THE UNIVERSITY OF CHICAGO

NEW METHODOLOGIES FOR HIGH-DIMENSIONAL DATA

A DISSERTATION SUBMITTED TO
THE FACULTY OF THE DIVISION OF THE PHYSICAL SCIENCES
IN CANDIDACY FOR THE DEGREE OF
DOCTOR OF PHILOSOPHY

DEPARTMENT OF STATISTICS

BY
HUY DANG TRAN

CHICAGO, ILLINOIS
AUGUST 2024

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGMENTS

I would like to express my gratitude to my advisor, Professor Claire Donnat, for her unwavering encouragement and support. Under her guidance, I have been able to develop myself as an independent researcher, and to explore interesting topics that have both theoretical and practical significance. I will always be fond of our conversations and discussions.

I would also like to thank my collaborators, Sansen Wei and Yating Liu, whose contributions have made this thesis possible. It has been a pleasure to work with both of you, and I hope you will achieve much success in your graduate studies and beyond.

ABSTRACT

High-dimensional statistics focuses on data whose ambient dimension is extremely large relative to the number of data points. In such regimes, classical asymptotic theory and traditional estimators often break down. This thesis proposes and analyzes new estimation procedures for two statistical settings where the number of parameters to be estimated is large but there exist special sparsity structures that can be exploited.

Part I of this thesis studies the estimation of the topic-word matrix under the probabilistic Latent Semantic Indexing model. Here, we assume that the ordered entries of the topic-word matrix's columns rapidly decay to zero; this assumption is partly motivated by the empirical observation that word frequencies in a text often follow Zipf's law. We introduce a new spectral estimation procedure that thresholds words based on their corpus frequencies, and show that its ℓ_1 error rate depends on the vocabulary size p only via a logarithmic term. The vocabulary size p is typically very large in practice, and prior works have not adequately addressed this high-dimensional data regime.

Part II of this thesis studies regression problems where the feature vectors are indexed by the vertices of a given graph. We propose a generalization of the Elastic Net penalty that is applicable when the true signal is believed to be smooth or piecewise constant with respect to the given graph. We derive the estimation and prediction error bounds under the assumptions of correlated Gaussian design and network alignment, and also provide a coordinate descent procedure based on the Lagrange dual objective to facilitate computations for large-scale problems.

CHAPTER 1

SPARSE TOPIC MODELING

1.1 Introduction

Topic modeling has proven to be a useful tool for dimensionality reduction and exploratory analysis in natural language processing. Beyond text analysis, it has also been successfully applied in areas such as population genetics [Pritchard et al., 2000, Bicego et al., 2012], social networks [Curiskis et al., 2020] and image analysis [Li et al., 2010].

1.1.1 The statistical model

In this project, we focus on the probabilistic Latent Semantic Indexing (pLSI) model introduced in Hofmann [1999]. This simple bag-of-words model involves three variables, namely topics (which are unobserved), words and documents.

Suppose we observe n documents written using a vocabulary of p words. For each $1 \leq i \leq n$, let N_i denote the length of document i. The corpus matrix $D \in \mathbb{R}^{p \times n}$, which is a sufficient statistic under the pLSI model and which records the empirical frequency of each word in each document, is defined by

$$D_{ji} = \frac{\text{count of word } j \text{ in document } i}{N_i} \quad \text{for all } 1 \leq i \leq n, 1 \leq j \leq p$$

Let $\{D_{*i} : 1 \leq i \leq n\}$ denote the columns of D, each of which contains only non-negative entries that sum up to 1. The pLSI model specifies that the raw word counts for each document $\{N_i D_{*i} : 1 \leq i \leq n\}$ are independently generated, with

$$N_i D_{*i} \sim \text{Multinomial}(N_i, [D_0]_{*i}) \tag{1.1}$$

for some matrix $D_0 \in \mathbb{R}^{p \times n}$ whose columns are $\{[D_0]_{*i} : 1 \leq i \leq n\}$. Here, the columns

of D_0 specify how words are assigned to documents, and these columns are required to be probability vectors with non-negative entries summing up to 1. Note that (1.1) implies $\mathbb{E}(D) = D_0$. If we let $Z := D - D_0$, we can write the observation model in a "signal plus noise" form:

$$D = D_0 + Z \tag{1.2}$$

The pLSI model further assumes that, for some unobserved $K \in \mathbb{N}$ (which denotes the number of topics), we can factorize D_0 as

$$\mathbb{E}(D) = D_0 = AW \tag{1.3}$$

for some matrices $A \in \mathbb{R}^{p\times K}$ and $W \in \mathbb{R}^{K\times n}$. Like D_0, the columns of A and W are required to be probability vectors, so that they can only contain non-negative entries that sum up to 1. A assigns words to topics, while W assigns topics to documents. In this project, we focus more specifically on estimating the topic-word matrix A.

One can think of (1.3) as equivalent to requiring that the following Bayes formula holds for any word j and document i:

$$\mathbb{P}(\text{word } j|\text{ document } i) = \sum_{k=1}^{K} \mathbb{P}(\text{word } j|\text{ topic } k) \cdot \mathbb{P}(\text{topic } k|\text{ document } i) \tag{1.4}$$

In most applications, $K \ll \min(n, p)$ and thus (1.3) impose a low-rank structure on $\mathbb{E}(D)$. We note that the number of topics K plays a role similar to that of the number of principal components in principal component analysis. For technical reasons, we will assume throughout that K is fixed as n, p and the document lengths N_i's vary. This is reasonable if one expects *a priori* that the number of topics covered by the corpus is small and bounded.

1.1.2 Related works and unaddressed issues

Before outlining our contributions in Section 1.1.3, it is important to provide context by discussing previous works that are relevant to the estimation of A under the pLSI model. In particular, we want to highlight some of the unaddressed issues from prior papers that our work aims to resolve.

1.1.2.1 The separability condition

We first present the definition of anchor words and the separability condition.

Definition 1 (Anchor words and separability)**.** *We call word j an anchor word for topic k if row j of A has exactly one nonzero entry at column k. The separability condition is said to be satisfied if there exists at least one anchor word for each topic $k \in \{1, \ldots, K\}$.*

Observe that the decomposition $D_0 = AW$ in general may not be unique, but under the separability condition, A is identifiable. The separability condition was first introduced in Donoho and Stodden [2003] to ensure uniqueness in the Non-negative Matrix Factorization (NMF) framework. The interpretation in our context is that, for each topic, there exist some words which act as unique signatures for that topic.

The separability condition greatly simplifies the problem of estimating A, as one can identify the anchor words for each topic as a first step. Prior works exploiting anchor words mainly differ in how anchor words are used to estimate the remaining non-anchor rows of A. Arora et al. [2012] start from the *word co-occurrence matrix* DD^T and apply a successive projection algorithm to rows of DD^T to find one anchor word per topic. The matrix DD^T is then re-arranged into four blocks where the top left $K \times K$ block corresponds to the anchor words identified, and A is estimated by taking advantage of the special structure of this block partition. More recently, Bing et al. [2020b] consider a matrix $B \in \mathbb{R}^{p \times K}$ obtained from A via multiplication by diagonal matrices. Unlike A, all rows of B sum up to 1, so anchor

rows of B are simply canonical basis vectors in $\mathbb{R}^K$. The non-anchor rows of B are then obtained via regression given the anchor rows of B. The topic matrix A can subsequently be recovered through an appropriate normalization of B.

A major drawback of these methods is that they rely heavily on the separability assumption, which suffices for uniqueness of the decomposition (1.3) but is far from necessary. This issue is related to the following question, which is of central importance in the NMF literature: given a collection of points $\{r_1, \ldots, r_m\} \subseteq \mathbb{R}^{K-1}$ presumed to lie within the convex hull of unobserved vertices $\{v_1^*, \ldots, v_K^*\}$, when is recovery of these vertices possible? In the NMF context, separability means that each vertex coincides with a point in the observed point cloud, in which case we only need to identify which of the r_i's correspond to simplex vertices. However, this is a very strong assumption and several efforts have been made to relax it. Javadi and Montanari [2020] show that vertex recovery is still possible under a uniqueness assumption that generalizes separability. Ge and Zou [2015] introduce the notion of *subset separability* which is also much weaker than separability. We note that many of the separability-based methods proposed in topic modeling, such as those in Arora et al. [2012], Bing et al. [2020a] and Bing et al. [2020b], have no obvious extension if the separability assumption is relaxed. This may not be important if the given corpus contains many specialized words and the topics are sufficiently distinct (an example is a collection of research papers), but may matter more if the topics overlap significantly and the vocabulary is generic (for instance, a collection of high school English essays).

1.1.2.2 The SVD-based approach in Ke and Wang [2022]

Ke and Wang [2022] are the first to establish the minimax-optimal rate of $\sqrt{\frac{p}{nN}}$ for the ℓ_1-loss $\|\hat{A} - A\|_1 := \sum_{j=1}^{p} \sum_{k=1}^{K} |\hat{A}_{jk} - A_{jk}|$ where, for simplicity, all document lengths are assumed to be equal to N. Their procedure links topic estimation to the NMF setting discussed in the previous subsection and is summarized as follows. Let $M := \mathrm{diag}(n^{-1} D \mathbf{1}_n)$

where $\mathbf{1}_n := (1, 1, \ldots, 1)^T \in \mathbb{R}^n$. Given K, the approach proposed in Ke and Wang [2022] considers the first K left singular vectors $\check{\xi}_1, \ldots, \check{\xi}_K \in \mathbb{R}^p$ of $M^{-1/2}D$. Elementwise division of $\check{\xi}_2, \ldots, \check{\xi}_K$ by $\check{\xi}_1$ (also known as *SCORE normalization* [Jin, 2015]) yields a matrix $\check{R} \in \mathbb{R}^{p\times(K-1)}$, whose rows $\check{r}_1, \ldots, \check{r}_p \in \mathbb{R}^{K-1}$ can be shown to form a point cloud contained in a K-vertex simplex (up to stochastic errors). Since this corresponds precisely to the NMF setup discussed in the previous subsection, the simplex vertices can now easily be recovered using a suitable *vertex hunting* algorithm. Once these vertices are identified, A can then be estimated via a series of normalizations.

The work by Ke and Wang [2022] is an important contribution that motivates several other methods for topic modeling, including ours. However, this method was developed using strong assumptions on the parameter regimes and the behavior of word frequencies. More specifically, Corollary 3.1 of Ke and Wang [2022] states that the error upper bound $\sqrt{\frac{p \log n}{nN}}$ is only applicable if we assume $N > p^{4/3}$ or $p \leq N < p^{4/3}$ and $n \geq \max(Np^2, p^3, N^2p^5)$. As the vocabulary size p is typically large, these are highly unrealistic assumptions on (n, N, p). For example, the Associated Press (AP) dataset used in Ke and Wang [2022] (a corpus of news articles frequently used for topic model evaluation) has $n = 2,134$ and $p = 7,000$. A typical AP article has between 300 and 700 words, so it is clear that none of the above assumptions holds. The error bound provided without these assumptions is $\frac{p^2}{N\sqrt{N}}\sqrt{\frac{p \log n}{nN}}$, which, when p is large and grows with n, may not necessarily converge to zero. Several other works that claim to establish minimax-optimal rates also do so by assuming $N > p$; see Theorem 4.1 of Wu et al. [2022] and Remark 10 of Bing et al. [2020a].

In this project, we do not seek to re-establish the rate $\sqrt{\frac{p}{nN}}$. Rather, we aim to provide a consistent error bound valid for all realistic parameter regimes (especially when $p > \max(n, N)$). We propose to resolve some of the outstanding issues of the estimator in Ke and Wang [2022] by leveraging a sparsity structure that is often empirically observed in text documents, resulting in:

1. **Improved error bounds:** We observe that even the minimax-optimal rate $\sqrt{\frac{p}{nN}}$ of Ke and Wang [2022] scales significantly with p. As the number of documents n increases, we can expect several previously unobserved words to be added to the corpus, whereas the average document length N may not change by much. However, many of these words may occur rarely, so the effective dimension of the parameter space may be quite small compared to the observed vocabulary size. This motivates us to restrict the parameter space by imposing a suitable column-wise sparsity assumption on A, which enables an error bound that does not scale with p except for log factors.

