	0.0582
	SSNMF	pizza	ride	brick	long	red	0.0518
	Whitening	coffee	starbucks	donuts	tea	time	0.2728
coffee	NMF	coffee	busy	starbucks	ice	cream	0.2613
	SSNMF	coffee	starbucks	drinks	latte	work	0.0974
	Whitening	sandwich	subway	sandwiches	bread	time	0.1714
sandwich	NMF	sandwich	subway	fresh	bread	location	0.1311
	SSNMF	sandwich	sandwiches	ham	chips	limited	0.0083
	Whitening	time	thai	rice	sauce	back	-0.2046
pho	NMF	pho	chicken	rice	sauce	back	-0.1096
	SSNMF	pho	rice	beef	vietnamese	sauce	-0.0911
	Whitening	classes	class	work	gym	yoga	0.1518
gym	NMF	link	open	isn	working	fast	-0.0304
	SSNMF	gym	fitness	work	open	time	0.1117
	Whitening	dog	park	dogs	area	kids	0.1099
park	NMF	park	dog	time	area	trail	0.1023
	SSNMF	park	dog	dogs	lake	area	0.1303
	Whitening	coffee	starbucks	drink	time	make	-0.1617
latte	NMF	coffee	busy	starbucks	ice	cream	0.0802
	SSNMF	latte	location	work	drink	drinks	-0.0539
	Whitening	park	area	phoenix	time	lot	0.1356
trail	NMF	park	dog	time	area	trail	0.1049
	SSNMF	trail	parking	street	major	easy	0.0267
	Whitening	office	years	dentist	experience	work	0.0734
dentist	NMF	office	dentist	time	work	years	0.1169
	SSNMF	dentist	office	insurance	made	teeth	0.0766
	Whitening	starbucks	drink	coffee	drinks	times	-0.0972
starbucks	NMF	coffee	busy	starbucks	ice	cream	-0.0477
	SSNMF	starbucks	drink	argue	smile	times	-0.1099
	Whitening	taco	bell	tacos	fast	sauce	0.0994
taco	NMF	mexican	fresh	burrito	tacos	time	0.1875
	SSNMF	taco	bell	ghetto	pizza	location	-0.0042
	Whitening	mexican	burrito	tacos	salsa	fresh	0.0887
salsa	NMF	mexican	fresh	burrito	tacos	time	0.0267
	SSNMF	salsa	fresh	tacos	baja	fish	-0.0697
	Whitening	thai	rice	chinese	hot	chicken	0.0691
thai	NMF	thai	chicken	rice	back	sauce	0.1164
	SSNMF	thai	pad	tea	dish	green	0.0275
	Whitening	yogurt	flavors	chocolate	cream	ice	0.1923
chocolate	NMF	gelato	flavors	chocolate	ice	cream	0.1641
	SSNMF	chocolate	caramel	factory	dark	covered	0.1943
	Whitening	bar	drinks	night	time	beer	0.0142
bar	NMF	pizza	brick	pretty	bar	box	-0.0143
	SSNMF	bar	bit	big	seating	beer	-0.0086
	Whitening	chicken	chinese	rice	thai	sauce	0.2423
noodle	NMF	pho	chicken	rice	sauce	back	0.2630
	SSNMF	chicken	noodle	rice	back	sauces	0.0910
	Whitening	burrito	mexican	stars	tacos	salsa	0.1320
burrito	NMF	mexican	fresh	burrito	tacos	time	0.0638
	SSNMF	stars	burrito	green	sauce	mexican	0.0467
	Whitening	salad	chicken	fresh	sandwich	bar	0.1780
salad							

Label word	Algo	topword-1	topword-2	topword-3	topword-4	topword-5	PMI
	NMF	pizza	brick	pretty	bar	box	-0.0220
	SSNMF	salad	bar	salads	soup	competitors	-0.0123
burger	Whitening	burger	fries	burgers	fast	time	0.1489
	NMF	link	open	isn	working	fast	0.0159
	SSNMF	stale	burger	meat	bite	king	0.0322
hike	Whitening	park	area	time	lot	back	0.0572
	NMF	park	dog	time	area	trail	0.0747
	SSNMF	hike	park	rock	mountain	water	0.1255
pedicure	Whitening	nails	nail	pedicure	job	salon	0.0189
	NMF	nails	nail	pedicure	time	salon	0.0158
	SSNMF	pedicure	job	nail	close	home	-0.0931
fries	Whitening	burger	fries	burgers	fast	cheese	-0.0413
	NMF	cut	wait	time	hair	manager	-0.2616
	SSNMF	fries	grease	dirty	dark	slow	-0.1629
dog	Whitening	dog	dogs	park	pet	hot	0.1501
	NMF	dog	tony	cut	dogs	style	0.0751
	SSNMF	dog	door	tie	made	serve	0.0080
panda	Whitening	chicken	fast	chinese	rice	time	-0.1488
	NMF	chicken	chinese	fast	rice	time	-0.1291
	SSNMF	panda	orange	rice	fried	bad	-0.1327
beans	Whitening	mexican	burrito	chicken	tacos	salsa	-0.0550
	NMF	mexican	fresh	burrito	tacos	time	-0.1419
	SSNMF	trouble	beans	rice	chicken	marinated	-0.1233
subway	Whitening	subway	sandwich	clean	fresh	location	-0.0074
	NMF	sandwich	subway	fresh	bread	location	-0.0445
	SSNMF	subway	location	clean	super	sandwich	-0.0524
car	Whitening	car	wash	back	time	work	0.1064
	NMF	car	wash	back	time	job	0.0874
	SSNMF	visited	car	back	job	weeks	0.0353
cake	Whitening	found	cake	chocolate	shop	yogurt	0.0754
	NMF	back	time	shop	cake	found	0.0099
	SSNMF	cake	wanted	wedding	flavor	perfect	0.0416
steak	Whitening	location	fast	makes	feel	quality	-0.0672
	NMF	prices	selection	quality	family	helpful	-0.1569
	SSNMF	difference	fast	steak	sandwiches	subs	-0.1672
curry	Whitening	thai	chicken	rice	chinese	hot	0.1482
	NMF	thai	chicken	rice	back	sauce	0.1903
	SSNMF	chicken	stew	brown	curry	rice	0.0047
massage	Whitening	massage	back	amazing	years	spa	0.1359
	NMF	massage	time	back	amazing	hour	-0.0035
	SSNMF	massage	arts	experience	amazing	hour	-0.0168
italian	Whitening	sandwich	pizza	time	back	bread	-0.0254
	NMF	gelato	flavors	chocolate	ice	cream	0.0241
	SSNMF	ice	italian	flavors	cream	chocolate	-0.0231

Appendix C. Computation of A, B for Different Models

This section outlines the construction of matrices A, B in various models via different moment computations. First we introduce some notations which we use in Appendices C, D, E, F, and G.

C.1 Notations

For a vector x, $\|x\|$ denotes its ℓ_2 norm. For a matrix X, $\|X\|$ represents the spectral norm of the matrix. We use the notation $\widehat{X}$ or $\widehat{\mathbb{E}}[X]$ to represent the sample estimate of a quantity X, unless mentioned otherwise. For a matrix M let $\sigma_k(M)$ denote the $k-$th largest singular value of M, and $\tilde{\sigma}_k(M)$ denote the $k-$th largest eigenvalue. n represents the number of samples used to obtain the sample estimates. Next, we introduce some basic tensor notations. Let $x, y, z \in \mathbb{R}^d$ be three d dimensional vectors. Then the order-3 tensor $T_3 = x \otimes y \otimes z$ is defined as $T_3(i,j,k) = x(i)y(j)z(k)$, for $i,j,k \in [d]$. Similarly the order-2 tensor $T_2 = x \otimes y$ is equivalent to the matrix outer product $T_2 = xy^T$. Finally let $v \in \mathbb{R}^d$ be another d dimensional vector, I be the d dimensional identity matrix. The tensor contraction $T_3(I,I,v)$ is equal to the order-2 tensor $T_3(I,I,v) = \langle z, v\rangle x \otimes y$, which is again equivalent to the matrix $T_3(I,I,v) = \langle z, v\rangle xy^T$. For order-2 tensors we will use the tensor and matrix notations interchangeably.

C.2 GMM Moments

In this section we prove how the required matrices A, B can be computed in the GMM model. We restate the following useful theorem from Hsu and Kakade (2013) which computes three tensor moments for the GMM model.

Theorem 7 (Hsu and Kakade (2013)) *Consider the GMM model with means $\{\mu_1, \dots, \mu_k\}$ and corresponding variances $\{\sigma_1^2, \dots, \sigma_k^2\}$, and α_i denote the proportion of the i-th component in the mixture. Let $\sigma^2 = \sum_{i=1}^k \alpha_i \sigma_i^2$ be the smallest eigenvalue of the covariance matrix $\mathbb{E}[(x - \mathbb{E}[x])(x - \mathbb{E}[x])^T]$ (note that since $\sum \alpha_i \mu_i \mu_i^T$ has rank k, this is the same as the $k+1$th-largest eigenvalue), and u be a unit norm eigenvector corresponding to the eigenvalue σ^2. Define*

$$\begin{aligned} \widetilde{m} &= \mathbb{E}[x(u^T(x - \mathbb{E}[x]))^2], \quad M_2 = \mathbb{E}[x \otimes x] - \sigma^2 I \\ M_3 &= \mathbb{E}[x \otimes x \otimes x] - \sum_{i=1}^d (\widetilde{m} \otimes e_i \otimes e_i + e_i \otimes \widetilde{m} \otimes e_i + e_i \otimes e_i \otimes \widetilde{m}) \end{aligned}$$

where $\{e_1, \dots, e_d\}$ form standard basis of $\mathbb{R}^d$. Then,

$$\widetilde{m} = \sum_{i=1}^k \alpha_i \sigma_i^2 \mu_i, \quad M_2 = \sum_{i=1}^k \alpha_i \mu_i \otimes \mu_i, \quad M_3 = \sum_{i=1}^k \alpha_i \mu_i \otimes \mu_i \otimes \mu_i.$$

Theorem 8 *In the GMM model define*

$$\begin{aligned} m &= \mathbb{E}[x], \quad A = \mathbb{E}[xx^T] - \sigma^2 I_d \\ B &= \mathbb{E}[\langle x, v\rangle xx^T] - \widetilde{m}v^T - v\widetilde{m}^T - \langle \widetilde{m}, v\rangle I_d \end{aligned}$$

Then, $m = \sum_i \alpha_i \mu_i$, $A = \sum_{i=1}^k \alpha_i \mu_i \mu_i^T$ and $B = \sum_{i=1}^k \alpha_i \langle \mu_i, v\rangle \mu_i \mu_i^T$

Proof The expression for m, A follows directly from Theorem 7 by noting that $A = M_2$ and $\mu_i \otimes \mu_i = \mu_i \mu_i^T$. To compute B consider the tensor contraction $M_3(I,I,v)$, M_3 as in

Theorem 7. Then,

$$\begin{aligned} M_3(I,I,v) &= \mathbb{E}[\langle x,v\rangle x\otimes x] - \sum_{i=1}^{d}(v(i)\widetilde{m}\otimes e_i + v(i)e_i\otimes\widetilde{m} + \langle\widetilde{m},v\rangle e_i\otimes e_i) \\ &= \mathbb{E}[\langle x,v\rangle xx^T] - \sum_{i=1}^{d}(v(i)\widetilde{m}e_i^T + v(i)e_i\widetilde{m}^T + \langle\widetilde{m},v\rangle e_ie_i^T) \\ &= \mathbb{E}[\langle x,v\rangle xx^T] - \widetilde{m}v^T - v\widetilde{m}^T - \langle\widetilde{m},v\rangle I_d = B \end{aligned}$$

Also from Theorem 7, $M_3(I,I,v) = \sum_{i=1}^k \alpha_i\langle\mu_i,v\rangle\mu_i\otimes\mu_i = \sum_{i=1}^k \alpha_i\langle\mu_i,v\rangle\mu_i\mu_i^T$. Therefore $B = \sum_{i=1}^k \alpha_i\langle\mu_i,v\rangle\mu_i\mu_i^T$. ■

C.3 LDA Moments

In this section we show the m, A, B computation corresponding to the LDA model. Again we restate the following theorem from Anandkumar et al. (2014) which computes the first three tensor moments for LDA distribution.

Theorem 9 (Anandkumar et al. (2014)) *In an LDA model with parameters $\bar{\alpha} = (\alpha_1,\ldots,\alpha_k)$, topic distributions $\mu_1,\ldots,\mu_k$. Let $\alpha_0 = \sum_{i=1}^k \alpha_i$. Define*

$$\begin{aligned} M_1 &= \mathbb{E}[x_1], \quad M_2 = \mathbb{E}[x_1\otimes x_2] - \frac{\alpha_0}{1+\alpha_0}M_1\otimes M_1 \\ M_3 &= \mathbb{E}[x_1\otimes x_2\otimes x_3] - \frac{\alpha_0}{\alpha_0+2}\left(\mathbb{E}[x_1\otimes x_2\otimes M_1] + \mathbb{E}[x_1\otimes M_1\otimes x_3] + \mathbb{E}[M_1\otimes x_2\otimes x_3]\right) \\ &\quad + \frac{2\alpha_0^2}{(\alpha_0+1)(\alpha_0+2)}M_1\otimes M_1\otimes M_1 \end{aligned}$$

Then,

$$\begin{aligned} M_1 &= \sum_{i=1}^{k}\frac{\alpha_i}{\alpha_0}\mu_i, \quad M_2 = \sum_{i=1}^{k}\frac{\alpha_i}{\alpha_0(\alpha_0+1)}\mu_i\otimes\mu_i \\ M_3 &= \sum_{i=1}^{k}\frac{2\alpha_i}{\alpha_0(\alpha_0+1)(\alpha_0+2)}\mu_i\otimes\mu_i\otimes\mu_i \end{aligned}$$

Theorem 10 *For an LDA model for any $v\in\mathbb{R}^d$ suppose m, A, B be defined as*

$$\begin{aligned} m &= \alpha_0\mathbb{E}[x_1] \\ A &= \alpha_0(\alpha_0+1)\mathbb{E}[x_1x_2^T] - mm^T \\ B &= \frac{\alpha_0(\alpha_0+1)(\alpha_0+2)}{2}\mathbb{E}[\langle x_3,v\rangle x_1x_2^T] - \frac{\alpha_0(\alpha_0+1)}{2}\big(\langle m,v\rangle\mathbb{E}[x_1x_2^T] + \mathbb{E}[\langle x_3,v\rangle x_1m^T] \\ &\quad + \mathbb{E}[\langle x_3,v\rangle mx_2^T]\big) + \langle m,v\rangle mm^T. \end{aligned}$$

Then we can express m, A, B as follows.

$$m = \sum_{i=1}^{k} \alpha_i \mu_i, \quad A = \sum_{i=1}^{k} \alpha_i \mu_i \mu_i^T, \quad B = \sum_{i=1}^{k} \alpha_i \langle \mu_i, v \rangle \mu_i \mu_i^T$$

Proof The expressions for m and A follows easily from Theorem 9 since $m = \alpha_0 M_1$ and $A = \alpha_0(\alpha_0+1)M_2$. To show the expression for B consider the tensor contraction $M_3(I, I, v)$, M_3 defined as in Theorem 9. Then we have

$$\begin{aligned} M_3(I, I, v) &= \mathbb{E}[\langle x_3, v \rangle x_1 \otimes x_2] - \frac{\alpha_0}{\alpha_0 + 2} (\mathbb{E}[\langle M_1, v \rangle x_1 \otimes x_2] + \mathbb{E}[\langle x_3, v \rangle x_1 \otimes M_1] \\ &\quad + \mathbb{E}[\langle x_3, v \rangle M_1 \otimes x_2 \otimes x_3]) + \frac{2\alpha_0^2}{(\alpha_0 + 1)(\alpha_0 + 2)} \langle M_1, v \rangle \otimes M_1 \otimes M_1 \\ &= \frac{2}{\alpha_0(\alpha_0 + 1)(\alpha_0 + 2)} B \end{aligned}$$

where we used $x_1 \otimes x_2$ is same as $x_1 x_2^T$ and so on. We also get from Theorem 9 $M_3(I, I, v) = \sum_{i=1}^{k} \frac{2\alpha_i}{\alpha_0(\alpha_0+1)(\alpha_0+2)} \langle \mu_i, v \rangle \mu_i \otimes \mu_i$. Therefore we have

$$B = \frac{\alpha_0(\alpha_0 + 1)(\alpha_0 + 2)}{2} M_3(I, I, v) = \sum_{i=1}^{k} \alpha_i \langle \mu_i, v \rangle \mu_i \mu_i^T.$$

■

C.4 Mixed Regression Moments

Recall in mixed regression we have $y = \langle x, \mu_i \rangle + \xi$ where $x \sim \mathcal{N}(0, I)$ and $\xi \sim \mathcal{N}(0, \sigma^2)$. In the following Lemmas we compute the various moments $M_{1,1}, M_{2,2}, M_{3,1}, M_{3,3}$ and show how they are used to compute m, A, B.