2. **An increased signal-to-noise ratio:** The approach in Ke and Wang [2022] may not be suitable if many words in the corpus occur with low frequency. If for each word j we define $h_j := \sum_{k=1}^{K} A_{jk}$, the theoretical guarantees in Ke and Wang [2022] require $\min_{1\leq j\leq p} h_j \geq \frac{cK}{p}$ for some $c \in (0, 1)$. Note that since the columns of A sum up to 1, we always have $\frac{1}{p}\sum_{j=1}^{p} h_j = \frac{K}{p}$. Therefore, since h_j roughly indicates the frequency of word j in the corpus, this assumption restricts the frequencies of the least frequent words to be of the same order as the average frequency of all words.

   Such a restrictive assumption is needed in Ke and Wang [2022] because when many low-frequency words exist in the corpus, their procedure involves division by small and noisy numbers. This is a problem with their *pre-SVD normalization* step where D is pre-multiplied by the $p \times p$ diagonal matrix $M^{-1/2}$, as the diagonal entries of $M := \mathrm{diag}(n^{-1}D\mathbf{1}_n)$ corresponding to infrequent words are usually small. This is also an issue with their elementwise division step, thus leading to higher errors from infrequent words in the point cloud obtained from their procedure (see Figures 1.8 and 1.17 for illustration). Although we also use SCORE normalization [Jin, 2015], our removal of infrequent words leads to a point cloud with a higher signal-to-noise ratio.

1.1.2.3 Sparse topic modeling approaches

To our knowledge, Bing et al. [2020b] and Wu et al. [2022] are the only two prior works that, like ours, impose additional sparsity constraints on A. However, the sparsity assumptions proposed in these papers are not appropriate for dealing with large p; rather, they are more suitable for dealing with large K.

1. Bing et al. [2020b] assume that A is elementwise sparse, in the sense that the total number of nonzero entries of A (denoted as $\|A\|_0$) is small. Their proposed procedure is then shown to satisfy the error upper bound

$$\|\hat{A} - A\|_1 \lesssim K\sqrt{\frac{\|A\|_0 \log(p \vee n)}{nN}} \tag{1.5}$$

   We note here that $\|A\|_0$ can still be very large. Indeed, let $\tilde{p}$ denote the number of words whose corresponding rows in A are *not* entirely zero. Technically we can have $p > \tilde{p}$, but words corresponding to zero rows of A are not observed with probability one, so $\tilde{p}$ covers the entire set of all distinct words observed in the corpus. We have

$$\tilde{p} \leq \|A\|_0 \leq Kp \tag{1.6}$$

   In fact, one can see that their error bound depends on p from the error decomposition $\|\hat{A} - A\|_1 \lesssim \mathrm{I} + \mathrm{II} + \mathrm{III}$ in Theorem 2 of Bing et al. [2020b]. For example, $\mathrm{I} = \frac{K}{\underline{\gamma}}\sqrt{\frac{p\log(n\vee p)}{nN}} + \frac{pK\log(p\vee n)}{\underline{\gamma}nN}$ for some constant $\underline{\gamma}$. This, together with (1.6), shows that the bound (1.5) is not very different from the rate $\sqrt{\frac{p}{nN}}$ in Ke and Wang [2022], except for possibly better dependence on K. Moreover, their theoretical results depend on several strong assumptions on the frequency of anchor words selected by their procedure. In contrast, our procedure is less affected by the frequency of anchor words, both in theory and in practice.

2. Wu et al. [2022] assume that each row of A has at most s_A nonzero entries. Since A has K columns, this sparsity assumption is only useful if K is large. Theorem 4.1 of Wu et al. [2022] then shows that their proposed estimator of A satisfies

$$\|\hat{A} - A\|_1 \lesssim K\sqrt{\frac{s_A \log n}{nN}} \tag{1.7}$$

However, upon close examination of their proof, the ℓ_1 bound they achieve is actually $K\sqrt{\frac{\|A\|_0 \log n}{nN}}$ (similar to (1.5)) so (1.7) is only possible by assuming that $p = O(1)$ and using $\|A\|_0 \leq ps_A$. Furthermore, their result assumes $N^{3/4} \geq p$ which, as we have noted in our discussion of Ke and Wang [2022], is highly restrictive.

In comparison with these two papers, our sparsity assumption is more compatible with the "large p" setting, and we do not assume $p = O(1)$ as in Wu et al. [2022].

1.1.3 Our contributions

We summarize the main contributions of this project below.

- We propose a new spectral procedure (Definition 5) for estimating A. This procedure takes into account the observation that, in most text datasets, the vocabulary size p is often large but many words occur very infrequently in the corpus. When K is unknown, a new estimator of K is also proposed (see Lemma 8).

- We introduce a new column-wise ℓ_q-sparsity assumption (Assumption 5) for A. This assumption is motivated by Zipf's law [Zipf, 1936] and links a word's frequency of occurrence in a topic to its rank. Our proposed procedure is then shown to be adaptive to the unknown sparsity level s in the ℓ_q-sparsity definition (1.19).

- We provide an error bound for our procedure using the ℓ_1 loss $\|\hat{A} - A\|_1$ in Theorem 7. Under our sparsity assumption (1.19), our error bound is shown to be valid for

all parameter regimes and only depends on p via weak factors. The common preprocessing step of removing infrequent words is incorporated into our procedure and accounted for in our analysis.

- Finally, in Section 1.2.4, we show that our theoretical results may still hold when the separability assumption is relaxed if we choose a suitable vertex hunting procedure for non-separable point clouds in Definition 5.

Extensive experiments with synthetic datasets to confirm the effectiveness of our estimation procedure under a wide variety of parameter regimes are presented in Section 1.3. Furthermore, we also demonstrate the usefulness of our method for text analysis, as well as for other applications where the pLSI model is also relevant, in Section 1.4.

1.1.4 Notations

For any set S, let $|S|$ denote its cardinality, and let S^c denote its complement if it is clear in context with respect to which superset. For any $k \in \mathbb{N}$, let $[k]$ denote the index set $\{1, \ldots, k\}$. We use $\mathbf{1}_d$ to denote the vector in $\mathbb{R}^d$ with all entries equal to 1. For a general vector $v \in \mathbb{R}^d$, let $\|v\|_r$ denote the vector ℓ_r norm, for $r = 0, 1, \ldots, \infty$, and let $\mathrm{diag}(v)$ denote the $d \times d$ diagonal matrix with diagonal entries equal to entries of v. For any $a, b \in \mathbb{R}$, let $a \vee b := \max(a, b)$ and $a \wedge b := \min(a, b)$.

Let I_m denote the $m \times m$ identity matrix. For a general matrix $Q \in \mathbb{R}^{m \times l}$ and $r = 0, 1, \ldots, \infty$, let $\|Q\|_r$ denote the vector ℓ_r-norm of Q if one treats Q as a vector. Let $\|Q\|_F$ and $\|Q\|_{\mathrm{op}}$ denote the Frobenius (i.e. $\|Q\|_F = \|Q\|_2$) and operator norms of Q respectively. For any index $j \in [m]$ and $i \in [l]$, let Q_{ji} or $Q(j, i)$ denote the (j, i)-entry of Q. For index sets $J \subseteq [m]$ and $I \subseteq [l]$, let Q_{JI} denote the submatrix of Q obtained by selecting only rows in J and columns in I (in particular, either J or I can be a single index). Also, let Q_{J*} denote the submatrix of Q obtained by selecting rows in J and *all* columns of Q; Q_{*I} is similarly defined. This means Q_{j*} and Q_{*i} denote the j^{th} row and i^{th} column of Q respectively. For

an integer $k \leq m \wedge l$, let $\gamma_k(Q)$ denote the k^{th} largest singular value of Q, and if Q is a square matrix then, if applicable, let $\lambda_k(Q)$ denote the k^{th} largest eigenvalue of Q. If $m = l$, then $\text{tr}(Q)$ denotes the trace of Q.

Let $C, c > 0$ denote absolute constants that may depend on K and q; we assume that K and q are fixed, unobserved constants. Let $C^*, c^* > 0$ denote numerical constants that do not depend on the unobserved quantities like K and q (this only matters when we discuss the estimation of K). The constants C, c, C^*, c^* may change from line to line.

In our paper, for ease of presentation, we assume $N_1 = \cdots = N_n = N$. Our results also hold if we assume the document lengths satisfy $\max_{i\in[n]} N_i \leq C^* \min_{i\in[n]} N_i$ (i.e. if $N_1 \asymp \cdots \asymp N_n$), in which case $N = \frac{1}{n}\sum_{i=1}^n N_i$ denotes the average document length.

1.2 Our procedure for estimating A and its theoretical properties

For simplicity, we will first assume separability in order to explain our procedure. A discussion of possible relaxations of this condition will be deferred to Section 1.2.4.

1.2.1 The oracle procedure to estimate A given D_0

Our oracle procedure concerns how A can be estimated if the non-stochastic matrix D_0, rather than D, is observed. Let $J \subseteq [p]$ be an arbitrary collection of words in our vocabulary. We first need the definition of a vertex hunting procedure, which is relevant to the NMF setup discussed in Section 1.1.2.1.

Definition 2 (Vertex hunting)**.** *Given K, a vertex hunting procedure is a function that takes a collection of points in $\mathbb{R}^{K-1}$ and returns K points in $\mathbb{R}^{K-1}$.*

Remark 1. A good vertex hunting procedure should return the vertices of the smallest K-simplex containing the given point cloud. We will use $\mathcal{V}(\cdot)$ to denote such a procedure.

The following definition of an ideal point cloud is based on the separability assumption. Any reasonable vertex hunting procedure should be able to successfully recover the simplex vertices from an ideal point cloud.

Definition 3 (Ideal point cloud)**.** *Given K, an ideal point cloud is a collection of points in $\mathbb{R}^{K-1}$ contained in the simplex defined by K vertices, such that the K vertices themselves belong to the point cloud.*

We are now ready to define the oracle procedure.

Definition 4 (Oracle procedure)**.** *Given inputs K, D_0, vertex hunting procedure $\mathcal{V}(\cdot)$ and a set of words $J \subseteq [p]$, the oracle procedure returns $\tilde{A} \in \mathbb{R}^{p\times K}$ defined as follows:*

1. *(SVD) Perform SVD on $[D_0]_{J*}$ to obtain $\Xi = [\xi_1, \ldots, \xi_K] \in \mathbb{R}^{|J|\times K}$ containing the first K left singular vectors of $[D_0]_{J*}$.*

2. *(Elementwise division) Divide $\xi_2, \ldots, \xi_K$ elementwise by ξ_1 to obtain $R \in \mathbb{R}^{|J|\times(K-1)}$. This means $R_{jk} = \xi_{k+1}(j)/\xi_1(j)$, for $k = 1, \ldots, K-1$ and $j \in J$.*

3. *(Vertex hunting) Treat the rows $\{r_j : j \in J\}$ of R as a point cloud in $\mathbb{R}^{K-1}$. Apply the vertex hunting procedure $\mathcal{V}(\cdot)$ on this point cloud to obtain vertices $v_1^*, \ldots, v_K^*$.*

4. *(Recovery of Π) For each $j \in J$, solve for $\pi_j \in \mathbb{R}^K$ from the linear equation*

$$\begin{pmatrix} 1 & \ldots & 1 \\ v_1^* & \ldots & v_K^* \end{pmatrix} \pi_j = \begin{pmatrix} 1 \\ r_j \end{pmatrix} \tag{1.8}$$

   In other words, π_j satisfies $\sum_{k=1}^K \pi_j(k) = 1$ and $r_j = \sum_{k=1}^K \pi_j(k) v_k^*$, for each $j \in J$. Let $\Pi \in \mathbb{R}^{|J|\times K}$ be the matrix whose rows are $\{\pi_j : j \in J\}$.