Lemma 11 *In mixed linear regression define $M_{1,1} = \mathbb{E}[yx]$, $M_{2,2} = \mathbb{E}[y^2 x x^T]$, $M_{3,1} = \mathbb{E}[y^3 x]$ and $M_{3,3} = \mathbb{E}[y^3 \langle x, v \rangle x x^T]$. Then,*

$$\begin{aligned} M_{1,1} &= \sum_{i=1}^{k} \alpha_i \mu_i \\ M_{2,2} &= 2 \sum_{i=1}^{k} \alpha_i \mu_i \mu_i^T + (\sigma^2 + \sum_{i=1}^{k} \alpha_i \|\mu_i\|^2) I \\ M_{3,1} &= 3 \sum_{i=1}^{k} \alpha_i (\sigma^2 + \|\mu_i\|^2) \mu_i \\ M_{3,3} &= 6 \sum_{i=1}^{k} \alpha_i \langle \mu_i, v \rangle \mu_i \mu_i^T + \left(M_{3,1} v^T + v M_{3,1}^T + \langle M_{3,1}, v \rangle I \right) \end{aligned}$$

Proof

We compute the moments as shown below.

$$M_{1,1} = \mathbb{E}[yx] = \sum_{i=1}^{k} \alpha_i \mathbb{E}[x^T \mu_i x + \xi x] = \sum_{i=1}^{k} \alpha_i \mu_i$$

$$\begin{aligned} M_{2,2} &= \mathbb{E}[y^2 xx^T] = \sum_{i=1}^{k} \alpha_i \mathbb{E}[\langle \mu_i, x\rangle^2 xx^T] + \mathbb{E}[\xi^2]\mathbb{E}[xx^T] \\ &= \sum_{i=1}^{k} \alpha_i \mathbb{E}[\langle \mu_i, x\rangle^2 xx^T] + \sigma^2 I \\ &= \sum_{i=1}^{k} \alpha_i (2\mu_i\mu_i^T + \|\mu_i\|^2 I) + \sigma^2 I \\ &= 2\sum_{i=1}^{k} \alpha_i \mu_i \mu_i^T + \sum_{i=1}^{k} \alpha_i (\sigma^2 + \|\mu_i\|^2) I \end{aligned}$$

Using the fact that all odd moments of normal random variable are zero.

$$\begin{aligned} M_{3,1} &= \mathbb{E}[y^3 x] = \sum_{i=1}^{k} \alpha_i \mathbb{E}[(\langle x, \mu_i\rangle + \xi)^3 x] \\ &= \sum_{i=1}^{k} \alpha_i \mathbb{E}[\langle x, \mu_i\rangle^3 x] + 3\sum_{i=1}^{k} \alpha_i \mathbb{E}[\xi^2]\mathbb{E}[\langle x, \mu_i\rangle x] \\ &= 3\sum_{i=1}^{k} \alpha_i \|\mu_i\|^2 \mu_i + 3\sum_{i=1}^{k} \alpha_i \sigma^2 \mu_i = 3\sum_{i=1}^{k} \alpha_i (\sigma^2 + \|\mu_i\|^2)\mu_i \end{aligned}$$

We use the fact that for even p the moment $\mathbb{E}[z^p] = (p-1)!!$ for a standard normal random variable z and !! denote the double factorial. Next we compute $M_{3,3}$.

$$\begin{aligned} M_{3,3} &= \mathbb{E}[y^3 \langle x, v\rangle xx^T] = \sum_{i=1}^{k} \alpha_i \mathbb{E}[(\langle x, \mu_i\rangle + \xi)^3 \langle x, v\rangle xx^T] \\ &= \sum_{i=1}^{k} \alpha_i \mathbb{E}[\langle x, \mu_i\rangle^3 \langle x, v\rangle xx^T] + 3\sum_{i=1}^{k} \alpha_i \mathbb{E}[\xi^2]\mathbb{E}[\langle x, v\rangle \langle x, \mu_i\rangle xx^T] \\ &= \sum_{i=1}^{k} \alpha_i \mathbb{E}[\langle x, \mu_i\rangle^3 \langle x, v\rangle xx^T] + 3\sigma^2 \sum_{i=1}^{k} \alpha_i \mathbb{E}[\langle x, v\rangle \langle x, \mu_i\rangle xx^T] \end{aligned} \tag{5}$$

Now we compute these individual moments.

$$\mathbb{E}[\langle x, v\rangle \langle x, \mu_i\rangle xx^T] = \mu_i^T v + v\mu_i^T + \langle \mu_i, v\rangle I$$

Using the fact that any odd combination of the variables in x will be zero in expectation. Also,

$$\mathbb{E}[\langle x, \mu_i\rangle^3 \langle x, v\rangle x x^T] = 6\langle v, \mu_i\rangle \mu_i \mu_i^T + 3\|\mu_i\|^2[\mu_i^T v + v\mu_i^T + \langle \mu_i, v\rangle I]$$

Again by using the moments of standard normal variable. This can be verified by considering the (a, b)-th entry of the matrix on the right as a polynomial in $\mu_i(l)$, the l-th component of μ_i, and matching the corresponding coefficients from both sides of the equation.

Combining with equation (5) we get,

$$\begin{aligned}
M_{3,3} &= \sum_{i=1}^k \alpha_i \left[6\langle v, \mu_i\rangle \mu_i \mu_i^T + 3\|\mu_i\|^2(\mu_i^T v + v\mu_i^T + \langle \mu_i, v\rangle I)\right] \\
&\quad +3\sigma^2 \sum_{i=1}^k \alpha_i[\mu_i^T v + v\mu_i^T + \langle \mu_i, v\rangle I] \\
&= 6\sum_{i=1}^k \alpha_i \langle v, \mu_i\rangle \mu_i \mu_i^T + 3\sum_{i=1}^k \alpha_i(\sigma^2 + \|\mu_i\|^2)[\mu_i^T v + v\mu_i^T + \langle \mu_i, v\rangle I] \\
&= 6\sum_{i=1}^k \alpha_i \langle v, \mu_i\rangle \mu_i \mu_i^T + \left(M_{3,1} v^T + v M_{3,1}^T + \langle M_{3,1}, v\rangle I\right)
\end{aligned}$$

■

Theorem 12 *Let m, A, B be defined as*

$$\begin{aligned}
m &= M_{1,1}, \quad A = \frac{1}{2}(M_{2,2} - \tau^2 I), \\
B &= \frac{1}{6}(M_{3,3} - (M_{3,1} v^T + v M_{3,1}^T + \langle M_{3,1}, v\rangle I))
\end{aligned}$$

where τ^2 is the smallest singular value of $M_{2,2}$. Then,

$$m = \sum_{i=1}^k \alpha_i \mu_i, \quad A = \sum_{i=1}^k \alpha_i \mu_i \mu_i^T, \quad B = \sum_{i=1}^k \alpha_i \langle \mu_i, v\rangle \mu_i \mu_i^T$$

Proof The proof follows directly from Lemma 11. Note that since μ_i-s are linearly independent the smallest singular vector τ^2 of $M_{2,2}$ is equal to $\sum_{i=1}^k \alpha_i(\sigma^2 + \|\mu_i\|^2)$. Then $A = \frac{1}{2}\left(M_{2,2} - \tau^2 I\right) = \sum_{i=1}^k \alpha_i \mu_i \mu_i^T$. Similarly the expression for B holds. ■

C.5 Subspace Clustering Moments

In this section we derive the necessary moments required for subspace clustering. Recall that in the subspace clustering model we have k dimension—m subspaces $U_1, \ldots, U_k \in \mathbb{R}^{d\times m}$ (matrices $U_1, \ldots, U_k$ have orthonormal columns). The data is generated as follows. We sample $y \sim \mathcal{N}(0, I_d)$ and set $x = U_i U_i^T y + \xi$, where $\xi \sim \mathcal{N}(0, \sigma^2 I_d)$ is additive noise.

Theorem 13 *Consider the subspace clustering model. Let M_2, A, B be defined as,*

$$\begin{aligned} M_2 &:= \mathbb{E}[xx^T], \quad A := M_2 - \sigma^2 I_d \\ B &:= \mathbb{E}[\langle x, v\rangle^2 xx^T] - \sigma^2(v^T A v) I_d - \sigma^2 \|v\|^2 A - \sigma^4(\|v\|^2 I_d + vv^T) - 2\sigma^2(Avv^T + vv^T A) \end{aligned}$$

where $\sigma^2 = \sigma_{mk+1}(M_2)$. Then,

$$\begin{aligned} A &= \sum_{i=1}^k \alpha_i U_i U_i^T \\ B &= \sum_{i=1}^k \alpha_i \|U_i^T v\|^2 U_i U_i^T + 2\sum_{i=1}^k \alpha_i U_i U_i^T vv^T U_i U_i^T \end{aligned}$$

Proof First we compute M_2.

$$M_2 = \mathbb{E}(xx^T) = \sum_{i=1}^k \alpha_i \mathbb{E}\left[U_i U_i^T yy^T U_i U_i^T\right] + \mathbb{E}[\xi\xi^T] = \sum_{i=1}^k \alpha_i U_i U_i^T + \sigma^2 I_d$$

Using $\mathbb{E}[yy^T] = I$ as $y \sim \mathcal{N}(0, I)$ and $U_i^T U_i = I$ since the columns are orthogonal. Since $\alpha_i > 0$, the $mk+1$-th singular value of M_2, $\sigma_{mk+1}(M_2) = \sigma^2$. Therefore it follows that,

$$A = M_2 - \sigma^2 I_d = \sum_{i=1}^k \alpha_i U_i U_i^T$$

Now we compute the moment $\mathbb{E}[\langle x, v\rangle^2 xx^T]$. Given a sample $x = U_i U_i^T y + \xi$ from the i-th subspace we have,

$$\begin{aligned} \langle x, v\rangle^2 &= v^T U_i U_i^T yy^T U_i U_i^T v + v^T \xi\xi^T v + 2v^T \xi v^T U_i U_i^T y \\ xx^T &= U_i U_i^T yy^T U_i U_i^T + U_i U_i^T y\xi^T + \xi y^T U_i U_i^T + \xi\xi^T \end{aligned}$$

Then we can write,

$$\begin{aligned} &\mathbb{E}[\langle x, v\rangle^2 xx^T] \\ =& \sum_{i=1}^k \alpha_i \big(\mathbb{E}[v^T U_i U_i^T yy^T U_i U_i^T v U_i U_i^T yy^T U_i U_i^T] + \mathbb{E}[v^T U_i U_i^T yy^T U_i U_i^T v]\mathbb{E}[\xi\xi^T] \\ &+ \mathbb{E}[v^T \xi\xi^T v]\mathbb{E}[U_i U_i^T yy^T U_i U_i^T] + \mathbb{E}[v^T \xi\xi^T v \xi\xi^T] + 2\mathbb{E}[(v^T \xi v^T U_i U_i^T y) U_i U_i^T y \xi^T] \\ &+ 2\mathbb{E}[(v^T \xi v^T U_i U_i^T y)\xi y^T U_i U_i^T]\big) \\ =& T_1 + T_2 + T_3 + T_4 + T_5 + T_6 \end{aligned} \tag{6}$$

where $T_1, \ldots, T_6$ are as follows. We define $v_i := U_i U_i^T v$, we use the Gaussian moment results $\mathbb{E}[\langle v, z\rangle z] = \sigma^2 v$, and $\mathbb{E}[\langle v, z\rangle^2 z z^T] = \sigma^4(\|v\|^2 I_d + vv^T)$ whenever $z \sim \mathcal{N}(0, \sigma^2 I_d)$.

$$
\begin{aligned}
T_1 &= \sum_{i=1}^k \alpha_i \mathbb{E}\left[v^T U_i U_i^T y y^T U_i U_i^T v U_i U_i^T y y^T U_i U_i^T\right] \\
&= \sum_{i=1}^k \alpha_i \mathbb{E}[\langle y, v_i\rangle^2 U_i U_i^T y y^T U_i U_i^T] = \sum_{i=1}^k \alpha_i U_i U_i^T \mathbb{E}[\langle y, v_i\rangle^2 y y^T] U_i U_i^T \\
&= \sum_{i=1}^k \alpha_i U_i U_i^T (\|v_i\|^2 I_d + 2 v_i v_i^T) U_i U_i^T \\
&= \sum_{i=1}^n \alpha_i \|v_i\|^2 U_i U_i^T + 2\sum_{i=1}^k \alpha_i U_i U_i^T v v^T U_i U_i^T \\
&= \sum_{i=1}^k \alpha_i \|U_i^T v\|^2 U_i U_i^T + 2\sum_{i=1}^k \alpha_i U_i U_i^T v v^T U_i U_i^T
\end{aligned}
$$

since $\|v_i\| = \|U_i U_i^T v\| = \|U_i^T v\|$.

$$
\begin{aligned}
T_2 &= \sum_{i=1}^k \alpha_i \mathbb{E}[v^T U_i U_i^T y y^T U_i U_i^T v]\mathbb{E}[\xi\xi^T] = \sum_{i=1}^k \alpha_i v^T U_i U_i^T v \times \sigma^2 I_d = \sigma^2 (v^T A v) I_d \\
T_3 &= \sum_{i=1}^k \alpha_i \mathbb{E}[v^T \xi\xi^T v]\mathbb{E}[U_i U_i^T y y^T U_i U_i^T] = \sigma^2 \|v\|^2 \sum_{i=1}^k \alpha_i U_i U_i^T = \sigma^2 \|v\|^2 A \\
T_4 &= \sum_{i=1}^k \alpha_i \mathbb{E}[v^T \xi\xi^T v \xi\xi^T] = \mathbb{E}[\langle v, \xi\rangle^2 \xi\xi^T] = \sigma^4(\|v\|^2 I_d + 2 v v^T) \\
T_5 &= \sum_{i=1}^k \alpha_i 2\mathbb{E}[(v^T \xi v^T U_i U_i^T y) U_i U_i^T y \xi^T] = 2\sum_{i=1}^k \alpha_i \mathbb{E}[(v^T U_i U_i^T y) U_i U_i^T y]\mathbb{E}[\langle v, \xi\rangle \xi^T] \\
&= 2\sum_{i=1}^k \alpha_i \mathbb{E}[(v^T U_i U_i^T y) U_i U_i^T y] \times \sigma^2 v^T = 2\sigma^2 \sum_{i=1}^k \alpha_i \mathbb{E}[(v^T U_i U_i^T y) U_i U_i^T y v^T] \\
&= 2\sigma^2 \sum_{i=1}^k \alpha_i \mathbb{E}[U_i U_i^T \langle v, y\rangle y v^T] = 2\sigma^2 \sum_{i=1}^k \alpha_i U_i U_i^T v v^T = 2\sigma^2 A v v^T \\
T_6 &= 2\sum_{i=1}^k \alpha_i \mathbb{E}[(v^T \xi v^T U_i U_i^T y)\xi y^T U_i U_i^T] = 2\sum_{i=1}^k \alpha_i \mathbb{E}[\langle v, \xi\rangle \xi]\mathbb{E}[\langle v_i, y\rangle y^T U_i U_i^T] \\
&= 2\sigma^2 \sum_{i=1}^k \alpha_i v v_i^T U_i U_i^T = \sigma^2 \sum_{i=1}^k \alpha_i v v^T U_i U_i^T = \sigma^2 v v^T \sum_{i=1}^k \alpha_i U_i U_i^T = 2\sigma^2 v v^T A
\end{aligned}
$$

Therefore,

$$\begin{aligned} B &= \mathbb{E}[\langle x, v\rangle^2 xx^T] - \sigma^2(v^T A v) I_d - \sigma^2 \|v\|^2 A - \sigma^4(\|v\|^2 I_d + vv^T) - 2\sigma^2(Avv^T + vv^T A) \\ &= \mathbb{E}[\langle x, v\rangle^2 xx^T] - T_2 - T_3 - T_4 - T_5 - T_6 = T_1 \\ &= \sum_{i=1}^k \alpha_i \|U_i^T v\|^2 U_i U_i^T + 2\sum_{i=1}^k \alpha_i U_i U_i^T vv^T U_i U_i^T \end{aligned}$$

■

Appendix D. Finite-sample Analysis of the Whitening Method

Suppose that

$$A = \sum_i \alpha_i \mu_i \mu_i^T$$

$$B = \sum_i \beta_i \mu_i \mu_i^T$$

$$\|A - \hat{A}\| \leq \epsilon$$

$$\|B - \hat{B}\| \leq \epsilon,$$

where σ_k is the kth singular value of A. Let V be the $n \times k$ matrix whose columns are the first k singular vectors of A, and let $\hat{V}$ be the same for $\hat{A}$. Let D be the diagonal matrix of singular values of A, and let $\hat{D}$ be the diagonal matrix of the first k singular values of $\hat{A}$. Then $A = VDV^T$ and $V^T V = \hat{V}^T \hat{V} = I_k$. This entire section is under the assumptions of Theorem 1; in particular, recall that $\epsilon \leq \sigma_k(A)/4$.