5. *(Normalization) Normalize the columns of $\mathrm{diag}(\xi_1) \cdot \Pi \in \mathbb{R}^{|J|\times K}$ so that the entries of each column sum up to 1. This yields $\tilde{A}_{J*}$. Set $\tilde{A}_{J^c*} = 0$ to obtain $\tilde{A}$.*

Our oracle procedure makes use of the SCORE normalization idea which was originally proposed for network data analysis [Jin, 2015]. The elementwise division step (Step 2) is the most important step, as it provides a connection between singular vectors of D_0 (or associated variables) and the NMF setup described in Section 1.1.2.1. The words in J are represented by the point cloud $\{r_j : j \in J\}$, which can be shown to be contained entirely in some K-vertex simplex. If the simplex vertices are identifiable and the vertex hunting procedure is successful in recovering them in Step 3, then (1.8) allows us to exactly recover the probabilistic weights $\{\pi_j : j \in J\}$ associated with each word in J, which are connected to A via the relation

$$\text{diag}(\xi_1) \cdot \Pi = A_{J*} \cdot \text{diag}(V_1) \tag{1.9}$$

for some vector $V_1 \in \mathbb{R}^K$ containing only positive entries. This explains the column normalization step (Step 5), which essentially reverses the elementwise division step. For more details, we refer the reader to the proof of Lemma 13.

Based on the relation (1.9), we can show the following result.

Lemma 1. *Suppose the set J contains at least one anchor word for each topic $k \in [K]$, and the vertex hunting procedure $\mathcal{V}(\cdot)$ can successfully recover the simplex vertices from any ideal point cloud. The oracle procedure in Definition 4 then returns $\tilde{A}$ satisfying $\tilde{A}_{J^c*} = 0$ and*

$$\tilde{A}_{J*} = A_{J*} \cdot \mathit{diag}(\|A_{J1}\|_1^{-1}, \ldots, \|A_{JK}\|_1^{-1}) \tag{1.10}$$

The proof of Lemma 1 is identical to that of Lemma 13. The sole difference is that in Lemma 13, the set J is chosen as in (1.11) and we use Assumption 3.

Remark 2. Our oracle procedure differs from that of Ke and Wang [2022] in two important ways. First, note that Step 1 only requires SVD to be performed on a submatrix of D_0. In general, we want the set J to contain words that occur with sufficiently high frequencies in the corpus so that the point cloud generated from our procedure has a higher signal-to-

noise ratio. When p is large, we can often expect the corpus to contain many infrequently occurring words whose corresponding rows in A should be estimated as zero. Our oracle procedure yields $\tilde{A}$ which is a good oracle approximation of A if $\|A_{J^c *}\|_1$ is small, as in that case the diagonal matrix in (1.10) is close to the identity matrix.

Second, note that we consider the SVD of a submatrix of D_0 and not $M_0^{-1/2} D_0$ as in Ke and Wang [2022], where $M_0 := \text{diag}(n^{-1} D_0 \mathbf{1}_n)$. This simplifies some parts of our theoretical analysis and allows us to obtain error bounds that depend less strongly on p.

1.2.2 Estimation procedure for A given D

Our procedure to estimate A below is designed to closely approximate the oracle procedure. Here we first assume K is known. The estimation of K is deferred to Section 1.2.5, and the choice of the vertex hunting procedure will be discussed in conjunction with identifiability assumptions on A.

Definition 5 (Estimation procedure for A)**.** *Given inputs K, observation matrix D and vertex hunting procedure $\mathcal{V}(\cdot)$, our estimation procedure returns $\hat{A}$ defined as follows:*

1. *(Thresholding) Let $M := diag(n^{-1} D \mathbf{1}_n)$ and $p_n := p \vee n$. Compute the set of words*

$$J := \left\{ j \in [p] : M(j,j) \geq \alpha \sqrt{\frac{\log p_n}{nN}} \right\} \tag{1.11}$$

   Here, α is a user-specified universal constant (see Remark 3).

2. *(Spectral decomposition) Compute the first K eigenvectors $\hat{\xi}_1, \ldots, \hat{\xi}_K \in \mathbb{R}^{|J|}$ of the submatrix G_{JJ} of the $p \times p$ matrix G, where*

$$G := DD^T - \frac{n}{N} M \tag{1.12}$$

   Here, we assume all entries of $\hat{\xi}_1$ are of the same sign, in which case we can choose $\hat{\xi}_1$

to have all positive entries. If some entries of $\hat{\xi}_1$ are negative, choose $\hat{\xi}_1$ such that the majority of entries are positive, and apply Remark 4.

3. *(Elementwise division) Divide $\hat{\xi}_2, \ldots, \hat{\xi}_K$ elementwise by $\hat{\xi}_1$ to obtain $\hat{R} \in \mathbb{R}^{|J| \times (K-1)}$, with rows $\{\hat{r}_j : j \in J\}$. This means $\hat{R}_{jk} = \hat{\xi}_{k+1}(j)/\hat{\xi}_1(j)$, for $k \in [K-1]$ and $j \in J$.*

4. *(Vertex hunting) Treat the rows of $\hat{R}$ as a point cloud in $\mathbb{R}^{K-1}$. Apply the vertex hunting procedure $\mathcal{V}(\cdot)$ to this point cloud to obtain vertices $\hat{v}_1^*, \ldots, \hat{v}_K^*$.*

5. *(Estimation of Π) For each $j \in J$, solve for $\hat{\pi}_j^\diamond \in \mathbb{R}^K$ from*

$$\begin{pmatrix} 1 & \ldots & 1 \\ \hat{v}_1^* & \ldots & \hat{v}_K^* \end{pmatrix} \hat{\pi}_j^\diamond = \begin{pmatrix} 1 \\ \hat{r}_j \end{pmatrix} \tag{1.13}$$

   Obtain $\hat{\pi}_j$ from $\hat{\pi}_j^\diamond$ by first setting any negative entries to 0 and then normalizing so that the entries of $\hat{\pi}_j$ sum up to 1. Let $\hat{\Pi}$ be the matrix whose rows are $\{\hat{\pi}_j : j \in J\}$.

6. *(Normalization) Normalize all columns of $diag(\hat{\xi}_1) \cdot \hat{\Pi}$ so that they have unit ℓ_1-norm. This yields $\hat{A}_{J^*}$. Set all entries of $\hat{A}_{J^c *}$ to zero to obtain $\hat{A}$.*

As Steps 3-5 are also based on the SCORE normalization idea [Jin, 2015], we call this procedure the *Thresholded Topic-SCORE* (TTS). However, Step 1, Step 2 and Step 6 contain significant differences when compared with Topic-SCORE in Ke and Wang [2022].

Remark 3 (Choice of α)**.** The set J in (1.11) is chosen by examining the row sums of the observation matrix D, which indicate how frequently the words occur in the corpus. In (1.11), α is meant to be a universal constant and thus does not affect our error rates, which are not optimized over constants. In our theoretical discussion, we choose $\alpha = 8$ for convenience, but for most datasets this value of α may result in too many words not meeting the threshold.

In practice, a good choice of α is important for obtaining a good estimator of A. Based on our experiments, we recommend a smaller value of α, such as $\alpha = 0.005$. This choice of α should produce reasonable results for commonly observed values of (n, N, p). Based on what we observe from experiments, if $n \in [1000, 5000], N \in [300, 700], p \in [5000, 10000]$, we can typically expect around 10-40% of words to be removed.

Remark 4 (Signs of $\hat{\xi}_1$'s entries)**.** In the oracle procedure, ξ_1 is the first left singular vector of $[D_0]_{J*}$ and so by Perron's theorem, the entries of ξ_1 are all positive. In Step 2, $\hat{\xi}_1$ is the first eigenvector of G_{JJ} which is not necessary a Perron matrix, so $\hat{\xi}_1$ technically may contain negative entries. Any word j for which $\hat{\xi}_1(j)$ is negative should have corresponding rows of A set to zero after Step 2, and then in Step 3 we form the point cloud by computing $\hat{\xi}_{k+1}(j)/\hat{\xi}_1(j)$ for $k \in [K-1]$ and $j \in J$ with $\hat{\xi}_1(j) > 0$ only.

In our theoretical analysis as well as in practice, however, this scenario will not happen with high probability. This is because G is chosen so that $\max_{j \in J} |\hat{\xi}_1(j) - \xi_1(j)|$ is small. Since any word j that meets our threshold occurs with sufficiently high frequency, $\xi_1(j)$ will also be sufficiently large for any $j \in J$, which implies $|\hat{\xi}_1(j) - \xi_1(j)| \ll \xi_1(j)$ and thus $\hat{\xi}_1(j) \geq \xi_1(j)/2$ for all $j \in J$.

The set J as defined in (1.11) is data-dependent. It is quite useful to note that J can be approximated by the non-stochastic sets (1.14) with high probability. The proof of the lemma below can be found under Theorem 11(b).

Lemma 2. *Let $M_0 := diag(n^{-1} D_0 \mathbf{1}_n)$, and let*

$$J_{\pm} := \left\{ j \in [p] : M_0(j,j) > \alpha_{\pm} \alpha \sqrt{\frac{\log p_n}{nN}} \right\} \tag{1.14}$$

where α is from the definition of J in (1.11) and $\alpha_- > 1$ and $\alpha_+ \in (0,1)$ are some suitably chosen constants depending on α (for example if $\alpha = 8$, we can let $\alpha_+ = \frac{1}{2}$, $\alpha_- = 2$). Then the event $\mathcal{E} := \{J_- \subseteq J \subseteq J_+\}$ occurs with probability at least $1 - o(p_n^{-1})$.

The following lemma bounds the size of J, and is obtained by bounding $|J_+|$ and using Lemma 2.

Lemma 3 (Size of J)**.** *With probability at least* $1 - o(p_n^{-1})$*,*

$$|J| \leq \left(\frac{K}{\alpha \alpha_+} \sqrt{\frac{Nn}{\log p_n}} \right) \wedge p \tag{1.15}$$

Our procedure requires the eigenvalue decomposition of a symmetric $|J| \times |J|$ matrix. The bound (1.15) can be significantly smaller than $\min(n, p)$ if $nN \ll p^2$ and $N \ll n$ (ignoring weak factors), which are reasonable assumptions for many text datasets. We can therefore expect the eigenvalue decomposition step in our procedure to be more computationally scalable than the SVD step (on a $p \times n$ matrix) in Ke and Wang [2022].

1.2.3 Error bounds for $\hat{A}$ under separability

We first discuss our theoretical results under separability, which is assumed in all of our proofs in Section 1.5. We begin by listing the assumptions underlying our analysis.

Assumption 1 (A and W are well-conditioned)**.** *Let* $\Sigma_W := n^{-1} W W^T$*. For some constant* $c \in (0, 1)$*,*

$$\sigma_K(A) \geq c\sqrt{K} \quad \textit{and} \quad \sigma_K(\Sigma_W) \geq c \tag{1.16}$$

Assumption 2 (The topic-topic correlation matrix is regular)**.** *The entries of* $A^T A$ *satisfy the following for some constant* $c > 0$*:*

$$\min_{1 \leq k,l \leq K} A^T A(k, l) \geq c \tag{1.17}$$

Assumption 3 (Separability)**.** *Each topic* $k \in [K]$ *has at least one associated anchor word* j *belonging to the set* J_- *defined in* (1.14)*.*

Assumption 4 (Vertex hunting efficiency)**.** *Given K and an ideal point cloud defined in Definition 3, the vertex hunting function $\mathcal{V}(\cdot)$ recovers the K vertices correctly. Furthermore, whenever $\mathcal{V}(\cdot)$ is given as inputs two point clouds $\{x_1, \ldots, x_m\}$ and $\{x'_1, \ldots, x'_m\}$, the outputs $\{v_1, \ldots, v_K\}$ and $\{v'_1, \ldots, v'_K\}$ satisfy for some absolute constant $C > 0$ (up to a label permutation)*

$$\max_{k\in[K]} \|v_k - v'_k\|_2 \leq C \max_{j\in[m]} \|x_j - x'_j\|_2 \tag{1.18}$$

Assumption 5 (Column-wise ℓ_q-sparsity)**.** *Let the entries of each column A_{*k} of A be ordered as $A_{(1)k} \geq \cdots \geq A_{(p)k}$. For some $q \in (0, 1)$ and $s > 0$, the columns of A satisfy*

$$\max_{k\in[K]} \left(\max_{j\in[p]} j A^q_{(j)k} \right) \leq s \tag{1.19}$$

Here, we assume that q is a fixed constant, whereas s is allowed to grow with n.