It will be technically convenient for us to assume that $\|B\| \leq \|A\| = \sigma_1(A)$. This assumption holds without loss of generality: if not, simply rescale the side information, setting $v^{\text{new}} = v\frac{\|A\|}{\|B\|}$. This has the effect of rescaling B, so that $\|B^{\text{new}}\| = \|A\|$; define also $\hat{B}^{\text{new}} = \hat{B}\frac{\|A\|}{\|B\|}$. Note that

$$\|B^{\text{new}} - \hat{B}^{\text{new}}\| = \|B - \hat{B}\| \frac{\|A\|}{\|B\|} \leq \epsilon$$

under the assumption $\|B - \hat{B}\| \leq \epsilon$. Now, the algorithm is homogeneous in $\hat{B}$: it will produce the same output given either $\hat{B}$ or $\hat{B}^{\text{new}}$; hence, it suffices to prove Theorem 1 with v, B, and $\hat{B}$ replaced by their new versions. Since the new versions satisfy $\|B^{\text{new}}\| \leq \|A\|$, we may assume this without loss of generality. From now on, we will drop the notation B^{new}, and we will simply prove Theorem 1 under the assumption $\|B\| \leq \|A\|$.

Our basic tool is Wedin's theorem:

Theorem 14 *For a matrix A, let $P^A_{\geq s}$ be the orthogonal projection onto the subspace spanned by singular vectors of A with singular value at least s. Let $P^A_{\leq s}$ be the orthogonal projection onto the subspace spanned by singular vectors with singular value at most s. Then for any matrices A and B, and for any $s < t$,*

$$\|P^A_{\leq s} P^B_{\geq t}\| \leq \frac{2\|A - B\|}{t - s}.$$

In applying Wedin's theorem, the following geometric lemma will be useful. In what follows, P_E denotes the orthogonal projection onto E.

Lemma 15 *Let E and F be subspaces of $\mathbb{R}^n$ with $\|P_{E^\perp} P_F\| \le \delta$. Then $\|P_F v\|^2 \le \|P_E v\|^2 + 3\delta\|v\|^2$ for every $v \in \mathbb{R}^n$.*

Lemma 16 *If $\epsilon < \sigma_k/4$ then for any $u \in \mathbb{R}^k$,*

$$\sqrt{1 - \frac{16\epsilon^2}{\sigma_k^2}}\|u\| \le \|\hat{V}^T V u\| \le \|u\|.$$

By a simple change of variables, if we define

$$O = D^{-1/2}\hat{V}^T V D^{1/2}$$

then O is also an almost-isometry: for every $u \in \mathbb{R}^k$,

$$\sqrt{1 - \frac{16\epsilon^2}{\sigma_k^2}}\|u\| \le \|Ou\| \le \|u\|. \tag{7}$$

Proof First, note that $\sigma_k(\hat{A}) \ge \sigma_k(A) - \|A - \hat{A}\| \ge \sigma_k - \epsilon$. If $\epsilon < \sigma_k/4$, we also have $\sigma_{k+1}(\hat{A}) \le \sigma_{k+1}(A) + \epsilon \le \sigma_k/4 < \sigma_k - \epsilon$, which implies that $\hat{V}\hat{V}^T = P^{\hat{A}}_{\ge \sigma_k - \epsilon}$.

Let $\hat{W}$ be a $d \times (d-k)$ matrix whose columns form an orthonormal basis for the orthogonal complement of the column span of $\hat{V}$. Note that if $\epsilon < \sigma_k/2$ then the kth singular value of $\hat{A}$ is strictly larger than $\sigma_k/2$ and the $(k+1)$th singular value is at most ϵ. Then $P^{\hat{A}}_{\le \epsilon} = \hat{W}\hat{W}^T$. By Wedin's theorem,

$$\|\hat{W}\hat{W}^T V V^T\| = \|P^{\hat{A}}_{\le\epsilon} P^A_{\ge \sigma_k}\| \le \frac{2\epsilon}{\sigma_k - \epsilon} \le \frac{4\epsilon}{\sigma_k}$$

Now, $\hat{W}^T$ and V have norm 1, and so it follows that

$$\|\hat{W}^T V\| = \|\hat{W}^T(\hat{W}\hat{W}^T V V^T)V\| \le \frac{4\epsilon}{\sigma_k}.$$

For any $u \in \mathbb{R}^k$ with $\|u\| = 1$, we have

$$\|\hat{V}^T V u\|^2 = 1 - \|\hat{W}^T V u\|^2 \ge 1 - 16\epsilon^2/\sigma_k^2,$$

from which the claimed lower bound follows. On the other hand, $\|\hat{V}^T V u\| \le u$ because both $\hat{V}^T$ and V have norm 1. ■

Let $M = D^{-1/2} V^T B V D^{-1/2}$ and $\hat{M} = \hat{D}^{-1/2}\hat{V}^T \hat{B} \hat{V} \hat{D}^{-1/2}$. Then M is the infinite-sample version of A's whitening matrix applied to B, and $\hat{M}$ is the finite-sample analogue. Recall from (7) that $O = D^{-1/2}\hat{V}^T V D^{1/2}$ is an almost-isometry of $\mathbb{R}^k$.

Lemma 17

$$\|OMO^T - \hat{M}\| \le C\frac{\epsilon\sigma_1}{\sigma_k^2}.$$

Proof The first step is to approximate OMO^T by $D^{-1/2}\hat{V}^T B\hat{V}D^{-1/2}$. To this end, note that

$$OMO^T = D^{-1/2}\hat{V}^T VV^T BVV^T\hat{V}D^{-1/2}.$$

Now, $\hat{V}$ is an isometry of $\mathbb{R}^k$ into $\mathbb{R}^n$; hence,

$$\|\hat{V}^T VV^T - \hat{V}^T\| = \|\hat{V}\hat{V}^T VV^T - \hat{V}\hat{V}^T\| = \|P^{\hat{A}}_{\geq\sigma_k-\epsilon}P^{A}_{\geq\sigma_k} - P^{\hat{A}}_{\geq\sigma_k-\epsilon}\| = \|P^{\hat{A}}_{\geq\sigma_k-\epsilon}P^{A}_{\leq 0}\|,$$

where the last equality used the fact that A has rank exactly k, and hence $I - P^{A}_{\geq\sigma_k} = P^{A}_{\leq 0}$. Now, Wedin's theorem applied to the computation above implies that

$$\|\hat{V}^T VV^T - \hat{V}^T\| \leq \frac{2\epsilon}{\sigma_k - \epsilon} \leq \frac{4\epsilon}{\sigma_k}$$

(recalling that $\epsilon \leq \sigma_k/4$).

Now, for general matrices $X, Y, \tilde{Y}, Z$ we have

$$\begin{aligned}\|X^T Y^T ZYX - X^T\tilde{Y}^T Z\tilde{Y}X\| &\leq \|X^T(Y-\tilde{Y})^T ZYX\| + \|X^T\tilde{Y}^T Z(Y-\tilde{Y})X\| \\ &\leq \|Y-\tilde{Y}\|\|X\|^2\|Z\|(\|Y\| + \|\tilde{Y}\|).\end{aligned}$$

We apply this with $X = D^{-1/2}$, $Y = \hat{V}$, $\tilde{Y} = \hat{V}VV^T$, and $Z = B$; since $\|D^{-1/2}\| = \sigma_k^{-1/2}$, $\|B\| \leq \sigma_1$, and $\|\hat{V}\|, \|V\|, \|V^T\| = 1$,

$$\|OMO^T - D^{-1/2}\hat{V}^T B\hat{V}D^{-1/2}\| \leq \frac{8\epsilon\sigma_1}{\sigma_k^2}$$

Next, we will replace B by $\hat{B}$ in the above inequality. Since $\|\hat{V}\| = \|\hat{V}^T\| = 1$ and $\|D^{-1/2}\| = \sigma_k^{-1/2}$,

$$\begin{aligned}\|D^{-1/2}\hat{V}^T B\hat{V}D^{-1/2} - D^{-1/2}\hat{V}^T\hat{B}\hat{V}D^{-1/2}\| &= \|D^{-1/2}\hat{V}^T(B-\hat{B})\hat{V}D^{-1/2}\| \\ &\leq \sigma_k^{-1}\|B-\hat{B}\| \leq \frac{\epsilon}{\sigma_k}.\end{aligned}$$

Putting this together with the previous bound yields

$$\|OMO^T - D^{-1/2}\hat{V}^T\hat{B}\hat{V}D^{-1/2}\| \leq \frac{\epsilon}{\sigma_k} + \frac{8\epsilon\sigma_1}{\sigma_k^2} \tag{8}$$

It remains to relate $D^{-1/2}\hat{V}^T\hat{B}\hat{V}D^{-1/2}$ to $\hat{M}$ (which is the same, but with $\hat{D}$ instead of D). Now, Weyl's inequality implies that

$$\|D^{-1/2} - \hat{D}^{-1/2}\| \leq \sigma_k^{-1/2} - (\sigma_k - \epsilon)^{-1/2} \leq \epsilon\sigma_k^{-3/2},$$

where the second inequality follows from a first-order Taylor expansion and the fact that $\epsilon \leq \sigma_k/2$. Hence,

$$\begin{aligned}\|D^{-1/2}\hat{V}^T\hat{B}\hat{V}D^{-1/2} - \hat{M}\| &\leq \|D^{-1/2} - \hat{D}^{-1/2}\|\|\hat{V}^T\hat{B}\hat{V}D^{-1/2}\| \\ &\quad + \|\hat{D}^{-1/2}\hat{V}^T\hat{B}\hat{V}\|\|D^{-1/2} - \hat{D}^{-1/2}\| \\ &\leq 4\epsilon\sigma_1\sigma_k^{-2}.\end{aligned}$$

Combining this with (8) and the triangle inequality, we have

$$\|OMO^T - \hat{M}\| = \frac{\epsilon}{\sigma_k} + 12\frac{\epsilon\sigma_1}{\sigma_k^2} \leq C\frac{\epsilon\sigma_1}{\sigma_k^2}.$$

■

Since O is almost an isometry, it follows that there is an orthogonal matrix $\tilde{O}$ that is close to O (for example, if $UDV^T = O$ is an SVD, let $\tilde{O} = UV^T$). In this way, we may find an orthogonal $\tilde{O}$ such that

$$\|O - \tilde{O}\| \leq 1 - \sqrt{1 - \frac{16\epsilon^2}{\sigma_k^2}} \leq \frac{16\epsilon^2}{\sigma_k^2}.$$

Now let u be the top eigenvector of M and let u_O be the top eigenvector of OMO^T. Then $\tilde{O}u$ is the top eigenvector of $\tilde{O}M\tilde{O}^T$. The triangle inequality implies that

$$\|OMO^T - \tilde{O}M\tilde{O}^T\| \leq 2\|M\|\|O - \tilde{O}\| \leq \frac{32\epsilon^2}{\sigma_k^2}\|M\|.$$

On the other hand, M was assumed to have a spectral gap of $\delta\|M\|$. By Wedin's theorem, it follows that

$$\|u - \tilde{O}^T u_O\| = \|\tilde{O}u - u_O\| \leq \frac{64\epsilon^2}{\delta\sigma_k^2}.$$

Finally, let $\hat{u}$ be the top eigenvector of $\hat{M}$. By Lemma 17 and Wedin's theorem,

$$\|\hat{u} - u_O\| \leq \frac{C\epsilon\sigma_1}{\delta\sigma_k^2}.$$

Then

$$\|Ou - \hat{u}\| \leq \|O - \tilde{O}\| + \|\tilde{O}u - \hat{h}\| \leq C\max\left\{\frac{\epsilon\sigma_1}{\delta\sigma_k^2}, \frac{\epsilon^2}{\delta\sigma_k^2}\right\} \leq \frac{C\epsilon\sigma_1}{\delta\sigma_k^2}, \tag{9}$$

where the last inequality follows because $\epsilon \leq \sigma_k/2 \leq \sigma_1/2$.

Next, we unpack O. Weyl's inequality implies that

$$\|D^{-1/2} - \hat{D}^{-1/2}\| \leq \sigma_k^{-1/2} - (\sigma_k - \epsilon)^{-1/2} \leq \epsilon\sigma_k^{-3/2},$$

where the second inequality follows from a first-order Taylor expansion and the fact that $\epsilon \leq \sigma_k/4$. Hence,

$$\|O - \hat{D}^{-1/2}\hat{V}^T V D^{1/2}\| \leq \|D^{1/2}\|\|D^{-1/2} - \hat{D}^{-1/2}\| \leq \frac{\epsilon\sqrt{\sigma_1}}{\sigma_k^{3/2}}.$$

The right hand side is smaller than $\frac{\epsilon\sigma_1}{\sigma_k^2}$, and so we may plug it into (9) to obtain

$$\|\hat{D}^{-1/2}\hat{V}^T V D^{1/2} u - \hat{u}\| \leq \frac{C\epsilon\sigma_1}{\delta\sigma_k^2}.$$

Finally, (again because $\epsilon \le \sigma_k/2$), $\|\hat{D}^{-1/2}\| \le (\sigma_k/2)^{-1/2}$, and so

$$\|VD^{1/2}u - \hat{V}\hat{D}^{1/2}\hat{u}\| \le \frac{C\epsilon\sigma_1}{\delta\sigma_k^{5/2}}. \tag{10}$$

Setting $w = VD^{1/2}u$ and $\hat{w} = \hat{V}\hat{D}^{1/2}\hat{u}$ and comparing this to the setting of Algorithm 1, (10) shows that the finite-sample algorithm gets almost the same w as the infinite-sample version.

It remains to check the last few lines of Algorithm 1; i.e., to see that we recover the right scaling of w.

Lemma 18 *Let M be a symmetric matrix of rank $k-1$ and let E be the span of its columns. Then $\|w\| \operatorname{dist}(w, E) \ge \sigma_k(M + ww^T)$.*

Proof It suffices to consider the case $\|w\| = 1$ (for a general w, apply the special case of the lemma to $w/\|w\|$ and $M/\|w\|^2$). Let P_E denote the orthogonal projection onto E, and note that $\|w - P_E w\| = \operatorname{dist}(w, E)$ Let $F = \operatorname{span}\{E, w\}$. Since F has dimension k and $y \in F^\perp$ implies $\|(M + ww^T)y\| = 0$, it suffices to find some $y \in F$ such that $\|(M + ww^T)y\| \le \operatorname{dist}(w, E)\|y\|$. Choose $y = w - P_E w$. Then $My = 0$ and so

$$\|(M + ww^T)y\| = |w^T y| = \|w - P_E w\|^2 = \operatorname{dist}(w, E)\|y\|.$$

■

Lemma 19 *Let E be a subspace and take $w \notin E$. For $x \in \operatorname{span}\{E, w\}$, let $a(x) \in \mathbb{R}$ be the unique solution to $x = aw + e$, $e \in E$. Then $|a(x) - a(y)| \le \|x - y\| / \operatorname{dist}(w, E)$.*

Proof Given $x, y \in \operatorname{span}\{E, w\}$, we can write $x - y = (a(x) - a(y))w + e$, where $e \in E$. It follows that

$$\begin{aligned} \|x - y\| &= \|(a(x) - a(y))w + e\| \ge \inf_{e \in E} \|(a(x) - a(y))w + e\| \\ &= |a(x) - a(y)| \operatorname{dist}(w, E). \end{aligned}$$

■

Finally, we apply the preceding two lemmas to show that $\hat{\alpha}_1$ is accurate in Algorithm 1. Together with (10) (whose right hand side provides the value of η that we will use), this completes the proof of Theorem 1.

Lemma 20 *Let $m = \sum_i \alpha_i \mu_i$. If $\|\hat{A} - A\| \le \epsilon$, $\|\hat{m} - m\| \le \epsilon$ and $\|\hat{w} - \sqrt{\alpha_1}\mu_1\| \le \eta$ then*

$$|\hat{\alpha}_1 - \alpha_1| \le \frac{C\sqrt{\alpha_1}|\alpha_1 R + \eta|}{\sigma_k}\left(\eta + R\frac{\epsilon}{\sigma_k} + \epsilon\right),$$

where $R = \max_i \|\mu_i\|$, provided that the right hand side above is at most α_1.

Proof By Wedin's theorem,

$$\|VV^T - \hat{V}\hat{V}^T\| \le \frac{2\|\hat{A} - A\|}{\sigma_k - \|\hat{A} - A\|} \le 4\frac{\epsilon}{\sigma_k}$$

if $\epsilon \le \sigma_k/2$. Hence,

$$\begin{aligned} \|m - \hat{V}\hat{V}^T\hat{m}\| &= \|VV^Tm - \hat{V}\hat{V}^T\hat{m}\| \\ &\le \|(VV^T - \hat{V}\hat{V}^T)m\| + \|\hat{V}\hat{V}^T(m - \hat{m})\| \\ &\le 4\frac{\epsilon}{\sigma_k}\|m\| + \epsilon. \end{aligned}$$

Now, let $y = \sqrt{\alpha_1}\hat{w} + \hat{V}\hat{V}^T\sum_{i=2}^k \alpha_i\mu_i$. Then

$$\begin{aligned} \|m - y\| &\le \sqrt{\alpha_1}\|\hat{w} - \sqrt{\alpha_1}\mu_1\| + \left\|\sum_{i=2}^k \alpha_i(\mu_i - \hat{V}\hat{V}^T\mu_i)\right\| \\ &\le \eta + \max_i \|\mu_i\|\|VV^T - \hat{V}\hat{V}^T\| \\ &\le \eta + 4\max_i \|\mu_i\|\frac{\epsilon}{\sigma_k}. \end{aligned}$$

Defining $R = \max_i \|\mu_i\|$, we have

$$\|y - \hat{V}\hat{V}^T\hat{m}\| \le \eta + 8R\frac{\epsilon}{\sigma_k} + \epsilon.$$

Now, let $\hat{E}$ be the span of $\{\hat{V}\hat{D}^{1/2}v : v \in \mathbb{R}^k, v \perp \hat{u}\}$, and note that $\hat{E}$ may also be written as the column space of $\hat{V}\hat{D}^{1/2}(I_k - \hat{u}\hat{u}^T)\hat{D}^{1/2}\hat{V}^T = \hat{V}\hat{D}\hat{V}^T - \hat{w}\hat{w}^T$. Since $\hat{V}\hat{D}^{1/2}$ is injective, $\hat{E}$ has dimension $k-1$ and does not contain $\hat{w} = \hat{V}\hat{D}^{1/2}\hat{u}$. Hence, $y = \sqrt{\alpha_1}\hat{w} + e$ is the unique way to decompose y in $\mathrm{span}\{\hat{w}\} \oplus \hat{E}$. If we define a by the decomposition $\hat{m} = a\hat{w} + e$ then Lemma 19 implies

$$\begin{aligned} |a - \sqrt{\alpha_1}| &\le \|y - \hat{m}\|/\operatorname{dist}(\hat{w}, \hat{E}) \\ &\le \frac{1}{\operatorname{dist}(\hat{w}, \hat{E})}\left(\eta + 8R\frac{\epsilon}{\sigma_k} + \epsilon\right). \end{aligned}$$

On the other hand, Lemma 18 applied to $\hat{V}\hat{D}\hat{V}^T - \hat{w}\hat{w}^T$ and $\hat{w}$ implies (because the kth singular value of $\hat{V}\hat{D}\hat{V}^T \ge \sigma_k - \epsilon \ge \sigma_k/2$) that $\|\hat{w}\|\operatorname{dist}(\hat{w}, \hat{E}) \ge \sigma_k/2$. Therefore,

$$|a - \sqrt{\alpha_1}| \le \frac{2\|\hat{w}\|}{\sigma_k}\left(\eta + 8R\frac{\epsilon}{\sigma_k} + \epsilon\right) \le \frac{2(\alpha_1\|\mu_1\| + \eta)}{\sigma_k}\left(\eta + 8R\frac{\epsilon}{\sigma_k} + \epsilon\right).$$

Finally, note that $|\hat{\alpha}_1 - \alpha_1| = |a^2 - \alpha_1| = |a - \sqrt{\alpha_1}|(a + \sqrt{\alpha_1})$. We consider two cases: if $a \le C\sqrt{\alpha_1}$ then $|\hat{\alpha}_1 - \alpha_1| \le (1 + C)\sqrt{\alpha_1}|a - \sqrt{\alpha_1}|$, which completes the proof. In the other case, we have

$$|\hat{\alpha}_1 - \alpha_1| \sim \hat{\alpha}_1 \le C\sqrt{\hat{\alpha}_1}|a - \sqrt{\alpha_1}|,$$

which implies that

$$|\hat{\alpha}_1 - \alpha_1| \le C|a - \sqrt{\alpha_1}|^2$$

for some other constant C. This implies

$$|\hat{\alpha}_1 - \alpha_1| \leq C\left[\frac{(\alpha_1 R + \eta)}{\sigma_k}\left(\eta + R\frac{\epsilon}{\sigma_k} + \epsilon\right)\right]^2 \leq C\sqrt{\alpha_1}\left[\frac{(\alpha_1 R + \eta)}{\sigma_k}\left(\eta + R\frac{\epsilon}{\sigma_k} + \epsilon\right)\right],$$

where the second inequality comes from the assumption that the right hand side in the lemma is bounded by α_1. ■

As we pointed out in Section 2, spectral algorithms similar to Algorithm 1 has been proposed before for GMM [Hsu and Kakade 2013] and LDA [Anandkumar et al. 2012] models, the main difference being how the second matrix (equivalent to B) is constructed. Since the underlying whitening procedure is the same in all these algorithms, the proof approach presented above is similar to those in Hsu and Kakade (2013); Anandkumar et al. (2012). The proofs diverge when computing the perturbation of the second matrix, matrix B in our algorithm, which introduces different dependence on various parameter models in the overall error bound. For example the error bound in Theorem 4.1 of Anandkumar et al. (2012) has a slightly worse dependence on k and σ_k than Theorem 1.

Appendix E. Finite-sample Analysis of the Cancellation Method

In this section we analyze the performance of Algorithm 2 when we have finite sample estimates of the matrices A, B and vector m. For ease of exposition we replaced the quantities $V_{1:(k-1)}, v_i, a_i, c_i$ in Algorithm 2 with the notation representing estimate $\widehat{V}_{1:(k-1)}, \hat{v}_i, \hat{a}_i, \hat{c}_i$ respectively, since these are computed from sample estimates $\widehat{A}, \widehat{B}$. First, we show in Lemma 21 that we can have a good estimate for $\widehat{Z}_{\lambda^*}$ using good estimates for A, B and λ_1.

Lemma 21 *Let* $\widehat{Z}_\lambda = \widehat{A} - \lambda\widehat{B}, Z_\lambda = A - \lambda B$. *Suppose* $\max\{\|\widehat{A} - A\|, \|\widehat{B} - B\|\} < \epsilon$ *and* $\lambda_1 = 1/w_1$. *Then,*

$$\|\widehat{Z}_\lambda - Z_{\lambda_1}\| < \epsilon\left(2 + \frac{1}{w_1}\right) + \epsilon_1\sigma_1(B)$$

when $|\lambda_1 - \lambda| < \epsilon_1 < 1$.

Proof We have,

$$\begin{aligned}
\|\widehat{Z}_\lambda - Z_{\lambda_1}\| &\leq \|\widehat{A} - A\| + \|\lambda\widehat{B} - \lambda_1 B\| \\
&< \|\widehat{A} - A\| + \lambda_1\|\widehat{B} - B\| + |\lambda_1 - \lambda|\|\widehat{B}\| \\
&\leq \epsilon + \lambda_1\epsilon + \epsilon_1(\sigma_1(B) + \epsilon) \\
&< \epsilon(1 + 1/w_1 + \epsilon_1) + \epsilon_1\sigma_1(B) < \epsilon\left(2 + \frac{1}{w_1}\right) + \epsilon_1\sigma_1(B)
\end{aligned}$$

since $\epsilon_1 < 1$. ■

The following lemma will show that even with noisy estimates of A, B, the estimated λ^* is close to λ_1.

Lemma 22 *Let* $\max\{\|\widehat{A}-A\|, \|\widehat{B}-B\|\} < \epsilon < \sigma_k(A)/2$, *and* $\lambda_1 = 1/w_1 > 0$. *Then,*

$$|\lambda^* - \lambda_1| = O(\epsilon)$$

Proof Define $Z'_\lambda = VV^TAVV^T - \lambda VV^TBVV^T$, V being the $d \times k$ matrix of top k eigenvectors of A. The corresponding empirical estimate $\widehat{Z}'_\lambda = \widehat{V}\widehat{V}^T\widehat{A}\widehat{V}\widehat{V}^T - \lambda\widehat{V}\widehat{V}^T\widehat{B}\widehat{V}\widehat{V}^T$. The main proof idea is the following. We try to find $\lambda_2, \lambda_3 > 0$ such that:

1. $\forall \lambda > \lambda_2$, $\widehat{Z}'_\lambda$ is not PSD.
2. $\forall \lambda < \lambda_3$, $\widehat{Z}'_\lambda$ is PSD.

The above two conditions imply that the optimum λ^* is bounded as $\lambda_3 \le \lambda^* \le \lambda_2$. We then simply bound $\lambda^* - \lambda_1$ as $\lambda_3 - \lambda_1 \le \lambda^* - \lambda_1 \le \lambda_2 - \lambda_1$. We now elaborate the above two steps. First, we bound the perturbation of empirical matrix $\widehat{Z}'_\lambda$ as follows. Using Wedin's theorem we have $\|\widehat{V}\widehat{V}^T - VV^T\| \le \frac{4\epsilon}{\sigma_k(A)}$. Using this and the theorem assumptions we can compute the following bounds.

$$\begin{aligned}
\|\widehat{V}\widehat{V}^T\widehat{A}\widehat{V}\widehat{V}^T - VV^TAVV^T\| &\le 13\epsilon \\
\|\widehat{V}\widehat{V}^T\widehat{B}\widehat{V}\widehat{V}^T - VV^TBVV^T\| &\le \left(1 + \frac{12\sigma_k(B)}{\sigma_k(A)}\right)\epsilon
\end{aligned}$$

Combining, we have

$$\|\widehat{Z}'_\lambda - Z'_\lambda\| \le \|\widehat{V}\widehat{V}^T\widehat{A}\widehat{V}\widehat{V}^T - VV^TAVV^T\| + \lambda\|\widehat{V}\widehat{V}^T\widehat{B}\widehat{V}\widehat{V}^T - VV^TBVV^T\| \le c_1(1+\lambda)\epsilon \quad (11)$$

where $c_1 = \max\{13, 1 + \frac{12\sigma_k(B)}{\sigma_k(A)}\}$.

Step 1: Since matrices A and B share the same column and row space, $VV^TAVV^T = A$, $VV^TBVV^T = B$, and $Z'_\lambda = Z_\lambda = \sum_{i=1}^k (1 - \lambda w_i)\alpha_i\mu_i\mu_i^T$, $w_i = \langle \mu_i, v\rangle$. Recall, $\mathcal{V} = \text{span}\{\mu_2, \ldots, \mu_k\}$ and Π denote the projection onto $\mathcal{V}_\perp$, its perpendicular space. Let $x_1 = \Pi\mu_1/\|\Pi\mu_1\|$, and $x_1 = V\tilde{x}_1$, $\|x_1\| = \|\tilde{x}_1\| = 1$. Consider the eigenvalues of the $k \times k$ Hermitian matrix $V^TZ_\lambda V$. Using variational theorem we can write:

$$\tilde{\sigma}_k(V^TZ_\lambda V) = \min_{x \neq 0, \|x\|=1} x^TV^TZ_\lambda Vx \le \tilde{x}_1^TV^TZ_\lambda V\tilde{x}_1 = x_1^TZ_\lambda x_1 = (1 - \lambda w_1)\alpha_1 a'_1 \quad (12)$$

where $a'_1 = |\langle x_1, \mu_1\rangle|^2 > 0$. Now note that the matrices $Z'_\lambda = VV^TZ_\lambda VV^T$ and $V^TZ_\lambda V$ have the same set of non-zero eigenvalues since V forms an orthonormal basis of the row/column space of Z_λ. Therefore we can write from above,

$$\tilde{\sigma}_k(Z'_\lambda) = \tilde{\sigma}_k(V^TZ_\lambda V) \le (1 - \lambda w_1)\alpha_1 a'_1 \quad (13)$$

For $\lambda = \lambda_1 = 1/w_1$, Z'_{λ_1} is a rank $k-1$ matrix, and for any $\lambda > \lambda_1$, Z'_λ has at least one negative eigenvalue. Consider $\lambda_2 > \lambda_1$ such that Z'_{λ_2} has one negative eigenvalue and $k-1$ positive eigenvalues. Since $\widehat{Z}'_{\lambda_2}, Z'_{\lambda_2}$ are symmetric matrices, using Weyl's inequality we get,

$$\begin{aligned}
\tilde{\sigma}_k(\widehat{Z}'_{\lambda_2}) &\le \tilde{\sigma}_k(Z'_{\lambda_2}) + \|\widehat{Z}'_{\lambda_2} - Z'_{\lambda_2}\| \le \tilde{\sigma}_k(Z'_{\lambda_2}) + c_1(1+\lambda_2)\epsilon \\
&\le (1 - \lambda_2 w_1)\alpha_1 a'_1 + c_1(1+\lambda_2)\epsilon \\
&\le a'_1[(\alpha_1 + \epsilon) - \lambda_2(w_1\alpha_1 - \epsilon)] \quad (14)
\end{aligned}$$

using equations (11), (13), and assuming $a'_1 > c_1$ (else we can simply rescale ϵ). Now for any $\lambda > \lambda_2 = \frac{\alpha_1+\epsilon}{\alpha_1 w_1-\epsilon}$ we get

$$\tilde{\sigma}_k(\widehat{Z}'_\lambda) \leq a'_1[(\alpha_1+\epsilon) - \lambda(w_1\alpha_1 - \epsilon)] \leq a'_1[(\alpha_1+\epsilon) - \lambda_2(w_1\alpha_1 - \epsilon)] = 0$$