Remark 5. We justify why Assumptions 1, 2 and 3 are reasonable below.

1. Equation (1.16) assumes the topic vectors in A are not too correlated. The assumption on W in (1.16) is necessary even when W is known, as its role is similar to that of the design matrix in the regression setting. Note that since the columns of A and W sum up to 1, we always have $\sigma_1(A) \leq \sqrt{K}$ and $\sigma_1(\Sigma_W) \leq 1$ (see Lemma 12(a)).

2. The matrix $A^T A \in \mathbb{R}^{K\times K}$ can be thought of as the topic-topic correlation matrix, since its entries are inner products of the columns of A. Therefore, (1.17) is especially true if the K topics are related to one another. However, even if the corpus covers unrelated topics, we expect all columns of A to assign significant weights to grammatical function words (such as 'and', 'the' in English) and filler words, which occur frequently in all documents regardless of the topics involved.

3. In light of Lemma 2, Assumption 3 requires that each topic has at least one anchor word that occurs in the corpus frequently enough so that it is included in J. Such an

assumption on the frequency of anchor words is also commonly seen in other works that exploit the separability condition, and Assumption 3 is not strong since the threshold level of order $\sqrt{\frac{\log p_n}{nN}}$ in the definition of J_- is quite low. For comparison, Bing et al. [2020b] makes the same assumption but with the threshold level of order $\frac{\log p_n}{N}$, which may be higher than ours if the number of documents n far exceeds the average document length N.

Remark 6 (Vertex hunting for separable point clouds)**.** Ke and Wang [2022] mentions two vertex hunting algorithms which are suitable for separable point clouds, namely Successive Projection (SP) [Araújo et al., 2001] and Sketched Vertex Search (SVS) [Jin et al., 2017].

Given a point cloud $r_1, \ldots, r_m$, SP starts by finding the point r_j whose Euclidean norm is the largest and sets this as the first estimated vertex $\hat{v}_1$. Then, for each $2 \leq k \leq K$, we can obtain $\hat{v}_k$ from $\hat{v}_1, \ldots, \hat{v}_{k-1}$ by setting $\hat{v}_k$ as the point r_j that maximizes $\|(I - P_{k-1})r_j\|_2$, where P_{k-1} denotes the projection matrix on the linear span of $\hat{v}_1, \ldots, \hat{v}_{k-1}$. SP can be shown to satisfy Assumption 4 when the volume of the true simplex is lower bounded by a constant [Gillis and Vavasis, 2013], which is a simple consequence of Theorem 12(f).

On the other hand, SVS starts by applying k-means clustering on the point cloud $\{r_1, \ldots, r_m\}$ to obtain cluster centers $\hat{c}_1, \ldots, \hat{c}_L$, where L is a tuning parameter that is much larger than K. These clusters are meant to reduce the noise levels in the point cloud. Next, SVS exhaustively searches for all simplexes whose K vertices are located on these cluster centers, in order to find the simplex S such that the maximum distance from any $\hat{c}_l$ to S is minimized. In comparison to SP, SVS is more robust to noise in the point cloud but is computationally much slower if K is not small. SVS satisfies Assumption 4 under mild regularity conditions [Jin et al., 2017].

Note that these vertex hunting algorithms are only meant for separable point clouds, as the simplex vertices they produce are designed to belong to the convex hull of the point cloud. For more implementation details of SVS and SP, we refer the reader to Section A of

Ke and Wang [2022].

Remark 7 (ℓ_q-sparsity)**.** To our knowledge, our work is the first to consider the ℓ_q-sparsity assumption (1.19) in the topic modeling context, although similar assumptions have been adopted in other statistical settings such as sparse PCA and sparse covariance estimation (see for example Ma [2013] and Cai and Zhou [2012]). (1.19) imposes an assumption on the decay rate of the ordered entries of the columns of A, but does not restrict how small (or large) the smallest (or largest, assuming $s \geq 1$) entries of A's columns can be. Thus, our theoretical results are valid even in the presence of severe word frequency heterogeneity.

Note that if columns A_{*k} has s nonzero entries, then we always have $\max_{j\in[p]} jA^q_{(j)k} \leq s$. However, in light of (1.6) where we observe that $\|A\|_0 \geq \tilde{p}$, there exists at least one column of A with at least $\lfloor \tilde{p}/K \rfloor$ nonzero entries, and so s in (1.19) cannot be much smaller than p if we impose hard sparsity ($q = 0$) on *all* columns of A. Therefore, the ℓ_q-sparsity assumption (1.19) gives us more flexibility as it allows for the possibility that most entries of A are small but nonzero. When $q \approx 0$, we can approximate the assumption of hard sparsity on all columns of A, whereas when q is close to 1, then (1.19) with $s = O(1)$ corresponds to Zipf's law, which is the empirical observation that word frequency in text data is often inversely proportional to word rank.

The restriction that $q \in (0, 1)$ is primarily due to the fact that we use the ℓ_1 loss $\|\hat{A}-A\|_1$. Since the columns of A sum up to 1, the columns of A already satisfy ℓ_q-sparsity with $q = s = 1$, but this alone is not sufficient to control the error term $\|A_{J^c}\|_1$ resulting from our thresholding step.

We are now ready to discuss our main theoretical results. Let $\hat{\Xi} = [\hat{\xi}_1, \ldots, \hat{\xi}_K] \in \mathbb{R}^{|J|\times K}$ contains the first K eigenvectors of G_{JJ} where G is defined as in (1.12). Recall its oracle counterpart $\Xi = [\xi_1, \ldots, \xi_K] \in \mathbb{R}^{|J|\times K}$ which contains the first K left singular vectors of $[D_0]_{J*}$. Let $\{\Xi_j : j \in J\}$ and $\{\hat{\Xi}_j : j \in J\}$ denote the rows of Ξ and $\hat{\Xi}$ respectively.

Lemma 4 (Row-wise error bounds for $\hat{\Xi}$)**.** *For all $j \in [p]$, let $h_j := \sum_{k=1}^K A_{jk}$. With*

probability $1 - o(p_n^{-1})$*, there exist* $\omega \in \{\pm 1\}$ *and a* $(K-1) \times (K-1)$ *orthonormal matrix* Ω^* *such that, if we define* $\Omega := diag(\omega, \Omega^*) \in \mathbb{R}^{K \times K}$*, we have*

$$\|\Omega\hat{\Xi}_j - \Xi_j\|_2 \leq C\sqrt{\frac{h_j \log p_n}{nN}} \quad \textit{for all } j \in J \tag{1.20}$$

The proof can be found in Lemma 25 and is an application of the well-known Davis-Kahan theorem (more specifically, we need to use the row-wise perturbation version of the theorem as proven in Lemma F.1 of Ke and Wang [2022]). We note here that the bound (1.20) depends on p only via the log term, and the h_j's, which indicates how frequently one may encounter word j in the corpus, determines the magnitude of the bound (1.20).

As a consequence of the above lemma, one can provide error bounds for the point cloud obtained from our procedure. Again, recall that $\{r_j : j \in J\}$ is the oracle point cloud from Step 3 of Definition 4, and $\{\hat{r}_j : j \in J\}$ is the point cloud from Step 4 of Definition 5.

Corollary 5 (Error bounds for the point cloud)**.** *With probability* $1 - o(p_n^{-1})$*, there exists a* $(K-1) \times (K-1)$ *orthonormal matrix* Ω^* *such that*

$$\max_{j \in J} \|\Omega^* \hat{r}_j - r_j\|_2 \leq C \left(\frac{\log p_n}{nN}\right)^{1/4} \tag{1.21}$$

The proof can be found in Lemma 26. To elaborate further on (1.21), we can show that with high probability,

$$\|\Omega^* \hat{r}_j - r_j\|_2 \leq C\sqrt{\frac{\log p_n}{h_j nN}} \quad \text{for all } j \in J \tag{1.22}$$

Observe that unlike (1.20), the bound (1.22) is inversely proportional to $\sqrt{h_j}$ due to the fact that the point cloud is obtained from the elementwise division step. Since we do not restrict how small $\min_{1 \leq j \leq p} h_j$ can be, the error bound (1.22) may be uncontrollable without appropriate thresholding of infrequent words. However, with the choice of J as in (1.11),

one can show $\min_{j\in J} h_j \geq c\sqrt{\frac{\log p_n}{nN}}$ with high probability, which when combined with (1.22) leads to (1.21).

From (1.22), we can also obtain bounds on how much the probabilistic weights $\{\hat{\pi}_j : j \in J\}$ from Step 5 of Definition 5 deviate from the oracle weights $\{\pi_j : j \in J\}$ from Step 4 of Definition 4). The proof of the following corollary can be found under Lemma 27.

Corollary 6 (Error bounds for $\hat{\Pi}$)**.** *With probability* $1 - o(p_n^{-1})$*,*

$$\max_{j\in J} \|\hat{\pi}_j - \pi_j\|_1 \leq C\left(\frac{\log p_n}{nN}\right)^{1/4} \tag{1.23}$$

Note that while $\{\Xi_j : j \in J\}$ and $\{r_j : j \in J\}$ can be recovered only up to an orthonormal transformation Ω^*, the bound (1.23) does not depend on Ω^*. We also note that the bounds (1.20), (1.21) and (1.23) are derived without using the ℓ_q-sparsity assumption (Assumption 5).

The next theorem is our main result, which provides the error rate for estimating A using the ℓ_1 loss $\|\hat{A} - A\|_1$. Recall the definition of $\tilde{A}$ in Lemma 1.

Theorem 7 (Estimation error for $\hat{A}$)**.** *Suppose Assumptions 1-4 are satisfied. Then with probability* $1 - o(p_n^{-1})$*,*

$$\|\hat{A}_{J*} - \tilde{A}_{J*}\|_1 \leq C\left(\frac{\log p_n}{nN}\right)^{1/4} \tag{1.24}$$

If we further assume the ℓ_q*-sparsity assumption (Assumption 5) and* $s\left(\frac{\log p_n}{nN}\right)^{\frac{1-q}{2}} = o(1)$*, we also have with probability* $1 - o(p_n^{-1})$*,*

$$\|\tilde{A}_{J*} - A_{J*}\|_1 = \|A_{J^c*}\|_1 \leq Cs\left(\frac{\log p_n}{nN}\right)^{\frac{1-q}{2}} \tag{1.25}$$

and therefore with probability $1 - o(p_n^{-1})$,

$$\|\hat{A} - A\|_1 \leq C \left[\left(\frac{\log p_n}{nN} \right)^{1/4} + s \left(\frac{\log p_n}{nN} \right)^{\frac{1-q}{2}} \right] \tag{1.26}$$

for some constant C *that may depend on* K *and* q.

The proof of the above statements can be found in Section 1.5.5.