Therefore, when $\lambda > \lambda_2 = \frac{\alpha_1+\epsilon}{\alpha_1 w_1-\epsilon}$, $\widehat{Z}'_\lambda$ is not PSD. This implies that $\lambda_2 \geq \lambda^*$. Then,

$$\lambda^* - \lambda_1 \leq \lambda_2 - \lambda_1 = \frac{\alpha_1+\epsilon}{\alpha_1 w_1 - \epsilon} - \frac{1}{w_1} = \frac{\epsilon(w_1+1)}{(\alpha_1 w_1 - \epsilon)w_1} \tag{15}$$

Step 2: Consider $\lambda_3 < \lambda_1$ such that Z'_{λ_3} is PSD. Then we lower bound $\tilde{\sigma}_k(Z'_{\lambda_3})$ as follows. Let $\tilde{v}_{k,\lambda_3}$ be the $k-$th eigenvector of Z'_{λ_3} having eigenvalue $\tilde{\sigma}_k(Z'_{\lambda_3})$. Then,

$$\begin{aligned} \tilde{\sigma}_k(Z'_{\lambda_3}) &= \tilde{v}^T_{k,\lambda_3} Z'_{\lambda_3} \tilde{v}_{k,\lambda_3} = \sum_{i=1}^k \alpha_i(1-\lambda_3 w_i)\tilde{v}^T_{k,\lambda_3}\mu_i\mu_i^T\tilde{v}_{k,\lambda_3} \\ &\geq (1-\lambda_3 w_1)\sum_{i=1}^k \alpha_i|\langle \tilde{v}_{k,\lambda_3}, \mu_i\rangle|^2 \geq (1-\lambda_3 w_1)a'_2 \end{aligned} \tag{16}$$

since $w_1 > w_i$, $i \neq 1$, and where $a'_2 = \inf_{\lambda\geq 0}\sum_{i=1}^k \alpha_i|\langle \tilde{v}_{k,\lambda}, \mu_i\rangle|^2 > 0$. Now using the lower bound of Weyl's inequality,

$$\begin{aligned} \tilde{\sigma}_k(\widehat{Z}'_{\lambda_3}) &\geq \tilde{\sigma}_k(Z'_{\lambda_3}) - \|\widehat{Z}'_{\lambda_3} - Z'_{\lambda_3}\| \\ &\geq \tilde{\sigma}_k(Z'_{\lambda_3}) - c_1(1+\lambda_3)\epsilon \\ &\geq (1-\lambda_3 w_1)a'_2 - c_1(1+\lambda_3)\epsilon \\ &\geq c_1[(1-\epsilon) - \lambda_3(w_1+\epsilon)] \end{aligned}$$

using equation (16), and assuming $c_1 < a'_2$ (else we can simply rescale ϵ). Then, for any $\lambda < \lambda_3 = \frac{(1-\epsilon)}{(w_1+\epsilon)}$ we have $\tilde{\sigma}_k(\widehat{Z}'_\lambda) > 0$, or $\widehat{Z}'_\lambda$ is PSD. This implies $\lambda^* > \lambda_3$. Therefore,

$$\lambda^* - \lambda_1 \geq \lambda_3 - \lambda_1 = \frac{(1-\epsilon)}{(w_1+\epsilon)} - \frac{1}{w_1} = -\frac{(w_1+1)\epsilon}{(w_1+\epsilon)w_1} \tag{17}$$

Combining equations (15), (17) we get,

$$|\lambda^* - \lambda_1| \leq c_3\epsilon = O(\epsilon)$$

where $c_3 = \max\left(\frac{(w_1+1)}{(w_1+\epsilon)w_1}, \frac{(w_1+1)}{(\alpha_1 w_1-\epsilon)w_1}\right)$. ■

In Lemma 22 we assume $w_1 = \langle \mu_1, v\rangle$ is positive. When $w_1 < 0$, we have to modify the line search and find the smallest $\lambda < 0$ such that $\widehat{Z}'_\lambda$ is PSD. However we can still apply similar arguments and prove that as long as the estimates of A, B, are within ϵ in spectral norm, Algorithm 2 can estimate λ^* within an $O(\epsilon)$ accuracy of λ_1. Lemma 21 and 22 together implies that $\|\widehat{Z}_{\lambda^*} - Z_{\lambda_1}\| = O(\epsilon)$ as follows, which will be used to prove Theorem 3. We have,

$$\begin{aligned}\|\widehat{Z}_{\lambda^*} - Z_{\lambda_1}\| &< \epsilon\left(2 + \frac{1}{w_1}\right) + \epsilon_1 \sigma_1(B) \\ &\leq \epsilon\left(2 + \frac{1}{w_1}\right) + c_3 \epsilon \sigma_1(B) \\ &\leq 3\eta_3 \epsilon \end{aligned} \tag{18}$$

where in the last inequality we assume $\epsilon < \alpha_1 w_1/2$, and $\eta_3 = \max\left\{2, \frac{1}{w_1}, c_3\sigma_1(B)\right\}$.

Lemma 23 *Let $\|\hat{m} - m\| < \epsilon$, $\|\widehat{Z}_{\lambda^*} - Z_{\lambda_1}\| < \epsilon_2 < \sigma_{k-1}(Z_{\lambda_1})/2$ for $\lambda_1 = \alpha_1/\beta_1$. $V_{1:(k-1)}$ denote the $d \times (k-1)$ matrix of $k-1$ largest singular vectors of Z_{λ_1} and $\widehat{V}_{1:(k-1)}$ be the $d \times (k-1)$ matrix of $k-1$ largest singular vectors of $\widehat{Z}_{\lambda^*}$. Then,*

$$\begin{aligned}\|\hat{x}_1 - x_1\| &< 2\epsilon + \frac{4\epsilon_2 R}{\sigma_{k-1}(Z_{\lambda_1})} = \epsilon_3 \\ \|\hat{v}_1 - v_1\| &< \frac{2\epsilon_3}{\alpha_1 a_1} = \epsilon_4\end{aligned}$$

where $R = \max_{i\in[k]} \|\mu_i\|$.

Proof Since, $\|\widehat{Z}_{\lambda^*} - Z_{\lambda_1}\| < \epsilon_2 < \sigma_{k-1}(Z_{\lambda_1})/2$, applying Wedin's theorem we get,

$$\|\widehat{V}_{1:(k-1)}\widehat{V}^T_{1:(k-1)} - V_{1:(k-1)}V^T_{1:(k-1)}\| \leq \frac{2\|\widehat{Z}_{\lambda^*} - Z_{\lambda_1}\|}{\sigma_{k-1}(Z_{\lambda_1}) - \|\widehat{Z}_{\lambda^*} - Z_{\lambda_1}\|} \leq \frac{4\epsilon_2}{\sigma_{k-1}(Z_{\lambda_1})} \tag{19}$$

since $\epsilon_2 < \sigma_{k-1}(Z_{\lambda_1})/2$. Now,

$$\begin{aligned}\|\hat{x}_1 - x_1\| &= \|\hat{m} - \widehat{V}_{1:(k-1)}\widehat{V}^T_{1:(k-1)}\hat{m} - m + V_{1:(k-1)}V^T_{1:(k-1)}m\| \\ &\leq \|\hat{m} - m\| + \|(\widehat{V}_{1:(k-1)}\widehat{V}_{1:(k-1)} - V_{1:(k-1)}V^T_{1:(k-1)})m\| + \|\widehat{V}_{1:(k-1)}\widehat{V}^T_{1:(k-1)}(m - \hat{m})\| \\ &< 2\|m - \hat{m}\| + \frac{4\epsilon_2\|m\|}{\sigma_{k-1}(Z_{\lambda_1})} < 2\epsilon + \frac{4\epsilon_2 R}{\sigma_{k-1}(Z_{\lambda_1})} := \epsilon_3\end{aligned}$$

where we used equation 19 and $\|m\| \leq R$. Recall that $x_1 = \alpha_1 \prod_{\mathcal{V}} \mu_1 = \alpha_1 a_1 v_1$, where $\mathcal{V} = \text{span}\{\mu_2, \ldots, \mu_k\}$ and $a_1 = \langle \mu_1, v_1 \rangle$. To show the second bound,

$$\begin{aligned}\|\hat{v}_1 - v_1\| &= \left\|\frac{\hat{x}_1}{\|\hat{x}_1\|} - \frac{x_1}{\|x_1\|}\right\| \\ &\leq \frac{\|\hat{x}_1 - x_1\|}{\|x_1\|} + \|\hat{x}_1\|\left|\frac{1}{\|x_1\|} - \frac{1}{\|\hat{x}_1\|}\right| \\ &< \frac{\|\hat{x}_1 - x_1\|}{\|x_1\|} + \frac{|\|\hat{x}_1\| - \|x_1\||}{\|x_1\|} \leq 2\frac{\|\hat{x}_1 - x_1\|}{\|x_1\|} \\ &< \frac{2\epsilon_3}{\alpha_1 a_1} := \epsilon_4\end{aligned}$$

■

Lemma 24 *Let $\|\widehat{A} - A\| < \epsilon, \|\hat{v}_1 - v_1\| < \epsilon_4$. Define $d \times k$ matrices $V = [v_1 V_{1:(k-1)}]$ and $\widehat{V} = [\hat{v}_1 \widehat{V}_{1:(k-1)}]$. Then,*

$$\|\widehat{V}\widehat{V}^T \widehat{A}\hat{v}_1 - VV^T A v_1\| < \sigma_1(A)\left(3\epsilon_4 + \frac{4\epsilon}{\sigma_{k-1}(Z_{\lambda_1})}\right) + \epsilon(1+\epsilon_4)$$

Proof Similar to Lemma 23 we have from Wedin's theorem $\|\widehat{V}_{1:(k-1)}\widehat{V}^T_{1:(k-1)} - V_{1:(k-1)}V^T_{1:(k-1)}\| < \frac{4\epsilon}{\sigma_{k-1}(Z_{\lambda_1})}$. Then we can bound,

$$\begin{aligned}
\|\widehat{V}\widehat{V}^T - VV^T\| &\leq \|\hat{v}_1\hat{v}_1^T - v_1 v_1^T\| + \|\widehat{V}_{1:(k-1)}\widehat{V}_{1:(k-1)} - V_{1:(k-1)}V^T_{1:(k-1)}\| \\
&< 2\|\hat{v}_1 - v_1\| + \frac{4\epsilon}{\sigma_{k-1}(Z_{\lambda_1})} \\
&< 2\epsilon_4 + \frac{4\epsilon}{\sigma_{k-1}(Z_{\lambda_1})} \qquad (20)
\end{aligned}$$

Now,

$$\begin{aligned}
\|\widehat{V}\widehat{V}^T \widehat{A}\hat{v}_1 - VV^T A v_1\| &\leq \|(\widehat{V}\widehat{V}^T - VV^T)Av_1\| + \|\widehat{V}\widehat{V}^T(A - \widehat{A})v_1\| \\
&\quad + \|\widehat{V}\widehat{V}^T\widehat{A}(v_1 - \hat{v}_1)\| \\
&\leq \|\widehat{V}\widehat{V}^T - VV^T\|\|A\| + \|A - \widehat{A}\| + \|\widehat{A}\|\|v_1 - \hat{v}_1\| \\
&< \sigma_1(A)\left(2\epsilon_4 + \frac{4\epsilon}{\sigma_{k-1}(Z_{\lambda_1})}\right) + \epsilon + (\sigma_1(A) + \epsilon)\epsilon_4
\end{aligned}$$

where we use inequality (20), $\|Av_1\| \leq \sigma_1(A)$ as v_1 is unit norm, $\|\widehat{V}\widehat{V}^T\| < 1$ since $\widehat{V}$ is orthonormal, and $\|\widehat{A}\| < \|A\| + \epsilon$. Combining,

$$\|\widehat{V}\widehat{V}^T \widehat{A}\hat{v}_1 - VV^T A v_1\| < \sigma_1(A)\left(3\epsilon_4 + \frac{4\epsilon}{\sigma_{k-1}(Z_{\lambda_1})}\right) + \epsilon(1+\epsilon_4)$$

■

Lemma 25 *Let $\|\widehat{A} - A\| < \epsilon$, $\|\hat{x}_1 - x_1\| < \epsilon_3 < \frac{\alpha_1 a_1}{2}$, and $\|\hat{v}_1 - v_1\| < \epsilon_4$. Then,*

$$|\hat{a}_1 - a_1| < \frac{\alpha_1 a_1\left(2\sigma_1(A)\epsilon_4 + \epsilon(1+\epsilon_4)\right) + 2(\sigma_1(A)+\epsilon)\epsilon_3}{\alpha_1^2 a_1^2}$$

Proof We first compute,

$$\begin{aligned}
|\hat{v}_1^T\widehat{A}\hat{v}_1 - v_1^T A v_1| &\leq |(v_1^T - \hat{v}_1^T)Av_1| + |\hat{v}_1^T(A - \widehat{A})v_1| + |\hat{v}_1^T\widehat{A}(v_1 - \hat{v}_1)| \\
&\leq \|v_1^T - \hat{v}_1^T\|\sigma_1(A) + \|A - \widehat{A}\| + \sigma_1(\widehat{A})\|v_1 - \hat{v}_1\| \\
&< \sigma_1(A)\epsilon_4 + \epsilon + (\sigma_1(A)+\epsilon)\epsilon_4 = 2\sigma_1(A)\epsilon_4 + \epsilon(1+\epsilon_4) \qquad (21)
\end{aligned}$$

using the fact that $v_1, \hat{v}_1$ have unit norms. Now we can bound the error $|\hat{a}_1 - a_1|$ as follows.

$$\begin{aligned} |\hat{a}_1 - a_1| &= \left| \frac{\hat{v}_1^T \widehat{A} \hat{v}_1}{\|\hat{x}_1\|} - \frac{v_1^T A v_1}{\|x_1\|} \right| \\ &\leq \frac{1}{\|x_1\|} |\hat{v}_1^T \widehat{A} \hat{v}_1 - v_1^T A v_1| + |\hat{v}_1^T \widehat{A} \hat{v}_1| \frac{|\|x_1\| - \|\hat{x}_1\||}{\|x_1\|\|\hat{x}_1\|} \end{aligned}$$

From equation (21) and using $|\|x_1\| - \|\hat{x}_1\|| < \|\hat{x}_1 - x_1\| < \epsilon_3, \|x_1\| = \alpha_1 a_1$ we get,

$$\begin{aligned} |\hat{a}_1 - a_1| &< \frac{2\sigma_1(A)\epsilon_4 + \epsilon(1+\epsilon_4)}{\alpha_1 a_1} + \frac{(\sigma_1(A)+\epsilon)\epsilon_3}{\alpha_1 a_1(\alpha_1 a_1 - \epsilon_3)} \\ &< \frac{\alpha_1 a_1 \left(2\sigma_1(A)\epsilon_4 + \epsilon(1+\epsilon_4)\right) + 2(\sigma_1(A)+\epsilon)\epsilon_3}{\alpha_1^2 a_1^2} \end{aligned}$$

since $\epsilon_3 < \frac{\alpha_1 a_1}{2}$. ■

Note that from Lemma 23 taking $\frac{2\epsilon_3}{\alpha_1 a_1} = \epsilon_4$ the above bound becomes $|\hat{a}_1 - a_1| < \frac{6\sigma_1(A)\epsilon_3 + \epsilon\alpha_1 a_1 + 4\epsilon\epsilon_3}{\alpha_1^2 a_1^2}$.