Remark 8. The bounds (1.24) and (1.25) can be interpreted as the estimation error and the approximation error respectively for using an estimator of A whose row support is contained in the set J. Note that the approximation error (1.25) is smaller if q is closer to 0; here we assume s does not grow too quickly relative to nN. In the most favorable setting where $s = O(1)$ and $0 < q < 1/2$ (strong sparsity regime), the aggregate error (1.26) is of the order $\left(\frac{\log p_n}{nN} \right)^{1/4}$, which clearly converges to zero as $nN \to \infty$. On the other hand, if $s \geq 1$ and $1/2 < q < 1$ (weak sparsity regime), the bound (1.26) is dominated by the term $s \left(\frac{\log p_n}{nN} \right)^{\frac{1-q}{2}}$.

Remark 9. We note again that the bound (1.26), which does not depend on p except for log terms, is valid for all parameter regimes and in particular for the high-dimensional setting where $p \gg \max(n, N)$. This justifies the use of our method for many text datasets where the number of unique words observed across all documents is extremely large. Also, the bound (1.26) does not depend on $\max_{j \in [p]} h_j$ or $\min_{j \in [p]} h_j$ and is thus completely unaffected by variations in word frequencies. In these regards, our result improves upon the theoretical guarantees presented in prior works such as Ke and Wang [2022], Bing et al. [2020a], Arora et al. [2012] and Wu et al. [2022].

1.2.4 Relaxation of the separability condition

Our main result (Theorem 7) may also hold under alternative identifiability assumptions on A if we use a suitable vertex hunting procedure that is effective even for non-separable point

clouds. Recall $v_1^*, \ldots, v_K^*$ are the simplex vertices from the oracle point cloud $\{r_j : j \in J\}$ in Definition 4 and $\hat{v}_1^*, \ldots, \hat{v}_K^*$ are the estimated vertices based on the point cloud $\{\hat{r}_j : j \in J\}$ in Definition 5. The assumptions we made concerning separability and vertex hunting efficiency, namely Assumptions 3 and 4, are only useful in our analysis insofar as they allow the following bound to hold with high probability:

$$\max_{k \in [K]} \|\hat{v}_k^* - v_k^*\|_2 \leq \max_{j \in J} \|\hat{r}_j - r_j\|_2 \tag{1.27}$$

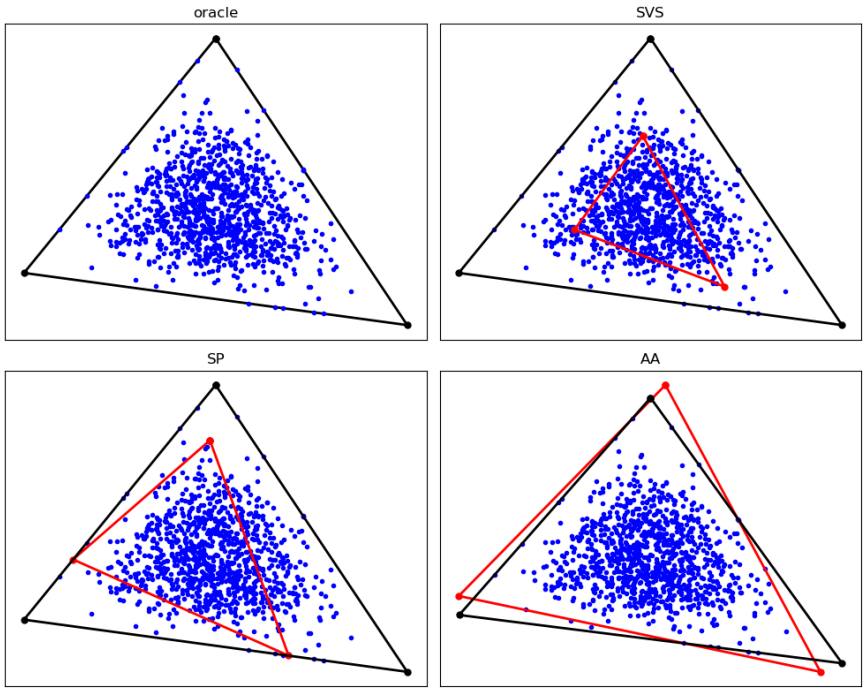


Figure 1.1: Top left: a non-separable point cloud (blue) contained in a simplex (black) with 3 vertices. Top right: Estimated vertices from SVS (red). Bottom left: Estimated vertices from SP (red). Bottom right: Estimated vertices from AA (red).

However, this bound may also hold if we adopt the identifiability assumption and the Archetype Analysis (AA) vertex hunting procedure proposed in Javadi and Montanari [2020]. Figure 1.1 provides an example of a non-separable point cloud where AA recovers the simplex vertices much more effectively than SP and SVS, which only search for possible vertices within the point cloud itself or its convex hull. Section 1.5.6 summarizes important results from

Javadi and Montanari [2020] that are relevant to this project. In our estimation procedure for A, once we obtain the matrix $\hat{R}$ whose rows are $\{\hat{r}_j : j \in J\}$ from Step 3 of Definition 5, the estimated simplex vertices can be obtained via AA by solving the following minimization problem:

$$\text{minimize } \mathscr{D}(V; \hat{R}) \text{ over } V \text{ s.t. } \mathscr{D}(\hat{r}_j; V) \leq \delta^2 \text{ for all } j \in J \tag{1.28}$$

Here the rows of V represent the simplex vertices; see Section 1.5.6 for the definition of the distance function $\mathscr{D}(\cdot, \cdot)$. The main theoretical result of Javadi and Montanari [2020] (Theorem 31) is that the AA algorithm is robust to noise in the point cloud under certain conditions. In particular, if we replace Assumptions 3 and 4 by the following assumptions:

(i) The matrix R from Step 2 of the oracle procedure (Definition 4) satisfies α-uniqueness for some absolute constant $\alpha > 0$. Here, α-uniqueness (described in Definition 6) is an identifiability assumption on the simplex vertices that is more general than separability.

(ii) The convex hull of the rows of R contains a $(K-1)$-dimensional ball of radius $\mu > 0$

(iii) The vertex hunting procedure $\mathcal{V}(\cdot)$ is defined by (1.28) with $\delta \asymp \left(\frac{\log p_n}{nN}\right)^{1/4}$. This value of δ is chosen based on Corollary 5 and Theorem 31.

then, in light of Theorem 31, (1.27) continues to hold and our main result, Theorem 7, remains valid. Alternatively, if we do not wish to use the α-uniqueness condition for identifiability, we can also assume that the distance from the oracle simplex vertices $\{v_1^*, \ldots, v_K^*\}$ to the convex hull of the oracle point cloud $\{r_j : j \in J\}$ is not larger than δ. In light of Theorem 32, this assumption can also be used to obtain (1.27).

Beside from Javadi and Montanari [2020], Ge and Zou [2015] also discusses an alternative identifiability assumption called *subset separability*. This notion can be illustrated by the point cloud in Figure 1.1 (top left), with $K = 3$. The point cloud (in blue) is contained in a triangle but is not separable as none of the triangle's vertices belongs to the point cloud.

However, each edge of the triangle contains several blue points and thus can clearly be identified from the point cloud. The vertices can then be identified by taking intersections of the edges. Ge and Zou [2015] also provides a vertex hunting procedure which, under subset separability and additional regularity assumptions, can also be shown to be robust to noise in the point cloud, in the sense of (1.27).

In terms of computation, Javadi and Montanari [2020] describes two algorithms to solve the following Lagrangian variant of (1.28):

$$\hat{V}_\lambda = \arg\min_V [\mathscr{D}(\hat{R}; V) + \lambda \mathscr{D}(V; \hat{R})] \tag{1.29}$$

Note that the objective function in (1.29) is non-convex and thus may have multiple minima. While AA may significantly reduce statistical error in the vertex hunting step when separability is not applicable, the trade-off is that its computational cost may be higher than that of the SP algorithm for separable point clouds.

1.2.5 Estimation of K

Our discussion so far assumes K is known. When K needs to be estimated, it is natural to examine the spectrum of any matrix that should be of rank K under the pLSI model.

Recall the definition of G in (1.12), and define $G_0 := \left(1 - \frac{1}{N}\right) D_0 D_0^T$. From Lemma 21, with probability $1 - o(p_n^{-1})$ we have (here C^* is a numerical constant not dependent on unobserved constants but may depend on the choice of α)

$$\|(G - G_0)_{JJ}\|_{\text{op}} \leq C^* K \sqrt{K} \sqrt{\frac{n \log p_n}{N}} \tag{1.30}$$

Furthermore one can show $[G_0]_{JJ}$ has rank K with high probability. By a simple application of Weyl's inequality, we then obtain the estimator (1.32) for K.

Lemma 8. *Let g_n be a quantity satisfying*

$$c\sqrt{\frac{nN}{\log p_n}} \geq g_n \geq C^* K\sqrt{K} \tag{1.31}$$

where C^* in (1.31) is the constant from (1.30) and c is another constant that may depend on K. If

$$\hat{K} := \max\left\{k : \lambda_k(G_{JJ}) > g_n\sqrt{\frac{n\log p_n}{N}}\right\} \tag{1.32}$$

then $\hat{K} = K$ with probability $1 - o(p_n^{-1})$.

The proof can be found under Corollary 22. In (1.31), the quantity g_n needs to be chosen to override the term $C^*K\sqrt{K}$ but cannot converge to $+\infty$ too quickly. Without any prior information on K, one can choose g_n to be a quantity that slowly converges to $+\infty$, such as $g_n = 8\log p_n$. If one has prior knowledge on an upper bound for K (for example if $K \leq 30$), the quantity g_n can be determined more specifically.

The estimator (1.32) is based on the bound (1.30), which depends on K and so we need to assume $g_n \geq C^*K\sqrt{K}$. However, one can also show that with probability at least $1-\frac{1}{n+p}$,

$$\|(D - D_0)_{J*}\|_{\text{op}} \leq 4\sqrt{\frac{n\log(n+p)}{N}} \tag{1.33}$$

(see Lemma 4 of Klopp et al. [2021]). This bound does not depend on K. Under similar assumptions on $\sigma_K(A)$ and $\sigma_K(W)$, we can consider the following estimator

$$\hat{K}' := \max\left\{k : \sigma_k(D_{J*}) > 4\sqrt{\frac{n\log(p+n)}{N}}\right\} \tag{1.34}$$

and also show that, based on (1.33), $\hat{K}' = K$ with high probability. The advantage of (1.32) over (1.34) is computational: both Step 2 of Definition 5 and (1.32) use the eigendecomposition of G_{JJ}, whereas (1.34) requires us to additionally perform SVD on D_{J*}.

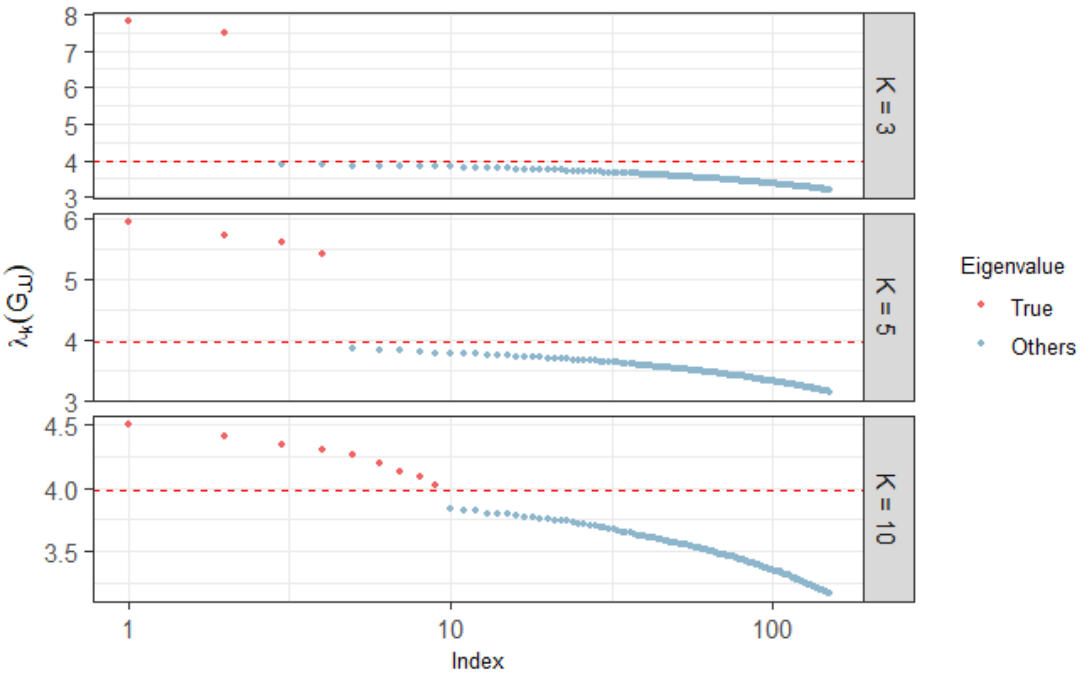


Figure 1.2: Scree plots of the eigenvalues of G_{JJ} for three synthetic datasets, with $K \in \{3, 5, 10\}$, $n = N = 500$ and $p = 5,000$. The x-axis is log-scaled. The red dots represent the largest K eigenvalues (excluding the largest one), while the blue dots represent all other eigenvalues.