E.1 Proof of Theorem 3

We now proof Theorem 3. Assume $\|\widehat{Z}_{\lambda^*} - Z_{\lambda_1}\| \leq \epsilon_2$. Under the assumptions we have using Lemma 23 $\|\hat{x}_1 - x_1\| < \epsilon_3 = 2\epsilon + \frac{4\epsilon_2 R}{\sigma_{k-1}(Z_{\lambda_1})}$, $\|\hat{v}_1 - v_1\| < \epsilon_4 = \frac{2\epsilon_3}{\alpha_1 a_1}$. Also from Lemma 24 we have $\|\widehat{V}\widehat{V}^T \widehat{A} \hat{v}_1 - VV^T A v_1\| < \sigma_1(A)\left(3\epsilon_4 + \frac{4\epsilon}{\sigma_{k-1}(Z_{\lambda_1})}\right) + \epsilon(1+\epsilon_4)$. Using these we compute the first bound as follows.

$$\begin{aligned} \|\hat{\mu}_1 - \mu_1\| &= \left\| \frac{\widehat{V}\widehat{V}^T \widehat{A} \hat{v}_1}{\|\hat{x}_1\|} - \frac{VV^T A v_1}{\|x_1\|} \right\| \\ &\leq \|\widehat{V}\widehat{V}^T \widehat{A} \hat{v}_1\| \left| \frac{1}{\|\hat{x}_1\|} - \frac{1}{\|x_1\|} \right| + \frac{1}{\|x_1\|} \|\widehat{V}\widehat{V}^T \widehat{A} \hat{v}_1 - VV^T A v_1\| \\ &\leq \|\widehat{A}\| \frac{\|\hat{x}_1 - x_1\|}{\|\hat{x}_1\|\|x_1\|} + \frac{1}{\|x_1\|} \|\widehat{V}\widehat{V}^T \widehat{A} \hat{v}_1 - VV^T A v_1\| \end{aligned}$$

Now using bounds from Lemma 23, 24 we get,

$$\begin{aligned} \|\hat{\mu}_1 - \mu_1\| &< \frac{(\sigma_1(A)+\epsilon)\epsilon_3}{\alpha_1 a_1(\alpha_1 a_1 - \epsilon_3)} + \frac{\sigma_1(A)\left(3\epsilon_4 + \frac{4\epsilon}{\sigma_{k-1}(Z_{\lambda_1})}\right) + \epsilon(1+\epsilon_4)}{\alpha_1 a_1} \\ &< \frac{2}{\alpha_1^2 a_1^2} \left[(\sigma_1(A)+\epsilon)\epsilon_3 + \alpha_1 a_1 \left((3\sigma_1(A)+\epsilon)\epsilon_4 \right. \right. \\ &\quad \left. \left. + \epsilon\left(1 + 4\sigma_1(A)/\sigma_{k-1}(Z_{\lambda_1})\right)\right)\right] \\ &< \frac{2}{\alpha_1^2 a_1^2} \left[(\sigma_1(A)+\epsilon)\epsilon_3 + 2\left(3\sigma_1(A)+\epsilon\right)\epsilon_3 \right. \\ &\quad \left. + \alpha_1 a_1 \epsilon \left(1 + 4\sigma_1(A)/\sigma_{k-1}(Z_{\lambda_1})\right)\right] \\ &\leq 2\frac{10\sigma_1(A)\epsilon_3 + 5\alpha_1 a_1 \epsilon \frac{\sigma_1(A)}{\sigma_{k-1}(Z_{\lambda_1})}}{\alpha_1^2 a_1^2} \end{aligned}$$

assuming $\epsilon_3 \leq \frac{\alpha_1 a_1}{2}$, $\sigma_1(A) \geq \epsilon$, and $\sigma_1(A) > \sigma_{k-1}(Z_{\lambda_1})$. Now expanding ϵ_3 and rearranging terms we have,

$$\begin{aligned}
\|\hat{\mu}_1 - \mu_1\| &< \frac{1}{\alpha_1^2 a_1^2}\left(\left(40 + 10\frac{\alpha_1 a_1}{\sigma_{k-1}(Z_{\lambda_1})}\right)\sigma_1(A)\epsilon + 80\frac{\sigma_1(A)R\epsilon_2}{\sigma_{k-1}(Z_{\lambda_1})}\right) \\
&< \frac{80}{\alpha_1^2 a_1^2}\left(\sigma_1(A)\epsilon\left(1 + \frac{\alpha_1 a_1}{\sigma_{k-1}(Z_{\lambda_1})}\right) + \frac{\sigma_1(A)\epsilon_2 R}{\sigma_{k-1}(Z_{\lambda_1})}\right) \qquad (22)
\end{aligned}$$

To prove the second bound from Lemma 25 and assuming $\epsilon < \sigma_1(A)$ we have $|\hat{a}_1 - a_1| \leq \frac{10\sigma_1(A)\epsilon_3 + \alpha_1 a_1 \epsilon}{\alpha_1^2 a_1^2}$. Then,

$$\begin{aligned}
\hat{a}_1(\alpha_1 - \hat{\alpha}_1) &= \hat{a}_1\alpha_1 - \hat{a}_1\hat{\alpha}_1 \\
&= a_1\alpha_1 - \hat{a}_1\hat{\alpha}_1 + \hat{a}_1\alpha_1 - a_1\alpha_1 \\
\hat{a}_1|\alpha_1 - \hat{\alpha}_1| &\leq |a_1\alpha_1 - \hat{a}_1\hat{\alpha}_1| + \alpha_1|\hat{a}_1 - a_1| \\
|\alpha_1 - \hat{\alpha}_1| &\leq \frac{1}{\hat{a}_1}\left(\|x_1 - \hat{x}_1\| + \alpha_1|\hat{a}_1 - a_1|\right) \\
&< \frac{\epsilon_3 + \alpha_1|\hat{a}_1 - a_1|}{a_1 - |\hat{a}_1 - a_1|} \\
&\leq 2\frac{\epsilon_3 + \frac{(10\sigma_1(A)\epsilon_3 + \alpha_1 a_1\epsilon)}{\alpha_1 a_1^2}}{a_1}
\end{aligned}$$

using $|\hat{a}_1 - a_1| < \frac{a_1}{2}$. We have,

$$\begin{aligned}
|\alpha_1 - \hat{\alpha}_1| &\leq 2\frac{\alpha_1 a_1^2\epsilon_3 + 10\sigma_1(A)\epsilon_3 + \alpha_1 a_1\epsilon}{\alpha_1 a_1^3} \\
&< \frac{2}{\alpha_1 a_1^3}\left(\left(\alpha_1 a_1^2 + 10\sigma_1(A)\right)\left(2\epsilon + 4R\epsilon_2/\sigma_{k-1}(Z_{\lambda_1})\right) + \alpha_1 a_1\epsilon\right) \\
&\leq \frac{4\sigma_1(A)}{\alpha_1 a_1^3}\left(\eta_1\epsilon + \frac{\eta_2 R\epsilon_2}{\sigma_{k-1}(Z_{\lambda_1})}\right) \qquad (23)
\end{aligned}$$

where $\eta_1 := \max\{\alpha_1 a_1(2a_1+1), 20\}$, and $\eta_2 := \max\{\alpha_1 a_1^2, 10\}$.

Finally using equation (18) we can bound $\|\widehat{Z}_{\lambda^*} - Z_{\lambda_1}\| \leq \epsilon_2 \leq 3\eta_3\epsilon$, where $\eta_3 = \max\left\{1, \frac{1}{w_1}, c_3\sigma_1(B)\right\}$. Using this in equations (22) and (23) proves the theorem.

E.2 Related Lemmas

In this section we prove a supporting lemma for Lemma 5.

Lemma 26 *Let $\{\mu_2, \dots, \mu_k\}$ be linearly independent. Suppose matrix Z_{λ^*} be expressed as,*

$$Z_{\lambda^*} = \sum_{i=2}^{k} \alpha_i(1 - \lambda^* w_i)\mu_i\mu_i^T = V_{1:(k-1)}\Sigma_{1:(k-1)}V_{1:(k-1)}^T = \sum_{i=2}^{k}\sigma_{i-1}(Z_{\lambda^*})v_i v_i^T, \qquad (24)$$

where $w_i = \langle \mu_i, v\rangle$, $V_{1:(k-1)} = [v_2, \dots, v_k]$ the matrix of $k-1$ singular vectors, and $\Sigma_{1:(k-1)}$ is a diagonal matrix of singular values of Z_{λ^*}. Then $\{v_2, \dots, v_k\}$ forms a basis of span$\{\mu_2, \dots, \mu_k\}$.

Proof Define $\mathcal{V}_{Z_{\lambda^*}}$ as the **column space** of matrix Z_{λ^*}. First observe that from equation (24) each column of Z_{λ^*} can be written as a linear combination of $\{\mu_2, \ldots, \mu_k\}$. Therefore any vector in the column space $\mathcal{V}_{Z_{\lambda^*}}$ can be written as a linear combination of $\{\mu_2, \ldots, \mu_k\}$. this implies,

$$\mathcal{V}_{Z_{\lambda^*}} \subseteq span\{\mu_2, \ldots, \mu_k\} \tag{25}$$

Now any vector $y \in \mathcal{V}_{Z_{\lambda^*}}$ can be written as $y = Z_{\lambda^*}x = \sum_{i=2}^{k} \sigma_{i-1}(Z_{\lambda^*})\langle v_i, x\rangle v_i$ using equation (24). This implies,

$$\mathcal{V}_{Z_{\lambda^*}} \subseteq span\{v_2, \ldots, v_k\} \tag{26}$$

Conversely any vector $s \in span\{v_2, \ldots, v_k\}$ can be written as $s = V_{1:(k-1)}r = Z_{\lambda^*}V_{1:(k-1)}\Sigma^{-1}_{1:(k-1)}r = Z_{\lambda^*}r'$, using equation (24), where $r' = V_{1:(k-1)}\Sigma^{-1}_{1:(k-1)}r$. This implies,

$$span\{v_2, \ldots, v_k\} \subseteq \mathcal{V}_{Z_{\lambda^*}} \tag{27}$$

Therefore combining equations (25),(26),(27) we get,

$$span\{v_2, \ldots, v_k\} = \mathcal{V}_{Z_{\lambda^*}} \subseteq span\{\mu_2, \ldots, \mu_k\} \tag{28}$$

Note that both the vector spaces $span\{v_2, \ldots, v_k\}$ and $span\{\mu_2, \ldots, \mu_k\}$ have rank $k-1$ since $\{v_2, \ldots, v_k\}$ are orthonormal, and $\{\mu_2, \ldots, \mu_k\}$ are linearly independent. Then from this rank constraint and equation (28) we must have:

$$span\{v_2, \ldots, v_k\} = span\{\mu_2, \ldots, \mu_k\}$$

This implies $\{v_2, \ldots, v_k\}$ forms a basis of $span\{\mu_2, \ldots, \mu_k\}$. ■

Appendix F. Subspace Clustering Proofs

In this section we prove Theorem 6 and the necessary lemmas. The main point is the following infinite-sample analysis, which shows that the top m eigenvectors of the whitened matrix B can be used to recover the subspace $\mathcal{U}_1$.

Theorem 27 *Suppose that there is some $\delta > 0$ such that $\|U_i v\|^2 \leq (1/3 - \delta)\|U_1 v\|^2$ for all $i \neq 1$. Let $Y = [u_1, ..., u_m]$ be the matrix of top m eigenvectors of $R = D^{-1/2}V^T BVD^{-1/2}$ and $Z = VD^{1/2}Y$. Let $\mathcal{Z}$ be the subspace spanned by columns of Z. Then,*

1. *$\mathcal{Z} = \mathcal{U}_1$*
2. *$\sigma_m(R) - \sigma_{m+1}(R) \geq 3\delta\|U_1 v\|^2$*

Proof Define $w_i = \|U_i U_i^T v\| = \|U_i^T v\|$, and $\tilde{U}_i := \sqrt{\alpha_i} D^{-1/2}V^T U_i$; note that $\sum_{i=1}^{k} \tilde{U}_i \tilde{U}_i^T$ is the $(km) \times (km)$ identity matrix, which implies that each $\tilde{U}_i$ has orthonormal columns. Consider the whitened B matrix. Using Theorem 13,

$$\begin{aligned} D^{-1/2}V^T BVD^{-1/2} &= \sum_{i=1}^{k} w_i^2 \tilde{U}_i \tilde{U}_i^T + 2\sum_{i=1}^{k} \tilde{U}_i U_i^T vv^T U_i \tilde{U}_i^T \\ &= \sum_{i=1}^{k} w_i^2 \tilde{U}_i \tilde{U}_i^T + 2\sum_{i=1}^{k} \tilde{v}_i \tilde{v}_i^T = \sum_{i=1}^{k} (w_i^2 \tilde{U}_i \tilde{U}_i^T + 2\tilde{v}_i \tilde{v}_i^T) \end{aligned}$$

where $\tilde{v}_i = \tilde{U}_i U_i^T v$. Note that $\tilde{v}_i$ are orthogonal to each other and each $\tilde{v}_i$ is in the space $\tilde{\mathcal{U}}_i$, the span of corresponding $\tilde{U}_i$. Moreover, $\|\tilde{v}_i\| = w_i$. Now for each i consider a different orthonormal basis $\tilde{V}_i$ of $\tilde{\mathcal{U}}_i$ such that in this basis the first unit vector is aligned along $\tilde{v}_i$. Define a rotation R_i such that $\tilde{V}_i = \tilde{U}_i R_i$. Then $\tilde{V}_i \tilde{V}_i^T = \tilde{U}_i \tilde{U}_i^T$. Therefore we can write the above equation as

$$R = D^{-1/2} V^T B V D^{-1/2} = \sum_{i=1}^{k} \tilde{V}_i \tilde{D}_i \tilde{V}_i^T \tag{29}$$

where each $\tilde{D}_i$ is a diagonal matrix with one maximum value of $3w_i^2$ and all other values w_i^2, and also the matrices $\tilde{V}_i$ are orthogonal. Under the assumption that $w_i^2 \leq (1/3 - \delta) w_1^2$, it follows that the top m eigenvectors of R are the columns of $\tilde{V}_i$, and that the corresponding eigenvalues are $3w_1^2$ and then w_1^2 repeated $m-1$ times. Therefore we can write $Y = \tilde{U}_i O$, where O is an $m \times m$ orthogonal matrix. Then,

$$Z = V D^{1/2} Y = V D^{1/2} \tilde{U}_i O = \sqrt{\alpha_1} U_1 O$$

This proves the first statement that $\mathcal{Z}$, the span of the columns of Z, is the subspace $\mathcal{U}_1$, the span of columns of U_1. The second statement follows from equation (29) since the maximum value of the $m+1$-th eigenvalue is $3w_i^2$ for some $i \neq 1$. Hence,

$$\sigma_m(R) - \sigma_{m+1}(R) \geq w_1^2 - 3 \max_{i \neq 1} w_i^2 \geq 3\delta w_1^2 = 3\delta \|U_1 v\|^2.$$

■

Lemma 28 *Let $\|\hat{A} - A\| < \epsilon < \sigma_{mk}(A)/4$. $A = VDV^T$ and $\hat{A} = \hat{V}\hat{D}\hat{V}^T$ be the eigen decompositions of $A, \hat{A}$. Let $\hat{W} = \hat{V}\hat{D}^{-1/2}$ be the whitening matrix. Then,*

$$\|I_k - (\hat{W}^T A \hat{W})^{-1/2}\| \leq \frac{4\epsilon}{\sigma_{mk}(A)}$$

Proof We prove this along the lines in Hsu and Kakade (2013). The matrix $\hat{W}$ whitens $\hat{A}$ since,

$$\hat{W}^T \hat{A} \hat{W} = \hat{D}^{-1/2} \hat{V}^T \hat{A} \hat{V} \hat{D}^{-1/2} = I_k$$

Also $\epsilon < \sigma_{mk}(A)/2$, hence using Weyl's inequality $\sigma_{mk}(\hat{A}) \geq \sigma_{mk}(A)/2$. This implies

$$\begin{aligned} \|I_k - \hat{W}^T A \hat{W}\| &= \|\hat{W}^T (\hat{A} - A) \hat{W}\| \leq \|\hat{W}\|^2 \|\hat{A} - A\| \\ &< \frac{2\epsilon}{\sigma_{mk}(A)} \end{aligned}$$