There are many choices of the quantity g_n that may satisfy (1.31) when nN is sufficiently large. In practice, the estimation of K may be sensitive to the choice of the eigenvalue cutoff, and moreover real datasets may not always adhere to our assumptions. As Lemma 8 suggests the spectrum of G_{JJ} is useful for estimating K, we note that it is often possible to determine the eigenvalue cutoff by inspecting the scree plot of G_{JJ}'s eigenvalues. Figure 1.2 displays the scree plots for several synthetic datasets with different values of K. In some situations, the top K eigenvalues of G_{JJ} are separated from the other eigenvalues by a discernible gap, thus helping one to visually determine K. When such a gap is unavailable, one can use the Kneedle algorithm [Satopaa et al., 2011] to find the point of maximum curvature of the scree plot; this is a common technique to determine the number of principal components in principal component analysis.

1.3 Experiments with synthetic data

In this section, we assess the empirical performance of our estimator through a series of synthetic experiments[1]. The controlled environment provided by these experiments allows us to better understand the behavior of our method in different parameter regimes.

Throughout this section, we benchmark our estimator's performance against the following well-established methods: (a) Latent Dirichlet Allocation [Blei et al., 2003]; (b) the anchor word recovery (AWR) approach in Arora et al. [2012], a procedure based on the non-negative factorization of the second-order moment DD^T; (c) the Topic-SCORE procedure in Ke and Wang [2022]; and (d) the Sparse Topic Model solver proposed in Bing et al. [2020b]. We note the following regarding the procedure in Bing et al. [2020b]:

- This procedure removes infrequently occurring words in the same manner as ours, but with the threshold $\alpha\sqrt{\frac{\log p_n}{nN}}$ in (1.11) replaced by $\frac{7\log p_n}{nN}$. This threshold is lower than ours if $\frac{\log p_n}{nN}$ is sufficiently small. In practice, however, the constant 7 used in their threshold is quite large and thus leads to excessive thresholding in some of our datasets, especially when the word frequencies decay according to Zipf's law.

- This procedure requires a list of anchor words for each topic $k \in [K]$ as input, rather than just the number of topics K. We therefore need to estimate a partition of anchor words using a special procedure which is included in their original implementation. Clearly, whether the anchor words are estimated and partitioned correctly has an impact on the overall estimation of A.

We therefore caution the reader that these factors put the Sparse Topic Model solver of Bing et al. [2020b] at a comparative disadvantage in our experiments.

1. The code for our method and all the experiments presented in this section can be found on Github at the following link: `https://github.com/yatingliu2548/topic-modeling`

Data generation mechanism. For simplicity, we ensure all documents are of the same length N. For each experiment, we create a document-to-topic matrix $W \in \mathbb{R}^{K\times n}$ by independently drawing the columns $W_{*i} \in \mathbb{R}^K, i = 1, \ldots, n$ from the Dirichlet distribution with parameter $\boldsymbol{\alpha}_W = \mathbf{1}_K$. We generate the matrix $A \in \mathbb{R}^{p\times K}$ either without anchor words or with 5 anchor words per topic, in which case whenever word j is an anchor word for topic k, we set $A_{jk} = \delta_{\text{anchor}}$ where $\delta_{\text{anchor}} \in \{0.0001, 0.001, 0.01\}$. In order to mimic the behavior of real text data, the entries of column k of A corresponding to non-anchor words are then chosen such that they decay according to Zipf's law. This means for each column k of A, we ensure that the frequency $f_{(j)}$ of the j^{th} most frequent non-anchor word follows the pattern

$$f_{(j)} \propto \frac{1}{(j + b_{\text{zipf}})^{a_{\text{zipf}}}} \tag{1.35}$$

where $a_{\text{zipf}} = 1, b_{\text{zipf}} = 2.7$. Each column of A is subsequently normalized to unit ℓ_1-norm. The pattern (1.35) has indeed been empirically shown to hold approximatively for word frequencies in real datasets; see Zipf [1936] and Piantadosi [2014]. Figure 1.12a in Section 1.5.7 illustrates the distribution of word frequencies generated under our data generation mechanism.

Having specified both A and W, the observation matrix D is then generated according to the pLSI model described in Section 1.1.1. We fit our method and the four benchmarks while varying the values of n, p, N, and K. In all of our experiments, unless otherwise specified, the constant α in the threshold (1.11) is fixed at $\alpha = 0.005$. We evaluate the estimation error of all methods relative to the true underlying A by computing the ℓ_1 loss per topic

$$\mathcal{L}_1(\hat{A}, A) = \min_{\Pi\in\mathcal{P}} \frac{1}{K}\|\hat{A}\Pi - A\|_1$$

where $\mathcal{P}$ denotes the set of all $K \times K$ permutation matrices.

Varying (p, N, K)**.** We first provide a snapshot of our method's relative performance in different parameter regimes by fixing $n = 500$ and varying (p, N, K). Here we specify 5 anchor words per topic and set the anchor word frequency to $\delta_{\text{anchor}} = 10^{-3}$. The median $\mathcal{L}_1(\hat{A}, A)$-errors over 50 trials are plotted in Figure 1.3. As Figure 1.3 shows, our method (in blue) outperforms all other methods in most parameter regimes considered here. Interestingly, the estimation errors of AWR and LDA often appear constant as a function of document length N. As N increases, the errors from both Topic-SCORE and our method display a clearer pattern of consistency relative to AWR and LDA; this observation is also made by Ke and Wang [2022] in a similar experimental setup. However, our method's errors decay to zero much faster than all other benchmarks when the vocabulary size is large ($p \in \{5000, 10000\}$).

We note that in these experiments, the approach proposed by Bing et al. [2020b] does not perform very well. In particular, for small p and small N, the number of topics returned by this method is smaller than the expected number of topics K, which prevents us from comparing its results with all four other methods. On inspection, we find that this is due to over-thresholding of the vocabulary, which leaves too few words to reliably estimate the matrix A. To provide a fair comparison with Bing et al. [2020b], we also compare all five methods using the data generation mechanism proposed in Bing et al. [2020b]. This means that the non-anchor entries of each column of A no longer display the Zipf's law pattern (1.35), but instead are generated from a Uniform distribution. We note that this data generation mechanism ensures all the non-anchor words for each topic are of roughly equal frequency and is thus also favorable to Topic-SCORE [Ke and Wang, 2022], which assumes $\min_{j \in [p]} h_j \geq c\bar{h}$ where $\bar{h} := \frac{1}{p} \sum_{j=1}^{p} h_j$. The results are displayed in Figure 1.14 of Section 1.5.7. Under this uniform data generation mechanism, our method (with $\alpha = 0.005$) displays identical performance relative to Topic-SCORE, and both SCORE-based methods still perform well relative to other benchmarks in most parameter regimes. As expected, we also find that fewer words are removed by thresholding, in comparison with the Zipf's

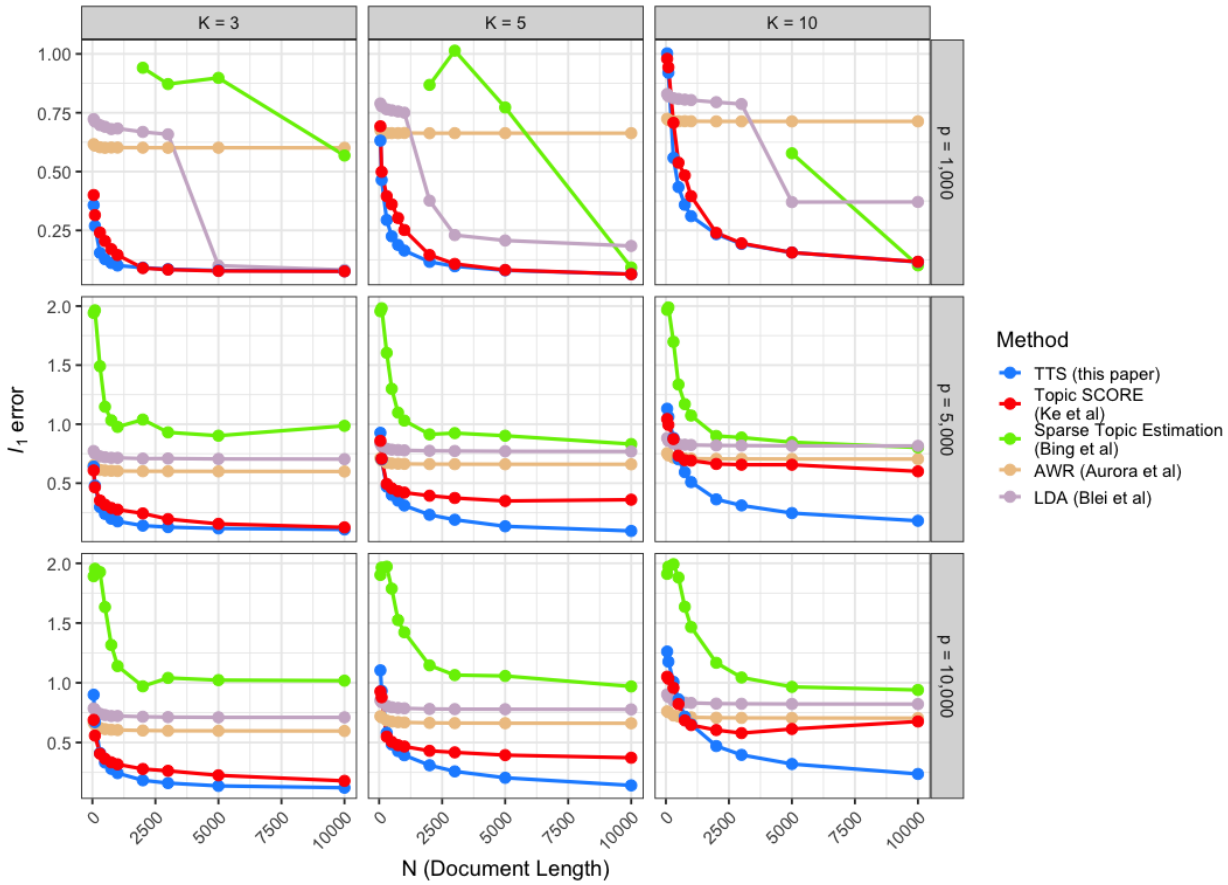


Figure 1.3: Median $\mathcal{L}_1(\hat{A}, A)$ errors for all methods based on 50 independent trials. Here the number of documents is fixed ($n = 500$). In each panel, the errors are plotted as a function of document length N (log-scaled on the x-axis). The panels display results for different values of (p, K), as specified by row and column labels.

law setting where our ℓ_q-sparsity assumption (1.19) is more likely to hold with small s and many more words occur infrequently. These experiments empirically suggest that 1) TTS improves upon the performance of Topic-SCORE when the columns of A exhibit a Zipf's law (or ℓ_q-sparsity) decay pattern, and 2) our procedure's performance remains reasonable and is similar to that of Topic-SCORE when the ℓ_q-sparsity assumption (1.19) is violated.

Varying the number of documents n. We now focus on the effect of varying n on the estimation error. Fixing this time $N = 500$ and $p = 10,000$, the $\mathcal{L}_1(\hat{A}, A)$-errors are presented in Figure 1.4 with $K = 5$ and $K = 10$. Our method (in blue) consistently outperforms other

methods and also displays a clear trend of consistency as n increases. When K increases, the estimation problem becomes more difficult due to the larger number of parameters, and so more documents are needed to achieve a reasonable performance. Nonetheless, our method still performs well when $K = 10$ and n is reasonably large, whereas the error from Topic-SCORE decays to zero very slowly with this larger value of K.

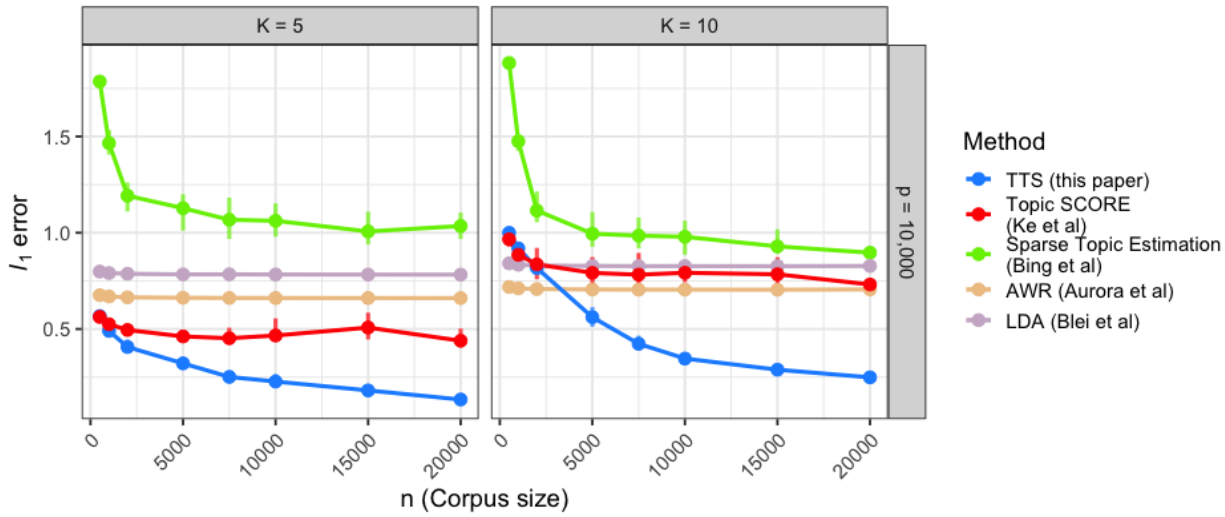
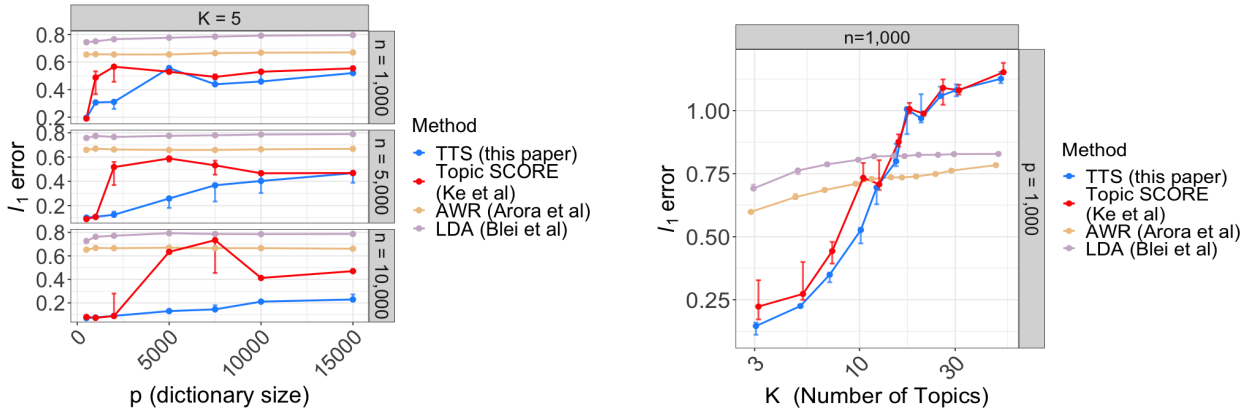


Figure 1.4: $\mathcal{L}_1(\hat{A}, A)$-errors from all methods as a function of n, for $K \in \{5, 10\}$ with p and N fixed. Vertical error bars centered about the median errors indicate the errors' interquartile ranges computed based on 50 independent trials.

Varying the dictionary size p. Figure 1.5a shows how the $\mathcal{L}_1(\hat{A}, A)$-errors vary as the vocabulary size p increases, with $N = 500$, $K = 5$ and $n \in \{1000, 5000, 10000\}$. We do not include the errors from the procedure in Bing et al. [2020b] as they are higher than those of LDA. As expected, the errors for all methods increase with the dictionary size p. However, our method mostly outperforms the other benchmarks, even in some high-dimensional parameter regimes where $p > \max(n, N)$. The performance of Topic-SCORE only converges to ours when p is too large relative to n, a setting which is challenging for all methods.

Additionally, our method also outperforms most other benchmarks in terms of compu-

tational runtime when p is large. Our method's runtime is similar to that of AWR and is consistently better than that of Topic-SCORE, primarily due to our thresholding of infrequent words before performing eigendecomposition.



(a) $\mathcal{L}_1(\hat{A}, A)$-errors as a function of p, with $K = 5$ and $N = 500$. Results are obtained based on 15 independent trials.

(b) $\mathcal{L}_1(\hat{A}, A)$-errors as a function of K, with $n = p = 10^3$ and $N = 500$. Results are obtained based on 50 independent trials.

Figure 1.5: $\mathcal{L}_1(\hat{A}, A)$-errors as a function of the dictionary size p (left) and the number of topics K (right). Vertical bars around median errors indicate interquartile ranges.

Varying the number of topics K. Figure 1.5b shows how the $\mathcal{L}_1(\hat{A}, A)$-errors vary as K increases, with $n = p = 1000$ and $N = 500$. The main observation here is that LDA and AWR may be preferable to our method if K is *a priori* known to be large while the dataset we possess is relatively small. As Figure 1.5b illustrates, the SCORE-based methods perform worse than LDA and AWR when $K > 15$, but this is because the number of documents is quite small in this experiment ($n = 1000$). If n and N are large enough, one can expect our method to accommodate a larger number of topics; see Figure 1.4 for an illustration.

Relaxation of the separability assumption. Section 1.2.4 suggests that the vertex hunting algorithm from Javadi and Montanari [2020] may reduce the vertex hunting error in some situations when separability fails to hold. Figure 1.6 compares the overall $\mathcal{L}_1(\hat{A}, A)$-errors

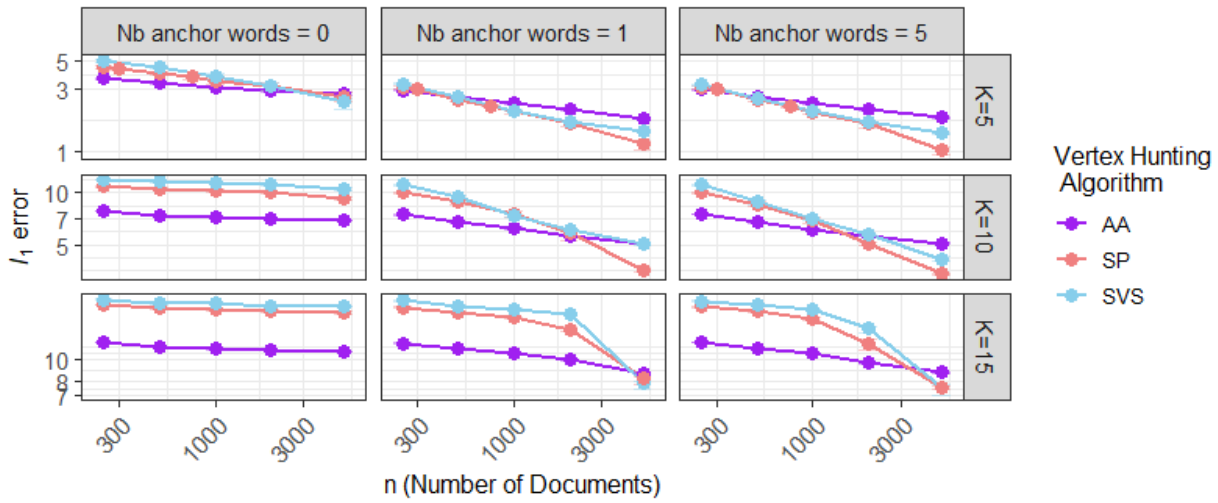
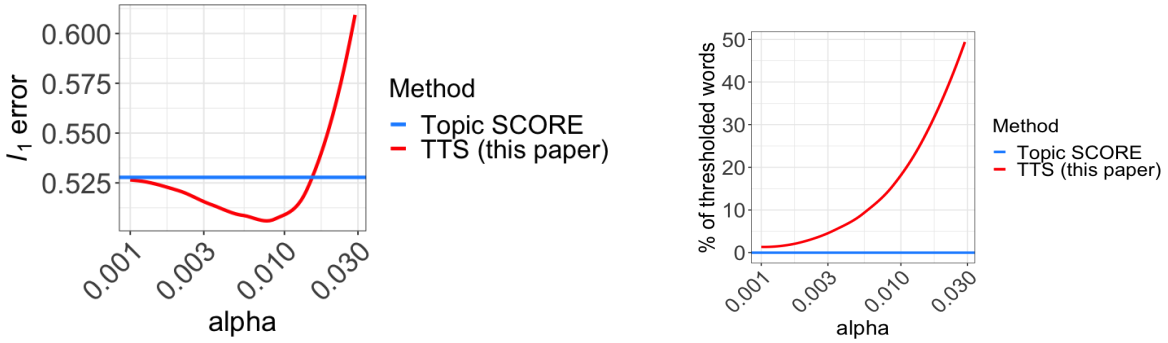


Figure 1.6: $\mathcal{L}_1(\hat{A}, A)$-errors as a function of n when we use three different vertex hunting algorithms in the vertex hunting step of TTS. Here, $p = 10^4$ and $N = 500$ are fixed, and $K \in \{5, 10, 15\}$. The number of topics per document is either 0, 1 or 5. Results are averaged over 50 independent experiments.

as a function of n when we use Successive Projection (SP), Sketched Vertex Analysis (SVS) and Archetype Analysis (AA) in the vertex hunting step of TTS. As expected, when there are no anchor words, using the AA algorithm rather than SP/SVS can significantly improve the estimation of $\hat{A}$, especially when K is large. Again, this is because SP and SVS are not designed for non-separable point clouds and also perform better with small K. In fact, the AA algorithm also often works well under separability, since the α-uniqueness condition in Javadi and Montanari [2020] is satisfied. The main trade-off for this stronger statistical performance is the computational cost of solving the non-convex optimization problem required by AA. Nonetheless, the fact that our method accommodates non-separable datasets makes TTS more widely applicable compared to methods based on anchor words identification, such as those proposed in Bing et al. [2020b] and Arora et al. [2012].

The importance of appropriate thresholding. Figure 1.7a shows how the $\mathcal{L}_1(\hat{A}, A)$-error varies as the threshold level in (1.11) increases from zero, and Figure 1.7b shows the

corresponding percentage of words removed. For this dataset, the performance of our method when $\alpha = 0$ (no thresholding) is not too different from Topic-SCORE. As the threshold level increases, infrequent words that contribute noise to the point cloud are removed, thus leading to an improvement in the estimation of A_{J*}. However, an excessively high threshold means we set too many rows of A to zero, and so the error from estimating A_{J^c*} becomes higher. This explains the pattern observed in Figure 1.7a, which demonstrates the importance of choosing a balanced threshold in our procedure.