Therefore all eigenvalues of the matrix $\hat{W}^T A \hat{W}$ lie in the interval $(1 - 2\epsilon/\sigma_{mk}(A), 1 + 2\epsilon/\sigma_{mk}(A))$. This implies the eigenvalues of $(\hat{W}^T A \hat{W})^{-1}$ lie in the interval $(1/(1 + 2\epsilon/\sigma_{mk}(A)), 1/(1 - 2\epsilon/\sigma_{mk}(A)))$. Then,

$$\begin{aligned}
(I_k - (\hat{W}^T A \hat{W})^{-1/2})(I_k + (\hat{W}^T A \hat{W})^{-1/2}) &= I_k - (\hat{W}^T A \hat{W})^{-1} \\
I_k - (\hat{W}^T A \hat{W})^{-1/2} &= \left(I_k - (\hat{W}^T A \hat{W})^{-1}\right)(I_k + (\hat{W}^T A \hat{W})^{-1/2})^{-1} \\
\|I_k - (\hat{W}^T A \hat{W})^{-1/2}\| &\le \|I_k - (\hat{W}^T A \hat{W})^{-1}\| \\
&\le \frac{1}{1 - 2\epsilon/\sigma_{mk}(A)} - 1 \le \frac{4\epsilon}{\sigma_{mk}(A)}
\end{aligned}$$

■

Lemma 29 (Whitening matrix perturbation) *Assume* $\|\hat{A} - A\| < \epsilon < \sigma_{mk}(A)/4$. *Let* $\hat{W} = \hat{V}\hat{D}^{-1/2}$ *be the whitening matrix. Define* $W := \hat{W}(\hat{W}^T A \hat{W})^{-1/2}$ *. Then,*

$$\|\hat{W} - W\| \le \frac{8\epsilon}{\sigma_{mk}(A)^{3/2}}$$

Proof We note that the matrix W whitens the matrix A, since

$$W^T A W = (\hat{W}^T A \hat{W})^{-1/2} \hat{W}^T A \hat{W} (\hat{W}^T A \hat{W})^{-1/2} = I_k$$

We can bound the perturbation as follows.

$$\begin{aligned}
\|\hat{W} - W\| &= \|\hat{W}(I_k - (\hat{W}^T A \hat{W})^{-1/2})\| \\
&\le \|\hat{W}\| \|I_k - (\hat{W}^T A \hat{W})^{-1/2}\| \\
&\le \frac{2}{\sqrt{\sigma_{mk}(A)}} \frac{4\epsilon}{\sigma_{mk}(A)} = \frac{8\epsilon}{\sigma_{mk}(A)^{3/2}}
\end{aligned}$$

where the last inequality follows from Lemma 28. ■

Lemma 30 *Let* $\max\{\|\hat{A} - A\|, \|\hat{B} - B\|\} < \epsilon$, *and also let* $\epsilon < \min\{\sigma_1(B)/2, \frac{\sigma_{mk}(A)}{16}\}$. $W = \hat{W}(\hat{W}^T A \hat{W})^{-1/2}$ *be the whitening matrix. Define* $R = W^T B W$ *as the whitened* B *matrix, and* $\hat{R} = \hat{W}^T \hat{B} \hat{W}$ *is its estimate. Then,*

$$\|\hat{R} - R\| < \frac{51\sigma_1(B)\epsilon}{\sigma_{mk}(A)^2} := \epsilon_1$$

Proof From Lemma 29 we have $\|\hat{W} - W\| \le \frac{8\epsilon}{\sigma_{mk}(A)^{3/2}} < \|\hat{W}\|/2$. Also we know $\|\hat{W}\| \le \sqrt{2/\sigma_{mk}(A)}$. We obtain the required bound as follows.

$$\begin{aligned}
\|\hat{R} - R\| &= \|\hat{W}^T \hat{B} \hat{W} - W^T B W\| \\
&\leq \|(\hat{W} - W)^T \hat{B} \hat{W}\| + \|W^T (\hat{B} - B) \hat{W}\| + \|W^T B (\hat{W} - W)\| \\
&\leq \frac{3}{2} \|\hat{W} - W\| \|B\| \|\hat{W}\| + \frac{3}{2} \|\hat{W}\|^2 \|\hat{B} - B\| + \frac{3}{2} \|\hat{W}^T\| \|B\| \|\hat{W} - W\| \\
&= 3 \|\hat{W} - W\| \|B\| \|\hat{W}\| + \frac{3}{2} \|\hat{W}\|^2 \|\hat{B} - B\| \\
&< 48 \frac{\sigma_1(B)\epsilon}{\sigma_{mk}(A)^2} + \frac{3\epsilon}{\sigma_{mk}(A)} < \frac{51\sigma_1(B)\epsilon}{\sigma_{mk}(A)^2}
\end{aligned}$$

■

Lemma 31 *Suppose $Y = [u_1, \ldots, u_m]$ be the matrix of m largest eigenvectors of $R = W^T B W$, and $\hat{Y}$ be that of $\hat{R} = \hat{W}^T \hat{B} \hat{W}$. Let $\hat{Z} = \hat{V} \hat{D}^{1/2} \hat{Y}$. Then,*

$$\|\hat{Z}\hat{Z}^T - ZZ^T\| \leq C_1 \frac{\sigma_1(A)\sigma_1(B)\epsilon}{(\sigma_m(R) - \sigma_{m+1}(R))\sigma_{mk}(A)^2}$$

where Z satisfies $Y = W^T Z$, and C_1 is a constant.

Proof First using Wedin's theorem for the matrix A and $\hat{A}$ we get

$$\|\hat{V}\hat{V}^T - VV^T\| < \frac{4\epsilon}{\sigma_{mk}(A)}. \tag{30}$$

From Lemma 30 we have $\|\hat{R} - R\| < \frac{51\sigma_1(B)\epsilon}{\sigma_{mk}(A)^2} = \epsilon_1$. Therefore we can again use Wedin's theorem on the matrices $R, \hat{R}$ to bound the perturbation of the subspace spanned by Y.

$$\begin{aligned}
\|\hat{Y}\hat{Y}^T - YY^T\| &\leq \frac{4\|\hat{R} - R\|}{\sigma_m(R) - \sigma_{m+1}(R)} \\
&= \frac{4\epsilon_1}{\sigma_m(R) - \sigma_{m+1}(R)}.
\end{aligned} \tag{31}$$

We now bound the following term.

$$\begin{aligned}
\|\hat{V}\hat{D}^{1/2} W^T - \hat{V}\hat{V}^T\| &= \|\hat{V}\hat{D}^{1/2} (\hat{W}^T A \hat{W})^{-1/2} \hat{W}^T - \hat{V}\hat{V}^T\| \\
&= \|\hat{V}\hat{D}^{1/2} (\hat{W}^T A \hat{W})^{-1/2} \hat{D}^{-1/2} \hat{V}^T - \hat{V}\hat{V}^T\| \\
&\leq \|\hat{D}^{1/2} (\hat{W}^T A \hat{W})^{-1/2} \hat{D}^{-1/2} - I_k\| \\
&\leq \|\hat{D}^{1/2}\| \|(\hat{W}^T A \hat{W})^{-1/2} - I_k\| \|\hat{D}^{-1/2}\| \\
&\leq \sqrt{\frac{\sigma_1(\hat{A})}{\sigma_{mk}(\hat{A})}} \frac{4\epsilon}{\sigma_{mk}(A)} \leq \frac{8\sigma_1(A)^{1/2}\epsilon}{\sigma_{mk}(A)^{3/2}}
\end{aligned} \tag{32}$$

where the second to last inequality follows from Lemma 28. Next we show that $\hat{Z}\hat{Z}^T$ is close to the projection of ZZ^T onto the subspace $\hat{V}\hat{V}^T$.

$$
\begin{aligned}
& \|\hat{Z}\hat{Z}^T - \hat{V}\hat{V}^T ZZ^T \hat{V}\hat{V}^T\| \\
= \; & \|\hat{V}\hat{D}^{1/2}\hat{Y}\hat{Y}^T\hat{D}^{1/2}\hat{V}^T - \hat{V}\hat{V}^T ZZ^T \hat{V}\hat{V}^T\| \\
\leq \; & \|\hat{V}\hat{D}^{1/2}(\hat{Y}\hat{Y}^T - YY^T)\hat{D}^{1/2}\hat{V}^T\| + \|\hat{V}\hat{D}^{1/2}YY^T\hat{D}^{1/2}\hat{V}^T - \hat{V}\hat{V}^T ZZ^T \hat{V}\hat{V}^T\| \\
\leq \; & \sigma_1(\hat{A})\|\hat{Y}\hat{Y}^T - YY^T\| + \|\hat{V}\hat{D}^{1/2}W^T ZZ^T W\hat{D}^{1/2}\hat{V}^T - \hat{V}\hat{V}^T ZZ^T \hat{V}\hat{V}^T\| \qquad (33)
\end{aligned}
$$

We bound the second term as follows. Observe that the matrix $D^{-1/2}V^T$ also whitens the matrix A. Therefore Z can be expressed as $Z = VD^{1/2}U'$ where U' is a matrix with orthonormal columns. This implies $\|ZZ^T\| = \|VD^{1/2}U'U'^T D^{1/2}V^T\| \leq \sigma_1(A)$.

$$
\begin{aligned}
& \|\hat{V}\hat{D}^{1/2}W^T ZZ^T W\hat{D}^{1/2}\hat{V}^T - \hat{V}\hat{V}^T ZZ^T \hat{V}\hat{V}^T\| \\
\leq \; & \|(\hat{V}\hat{D}^{1/2}W^T - \hat{V}\hat{V}^T)ZZ^T W\hat{D}^{1/2}\hat{V}^T\| + \|\hat{V}\hat{V}^T ZZ^T(W\hat{D}^{1/2}\hat{V}^T - \hat{V}\hat{V}^T)\| \\
\leq \; & \|(\hat{V}\hat{D}^{1/2}W^T - \hat{V}\hat{V}^T)ZY^T\hat{D}^{1/2}\hat{V}^T\| + \|ZZ^T\|\|W\hat{D}^{1/2}\hat{V}^T - \hat{V}\hat{V}^T\| \\
\leq \; & \|\hat{V}\hat{D}^{1/2}W^T - \hat{V}\hat{V}^T\|\|Z\|\|\hat{D}^{1/2}\| + \|ZZ^T\|\|W\hat{D}^{1/2}\hat{V}^T - \hat{V}\hat{V}^T\| \\
\leq \; & \frac{8\sigma_1(A)^{1/2}\epsilon}{\sigma_{mk}(A)^{3/2}} \times 2\sigma_1(A) + \sigma_1(A) \times \frac{8\sigma_1(A)^{1/2}\epsilon}{\sigma_{mk}(A)^{3/2}} \\
= \; & 24\frac{\sigma_1(A)^{3/2}\epsilon}{\sigma_{mk}(A)^{3/2}}
\end{aligned}
$$

The second to last step follows from equation 32. Now using the above bound in equation 33 we get,

$$
\begin{aligned}
\|\hat{Z}\hat{Z}^T - \hat{V}\hat{V}^T ZZ^T \hat{V}\hat{V}^T\| &\leq \sigma_1(\hat{A})\|\hat{Y}\hat{Y}^T - YY^T\| + 24\frac{\sigma_1(A)^{3/2}\epsilon}{\sigma_{mk}(A)^{3/2}} \\
&\leq \frac{8\sigma_1(A)\epsilon_1}{\sigma_m(R) - \sigma_{m+1}(R)} + 24\frac{\sigma_1(A)^{3/2}\epsilon}{\sigma_{mk}(A)^{3/2}} \qquad (34)
\end{aligned}
$$

where the last step follows from inequalities (31). We compute the required bound by combining equations (30) and (34) as follows.

$$
\begin{aligned}
\|\hat{Z}\hat{Z}^T - ZZ^T\| &= \|\hat{Z}\hat{Z}^T - VV^T ZZ^T VV^T\| \\
&\leq \|\hat{Z}\hat{Z}^T - \hat{V}\hat{V}^T ZZ^T \hat{V}\hat{V}^T\| + 3\|VV^T - \hat{V}\hat{V}^T\|\|ZZ^T\| \\
&\leq \frac{8\sigma_1(A)\epsilon_1}{\sigma_m(R) - \sigma_{m+1}(R)} + 24\frac{\sigma_1(A)^{3/2}\epsilon}{\sigma_{mk}(A)^{3/2}} + \frac{12\sigma_1(A)\epsilon}{\sigma_{mk}(A)} \\
&\leq C_1\frac{\sigma_1(A)\sigma_1(B)\epsilon}{(\sigma_m(R) - \sigma_{m+1}(R))\sigma_{mk}(A)^2}
\end{aligned}
$$

where C_1 is a constant. ■

F.1 Proof of Theorem 6

The proof follows from Theorem 27 and Lemma 31. Note that the matrix Z has all singular values equal to $\sqrt{\alpha_1}$, therefore ZZ^T has singular values α_1. Under the affinity condition from Theorem 27, we have

$$\sigma_m(R) - \sigma_{m+1}(R) \geq 3\delta \|U_1 v\|^2$$

Combining with Lemma 31 we get

$$\|\hat{Z}\hat{Z}^T - ZZ^T\| \leq \frac{C_2 \sigma_1(A)\sigma_1(B)\epsilon}{\delta \|U_1 v\|^2 \sigma_{mk}(A)^2}$$

where C_2 is a constant. Finally applying Wedin's theorem for the matrices $\hat{Z}\hat{Z}^T$ and ZZ^T, we have

$$\|\hat{U}\hat{U}^T - U_1 U_1^T\| \leq \frac{C_3 \sigma_1(A)\sigma_1(B)\epsilon}{\alpha_1 \delta \|U_1 v\|^2 \sigma_{mk}(A)^2} \leq \frac{C \sigma_1(A)^2 \epsilon}{\alpha_1 \delta \sigma_{mk}(A)^2}$$

where $C_3 = 4C_2$.

Appendix G. Sample Complexity Analysis

Since the basic application of our method requires the estimation of certain covariance matrices, we need to show that one can estimate these matrices. There is a large literature on estimating covariance matrices, but for simplicity we will only focus on the simplest estimator: the sample covariance matrix. By well-known matrix concentration inequalities, one can show that the sample covariance matrix will be close to the covariance matrix with high probability if the sample size is large enough:

Theorem 32 *Tropp (2015) Let $A_1, \ldots, A_n$ be i.i.d. symmetric random $d \times d$ matrices. If $\|A_1\| \leq L$ a.s. then*

$$\Pr\left(\left\|\frac{1}{n}\sum_{i=1}^{n} A_i - \mathbb{E}A_i\right\| \geq t\right) \leq 8d \exp\left(-\frac{nt^2}{L^2}\right).$$

G.1 Truncation

Unfortunately, the matrices we will be dealing with do not usually have almost sure bounds on their norm. Here, we develop some straightforward truncation arguments in order to adapt Theorem 32.

Theorem 33 *Suppose that $A_1, \ldots, A_n$ are i.i.d. symmetric random $d \times d$ matrices satisfying the tail bound*

$$\Pr(\|A_1\| \geq t) \leq Ce^{-ct^\alpha}$$

for some $\alpha > 0$. Then for any $\epsilon, \delta > 0$, if $n \geq \tilde{\Omega}_\alpha(\epsilon^{-2}\log(d/\delta))$ then

$$\Pr(\|\hat{\mathbb{E}}A - \mathbb{E}A\| \geq \epsilon) \leq \delta,$$

where $\tilde{\Omega}_\alpha(k)$ means $C(\alpha)\Omega(k \log^{C(\alpha)} k)$.

Proof Fix $L > 0$ (to be determined later) and define the random matrix B_i by $B_i = A_i 1_{\{\|A_i\| \le L\}}$. Then Theorem 32 applies to B_i: if $n \ge \Omega(L^2 \epsilon^{-2} \log(d/\delta))$ then

$$\Pr(\|\hat{\mathbb{E}}B - \mathbb{E}B\| \ge \epsilon) \le \delta.$$

To compare this with the similar quantity involving A, we will consider $\hat{\mathbb{E}}(A - B)$ and $\mathbb{E}(A - B)$ separately.

First, note that $\Pr(A_i \ne B_i) = \Pr(\|A\| \ge L) \le C \exp(-cL^\alpha)$. If $L = \Omega(\log^{1/\alpha}(n/\delta))$ then $\Pr(A_i \ne B_i) \le \delta/n$. By a union bound,

$$\Pr(\hat{\mathbb{E}}A \ne \hat{\mathbb{E}}B) \le \delta. \tag{35}$$

Now we fix $L = C' \log^{1/\alpha}(n/(\delta \vee \epsilon))$ and we consider $\|\mathbb{E}(A - B)\|$. By the triangle inequality,

$$\|\mathbb{E}(A - B)\| = \|\mathbb{E}A 1_{\{\|A\| \ge L\}}\| \le \mathbb{E}\|A\| 1_{\{\|A\| \ge L\}}.$$

On the other hand, we can bound

$$\mathbb{E}\|A\| 1_{\{\|A\| \ge L\}} = \int_L^\infty \Pr(\|A\| \ge t)\, dt \le C \int_L^\infty e^{-ct^\alpha}\, dt.$$

With the change of variables $t = u^{1/\alpha}$, we have

$$\mathbb{E}\|A\| 1_{\{\|A\| \ge L\}} \le \frac{1}{\alpha} \int_{L^\alpha}^\infty u^{1/\alpha} e^{-cu}\, du.$$

Now, if $u \ge C'' \frac{1}{\alpha} \log \frac{1}{\alpha}$ for large enough C'' then $u^{1/\alpha} e^{-cu} \le e^{-cu/2}$. Hence, if $L^\alpha \ge C'' \frac{1}{\alpha} \log \frac{1}{\alpha}$ then

$$\mathbb{E}\|A\| 1_{\{\|A\| \ge L\}} \le \frac{1}{\alpha} \int_{L^\alpha}^\infty e^{-cu/2}\, du \le C(\alpha) e^{-cL^\alpha/2} \le C(\alpha)\epsilon$$

where the last inequality holds if the constant C' in the definition of L is large enough compared to c. On the other hand, if $L^\alpha < C'' \frac{1}{\alpha} \log \frac{1}{\alpha}$ then we must have $\epsilon > c(\alpha)$ for some $c(\alpha) > 0$. In this case, $\mathbb{E}\|A\| 1_{\{\|A\| \ge L\}} \le C \le C(\alpha)\epsilon$ trivially. To summarize, in every case we have

$$\|\mathbb{E}(A - B)\| \le C(\alpha)\epsilon.$$

Putting this together with (35), we have that if $n \ge \Omega(L^2 \epsilon^{-2} \log(d/\delta))$ then with probability at least $1 - 2\delta$,

$$\begin{aligned} \|\hat{\mathbb{E}}A - \mathbb{E}A\| &\le \|\hat{\mathbb{E}}B - \mathbb{E}B\| + \|\hat{\mathbb{E}}A - \hat{\mathbb{E}}B\| + \|\mathbb{E}A - \mathbb{E}B\| \\ &\le (1 + C(\alpha))\epsilon. \end{aligned}$$

Finally, recalling that $L = \mathrm{polylog}(n, 1/\epsilon, 1/\delta)$ (with the polynomial depending on α), we see that $n = \tilde{\Omega}_\alpha(\epsilon^{-2} \log(d/\delta))$ suffices. Finally, we can absorb the constant $C(\alpha)$ into ϵ. ■

We will now show how Theorem 33 bounds the error in estimating the various matrices that we had to estimate for the various different models we considered. Essentially, we will repeatedly use the observation that if z is a standard Gaussian variable then $z^{2/\alpha}$ has a tail that decays like e^{-ct^α}. In other words, moments of Gaussians will naturally lead to a condition that the one assumed in Theorem 33.

G.2 Gaussian Mixture Model

For the following theorem, we revert to the notation of the Gaussian mixture model.

Theorem 34 *Fix $\epsilon, \delta > 0$. Let $\hat{A} = \hat{\mathbb{E}}[xx^T]$ and $\hat{B} = \hat{\mathbb{E}}[\langle x, v\rangle xx^T]$, where $\hat{\mathbb{E}}$ is taken with n i.i.d. samples. If $n \geq \tilde{\Omega}(d\epsilon^{-2}\log(d/\delta))$ then with probability at least $1-\delta$, $\|\hat{\mathbb{E}}A - \mathbb{E}A\| \leq \epsilon$ and $\|\hat{\mathbb{E}}B - \mathbb{E}B\| \leq \epsilon$.*

Proof To estimate A, first note that $\|xx^T\| = \|x\|^2$. Now, $\mathbb{E}\|x\|^2 \leq R^2 + d\sigma^2$, where $R = \max_i \|\mu_i\|$, and also $\Pr(\|x\|^2 \geq \mathbb{E}\|x\|^2 + t\sqrt{d}) \leq Ce^{-ct}$. Hence, we may apply Theorem 33 with $A_i = x_i x_i^T/\sqrt{d}$ and $\alpha = 1$; this yields the claimed bound on $\|\hat{\mathbb{E}}A - \mathbb{E}A\|$.

To estimate B, note that $\|\langle x, v\rangle^2 xx^T\| = \langle x, v\rangle^2\|x\|^2$. Now, the triangle inequailty implies that $\langle x, v\rangle^2\|x\|^2$ is stochastically dominated by

$$4R^4 + 4\mathbb{E}[\langle z, v\rangle^2\|z\|^2] = 4R^4 + 4\mathbb{E}[z_1^2\|z\|^2],$$

where z is a standard (i.e., centered) Gaussian vector. Then $\mathbb{E}[z_1^2\|z\|^2] = 2 + d$, and $z_1^2\|z\|^2$ has tails of order $e^{-ct^{1/2}}$; that is it satisfies the assumptions of Theorem 33 with $\alpha = 1/2$. Applying Theorem 33 with $A_i = \langle x_i, v\rangle^2 x_i x_i^T/\sqrt{d}$ then yields the claimed bound on $\|\hat{\mathbb{E}}B - \mathbb{E}B\|$. ■

G.3 LDA Topic Model

For the following theorem, we revert to the notation of the LDA topic model, where d is the size of the dictionary.

Theorem 35 *Fix $\epsilon, \delta > 0$. Let $\hat{A} = \hat{\mathbb{E}}[x_1 x_2^T]$ and $\hat{B} = \hat{E}[\langle x_3, v\rangle x_1 x_2^T]$, where $\hat{\mathbb{E}}$ is taken with n i.i.d. samples. If $n \geq \Omega(\epsilon^{-2}\log(d/\delta))$ then with probability at least $1-\delta$, $\|\hat{A} - \mathbb{E}A\| \leq \epsilon$ and $\|\hat{B} - \mathbb{E}B\| \leq \epsilon$.*

Proof We can apply Theorem 32 directly, since $\|x_1 x_2^T\| \leq 1$ and $\langle x_3, v\rangle x_1 x_2^T \leq 1$. ■

G.4 Mixed Regression

For the following theorem, we revert to the notation of the mixed regression model.

Theorem 36 *Fix $\epsilon, \delta > 0$. Let $\hat{A} = \hat{\mathbb{E}}[y^2 xx^T]$ and $\hat{B} = \hat{\mathbb{E}}[y^3\langle x, v\rangle xx^T]$, where $\hat{\mathbb{E}}$ is taken with n i.i.d. samples. Let $R = \max_i \|\mu_i\|$. If $n \geq \tilde{\Omega}((R^2 + \sigma^2)\epsilon^{-2} d\log(d/\delta))$ then with probability at least $1-\delta$, $\|\hat{A} - \mathbb{E}A\| \leq \epsilon$ and $\|\hat{B} - \mathbb{E}B\| \leq \epsilon$.*

Proof Recalling that in cluster i we have $y = \langle x, \mu_i\rangle + \xi$, we have

$$\|y^2 xx^T\| \leq 2\langle x, \mu_i\rangle^2\|x\|^2 + 2\xi^2\|x\|^2.$$

Hence, $\mathbb{E}\|y^2 xx^T\| \leq 2R^2(2+d) + \sigma^2 d$, with tails that decay at the rate $e^{-ct^{1/2}}$. Applying Theorem 33 implies the claimed bounds for A. The case of B is analogous, except that since it involves sixth moments the tails will decay at the rate $e^{-ct^{1/3}}$; this only effects the poly-logarithmic terms hidden in the $\tilde{\Omega}$ notation. ■

G.5 Subspace Clustering

For the following theorem, we revert to the notation of the subspace clustering model. We assume for simplicity that σ is known, since if it isn't then it can be easily and accurately learnt.

Theorem 37 *Fix $\epsilon, \delta > 0$. Let $\hat{A} = \hat{\mathbb{E}}[xx^T] - \sigma^2 I_d$ and*

$$\hat{B} = \hat{\mathbb{E}}[\langle x, v\rangle^2 xx^T] - \sigma^2(v^T \hat{A} v) I_d - \sigma^2 \|v\|^2 \hat{A} - \sigma^4(\|v\|^2 I_d + vv^T) - 2\sigma^2(\hat{A}vv^T + vv^T \hat{A})$$

where $\hat{\mathbb{E}}$ is taken with respect to n i.i.d. samples. If $n \geq \tilde{\Omega}(\epsilon^{-2}(1+\sigma^2)\|v\|^2 m \log(d/\delta))$ then with probability at least $1-\delta$, $\|\hat{A} - A\| \leq \epsilon$ and $\|\hat{B} - B\| \leq \epsilon$.

Proof Since x/σ is an m-dimensional Gaussian vector, $\|x\|^2/(\sigma^2 m)$ is concentrated around its mean (1) with tails of order e^{-ct}. In other words, Theorem 33 (with $\alpha = 1$) implies our claim for A. The claim for B is analogous, except that since it involves fourth moments, the tails will decay at the rate $e^{-ct^{1/2}}$. ■

References

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.

Animashree Anandkumar, Rong Ge, Daniel Hsu, Sham M Kakade, and Matus Telgarsky. Tensor decompositions for learning latent variable models. *The Journal of Machine Learning Research*, 15(1):2773–2832, 2014.

Pablo Arbelaez, Michael Maire, Charless Fowlkes, and Jitendra Malik. Contour detection and hierarchical image segmentation. *IEEE Trans. Pattern Anal. Mach. Intell.*, 33(5): 898–916, May 2011.

Sanjeev Arora, Rong Ge, Yonatan Halpern, David M Mimno, Ankur Moitra, David Sontag, Yichen Wu, and Michael Zhu. A practical algorithm for topic modeling with provable guarantees. In *Proceedings of ICML-2013*, pages 280–288, 2013.

Sugato Basu, Arindam Banerjee, and Raymond Mooney. Semi-supervised clustering by seeding. In *Proceedings of 19th International Conference on Machine Learning (ICML-2002*, 2002.

David M Blei, Andrew Y Ng, and Michael I Jordan. Latent dirichlet allocation. *Journal of machine Learning research*, 3:993–1022, 2003.

Olivier Chapelle, Bernhard Schölkopf, and Alexander Zien, editors. *Semi-supervised learning.* MIT press Cambridge, 2006.

Yudong Chen, Xinyang Yi, and Constantine Caramanis. A convex formulation for mixed regression with two components: Minimax optimal rates. In *COLT*, pages 560–604, 2014.

Arthur P Dempster, Nan M Laird, and Donald B Rubin. Maximum likelihood from incomplete data via the em algorithm. *Journal of the royal statistical society. Series B (methodological)*, pages 1–38, 1977.

Ehsan Elhamifar and René Vidal. Sparse subspace clustering. In *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*, pages 2790–2797. IEEE, 2009.

Moritz Hardt and Eric Price. Tight bounds for learning a mixture of two gaussians. In *Proceedings of ACM on Symposium on Theory of Computing, STOC*, pages 753–760, 2015.

Daniel Hsu and Sham M Kakade. Learning mixtures of spherical gaussians: moment methods and spectral decompositions. In *Proceedings of the 4th conference on Innovations in Theoretical Computer Science*, pages 11–20. ACM, 2013.

Furong Huang, UN Niranjan, Mohammad Umar Hakeem, and Animashree Anandkumar. Online tensor methods for learning latent variable models. *Journal of Machine Learning Research*, 16:2797–2835, 2015.

Pirkko Kuusela and Daniel Ocone. Learning with side information: Pac learning bounds. *Journal of Computer and System Sciences*, 68(3):521–545, 2004.

Yue Lu and Chengxiang Zhai. Opinion integration through semi-supervised topic modeling. In *Proceedings of the 17th International Conference on World Wide Web*, pages 121–130. ACM, 2008.

Christopher D Manning, Prabhakar Raghavan, Hinrich Schütze, et al. *Introduction to information retrieval*, volume 1. Cambridge university press Cambridge, 2008.

Jon D Mcauliffe and David M Blei. Supervised topic models. In *Advances in neural information processing systems*, pages 121–128, 2008.

Ankur Moitra and Gregory Valiant. Settling the polynomial learnability of mixtures of gaussians. In *Foundations of Computer Science (FOCS), 2010 51st Annual IEEE Symposium on*, pages 93–102. IEEE, 2010.

David Newman, Jey Han Lau, Karl Grieser, and Timothy Baldwin. Automatic evaluation of topic coherence. In *Human Language Technologies: The 2010 Annual Conf. of the North American Chapter of the Association for Computational Linguistics*, pages 100–108. Association for Computational Linguistics, 2010.

David Newman, Edwin V Bonilla, and Wray Buntine. Improving topic coherence with regularized topic models. In *Advances in neural information processing systems*, pages 496–504, 2011.

Dohyung Park, Constantine Caramanis, and Sujay Sanghavi. Greedy subspace clustering. In *Advances in Neural Information Processing Systems*, pages 2753–2761, 2014.

Karl Pearson. Contributions to the mathematical theory of evolution. *Philosophical Transactions of the Royal Society of London. A*, pages 71–110, 1894.

Daniel Ramage, David Hall, Ramesh Nallapati, and Christopher D Manning. Labeled lda: A supervised topic model for credit attribution in multi-labeled corpora. In *Proc. of the 2009 Conf. on Empirical Methods in Natural Language Processing: Volume 1-Volume 1*, pages 248–256. Association for Computational Linguistics, 2009.

Richard A Redner and Homer F Walker. Mixture densities, maximum likelihood and the em algorithm. *SIAM review*, 26(2):195–239, 1984.

M. Röder, A. Both, and A. Hinneburg. Exploring the space of topic coherence measures. In *Proceedings of the eighth ACM international conference on Web search and data mining*, pages 399–408. ACM, 2015.

Michal Rosen-Zvi, Thomas Griffiths, Mark Steyvers, and Padhraic Smyth. In *Proceedings of the 20th conference on Uncertainty in Artificial Intelligence*, pages 487–494, 2004.

Hanie Sedghi, Majid Janzamin, and Anima Anandkumar. Provable tensor methods for learning mixtures of generalized linear models. In *Proceedings of International Conference on Artificial Intelligence and Statistics, AISTATS 2016*, pages 1223–1231, 2016.

Mahdi Soltanolkotabi and Emmanuel J Candes. A geometric analysis of subspace clustering with outliers. *The Annals of Statistics*, pages 2195–2238, 2012.

TACC. Texas advanced computing center, 2018. `http://www.tacc.utexas.edu`.

Joel Tropp. An introduction to matrix concentration inequalities. *arXiv preprint arXiv:1501.01571*, 2015.

UCI. NY Times dataset, 2008. `http://mlr.cs.umass.edu/ml/machine-learning-databases/`.

Eric P Xing, Michael I Jordan, Stuart Russell, and Andrew Y Ng. Distance metric learning with application to clustering with side-information. In *Advances in neural information processing systems*, pages 505–512, 2002.

Tianbao Yang, Rong Jin, and Anil K Jain. Learning from noisy side information by generalized maximum entropy model. In *Proceedings of the 27th International Conference on Machine Learning (ICML-10)*, pages 1199–1206, 2010.

Yelp. Yelp dataset, 2014. `http://www.yelp.com/dataset_challenge/`.

Xinyang Yi, Constantine Caramanis, and Sujay Sanghavi. Alternating minimization for mixed linear regression. In *Proceedings of International Conference on Machine Learning, ICML 2014*, pages 613–621, 2014.