(a) Average $\mathcal{L}_1(\hat{A}, A)$-error as a function of the threshold parameter α.

(b) Corresponding percentage of the words discarded by thresholding as a function of α.

Figure 1.7: $\mathcal{L}_1(\hat{A}, A)$-error averaged over 20 independent trials and the percentage of words removed as α increases, for a synthetic dataset with $p = 5000, n = N = 500$.

As we mentioned, the universal parameter α should be independent of (p, n, N, K). Our recommended value of $\alpha = 0.005$ is obtained based on numerous such experiments with synthetic data where we vary the values of (p, n, N, K). This choice of α also works well in all real data applications of Section 1.4, where several parameter regimes are involved.

Additional experiments and conclusion. We also evaluate the impact of other aspects of the data generation mechanism on our estimator's performance. We find that changing δ_{anchor}, which controls the frequency of anchor words, does not significantly impact the overall performance of TTS. This is an advantage of SCORE-based methods over methods

that rely on anchor words identification, which are often affected by the frequency of anchor words both in theory and in practice. Additionally, when we increase the parameter a_{zipf} in (1.35), we find that our estimator's performance improves significantly. This is not surprising as a larger a_{zipf} means the ordered entries of A's columns decay to zero faster, and our theoretical results also show that a strong sparsity regime (when q is close to 0 in Assumption 5) is favorable to our method. Further details about these experiments are deferred to Section 1.5.7. Finally, we check the performance of our method on a set of semi-synthetic experiments based on the Associated Press dataset (included in the R package `tm` [Feinerer et al., 2015]), thereby allowing us to test a different data generating mechanism. The results are also presented in Section 1.5.7.

Overall, we have illustrated that our method *(a) performs well in a wide variety of parameter regimes, and notably in the high-dimensional setting where p is large*, and *(b) performs well even if our sparsity assumption is violated* (see the discussion on the uniform data generation mechanism, and also note that we use a weak sparsity regime with $a_{\text{zipf}} \approx 1$ in most of our experiments). This makes our method applicable to the vast majority of real-world text datasets, which often are high-dimensional and exhibit Zipf's law decay. However, alternative methods such as LDA and AWR may still be competitive in some settings, especially when the pLSI model fails to hold or if the number of documents n and the document length N are unusually small relative to the number of topics K.

1.4 Practical applications in text analysis and beyond

In this section, we deploy our method on real-world datasets. Given the results of the previous section, we focus here on the comparison of our method with Topic-SCORE [Ke and Wang, 2022] and LDA [Blei et al., 2003].

Real datasets seldom have ground truth for A, and some may even lack an obvious choice for the number of topics K. Consequently, in this section we evaluate the estimators'

performance using, when appropriate, the following metrics:

(a) *Topic Resolution* as a measure of topic consistency. We fit each estimator on two disjoint halves of the data and report the cosine similarity between estimated topics (after an appropriate permutation of the columns of A). Mathematically, letting $\hat{A}^{(i)}, i \in \{1, 2\}$ denote the estimated topic-word matrices obtained for each half of the data, we define the "average topic resolution" η as the mean cosine similarity (a classical similarity metric in natural language processing) between aligned topics:

$$\eta = \max_{\sigma \in \Pi_K} \frac{1}{K} \sum_{k=1}^{K} \frac{\hat{A}_{*k}^{(1)\top} \hat{A}_{*\sigma(k)}^{(2)}}{\|\hat{A}_{*k}^{(1)}\|_2 \|\hat{A}_{*\sigma(k)}^{(2)}\|_2}, \tag{1.36}$$

where Π_K denotes the set of all permutations of $[K]$. Thus, higher resolution indicates better-defined and more consistent topic vectors (although this does not necessarily mean better ℓ_1-error).

(b) *Multiscale Topic Refinement and Coherence* (Fukuyama et al. [2021]): In the absence of an obvious number of topics K, we fit the method for multiple values of K and analyze the resulting topic hierarchy to check the stability of our estimator. We follow in particular the methodology of Fukuyama et al. [2021], which was developed to guide the choice of an appropriate number of topics K for LDA [Blei et al., 2003] by investigating the relationships among topics of increasing granularity. Given a hierarchy of topics, the method evaluates which topics consistently appear, constantly split, or are merely transient. We use these tools here (and its associated package `alto` [Fukuyama et al., 2021]) to analyze our estimator. The method of Fukuyama et al. [2021] starts by computing the alignment of topics across the hierarchy using the transport distance: for each K, this method computes how the mass of topic $j \in \{1, \cdots, K\}$ is split amongst the $K + 1$ topics at the next level of the hierarchy. We refer the reader to the original work by Fukuyama et al. [2021] for a more detailed explanation of topic transport align-

ment. Once the relationships between consecutive topic models have been established, the method of Fukuyama et al. [2021] allows visualization of (a) topic refinement (i.e., whether topics increase in granularity, as indicated by a small number of ancestors in the hierarchy; or conversely, whether topics are perpetually recombined from one level of the hierarchy to the next); and (b) topic coherence (whether a topic appears across multiple values of K). We choose here to favour methods with improved topic coherence and topic refinement, since there are markers of topic stability.

We explore the comparison between our method, LDA and Topic-SCORE under diverse parameter regimes (with varying n, N and p).

1.4.1 Research articles (high p, high n, low N)

For our first experiment, we consider a corpus of 20,140 research abstracts belonging to (at least) one of four categories: Computer Science, Mathematics, Physics and Statistics[2]. After pre-processing of the data (including the removal of standard stop words, numbers, and punctuation), our dataset involves a dictionary of size $p = 81,649$ and $n = 20,140$ documents with an average document size of $N = 157$ words.

We first evaluate the topic consistency of all methods in estimating the topic-word matrix A using the mean topic resolution defined in equation (1.36). Table 1.1 displays the average topic resolution over 25 random splits of the data. As highlighted in the introductory para-

Methods	Average Topic Resolution(η)	Interquantile range
LDA (Blei et al)	0.304	(0.270,0.330)
TTS (this paper)	0.332	(0.310,0.360)
Topic-SCORE (Ke et al)	0.145	(0.093,0.179)

Table 1.1: Average Topic Resolution on research article data. The interquartile range for the average topic resolution was computed over 25 random splits of the data.

2. The data is available on Kaggle at this link. Although the original data set comprises six topics (with the addition of Quantitative Biology and Finance), due to the low representation of these last two topics ($< 4\%$ of the data), we drop them from our analysis.

graph to this section, topic resolution can be taken as an indicator of the stability of the estimator of $\hat{A}$ between two separate portions of the data. A method that produces higher topic resolution with a narrower interquartile range indicates a more stable estimation of the topic-word matrix A. As shown in Table 1.1, our approach consistently outperforms LDA and Topic-SCORE on this metric; it offers the highest average topic resolution score. Topic-SCORE's performance exhibits more significant fluctuations, as indicated by its larger interquartile range.

Taking a closer look at the estimation of A, we consider the 10 most representative words generated by each of the three methods for every topic (obtained by selecting the top 10 largest entries in each column of $\hat{A}$). The results are presented in Tables 1.2, 1.3, and 1.4. For the topics of Computer Science and Statistics, the top 10 most representative words produced by our method agree with 70% of LDA's most representative words in the corresponding topics. There is much less agreement for the topic of Physics, but upon closer inspection we find that some of the words produced by our method in that category (such as 'magnetic', 'energy') are more indicative of the topic of Physics, whereas all of the top 10 words for Physics produced by LDA are generic words that can appear in other categories.

In contrast, the results of Topic-SCORE (Table 1.4) seem to diverge substantially from those of LDA and our method. It appears that the top 10 most representative words for Physics, Mathematics and Statistics from Topic-SCORE are dominated by infrequently occuring words and foreign words; the foreign words can be traced back to a few rare abstracts written in English and followed by a foreign language translation. This supports our hypothesis that Topic-SCORE amplifies the effects of infrequent words, unless significant *ad hoc* data pre-processing (removal or merger of rare words, and removal of documents with significant numbers of rare words) is applied.

In order to further investigate the performance gap between TTS and Topic-SCORE, we visualize the point cloud from both methods in Figure 1.8. As expected, we observe that the

	Top 10 most representative words per topic
Computer Science	“learning" “network" “networks" “model" "can" “neural" “deep" “using" "models" “data"
Physics	“model" “can" “system" “field" “energy" “systems" “magnetic" “models" “using" “phase"
Mathematics	“problem" “can" “algorithm" “show" “method" “paper" “results" “also" “time" “using"
Statistics	“data" “model" “can" “learning" “using" “models" “method" “approach" “based" “paper"

Table 1.2: Most common words found by our method

	Top 10 most representative words per topic
Computer Science	“data" “network" “learning" “networks" “can" “model" “using" “new" “paper" “based"
Physics	“show" “data" “analysis" “two" “can" “problem" “results" “field" “system" “performance"
Mathematics	“can" “used" “models" “using" “model" “paper" “number" “method" “proposed" “approach"
Statistics	“model" “results" “show" “can" “learning" “method" “using" “based" “data" “also"

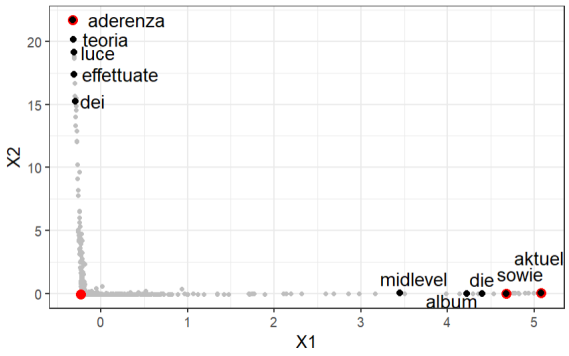
Table 1.3: Most common words found by LDA

Topic-SCORE point cloud is severely stretched by a set of low-frequency words that include several foreign words. Again, with the presence of many rare words in the dataset, the lack of thresholding and the use of the pre-SVD multiplication step in Topic-SCORE contribute to a significant distortion of the point cloud. In comparison, the thresholding approach we adopt yields a more compact point cloud. As demonstrated in Figure 1.8b, our method effectively recaptures the essential vertices of the point cloud simplex. A closer look at the words surrounding each vertex, as shown in Figure 1.8b, allows us to easily identify which simplex vertex belongs to which topic (Physics, Math, Computer Science and Statistics when moving in the anticlockwise direction). Under this “large p” regime and in the presence of a myriad of rare words that may introduce significant noise, our method not only distinguishes words effectively but also clusters them into well-defined topics.

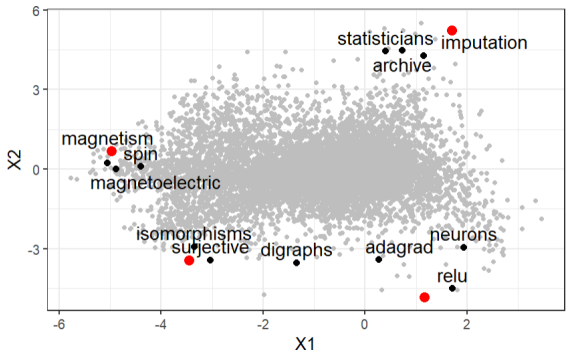
We note that this dataset comes with manually curated topic labels for each document.

	Top 10 most representative words per topic
Computer Science	“data" “can" “model" “using" "learning" “show" “results" “method" “paper" “also"
Physics	“della" “quantum" “theory" “del" “year" “teoria" “quantistica" “per" “nel" “delle"
Mathematics	“die" “der" “collectors" “problem" “able" “assumptions" “coupon" “wir" “based" “one"
Statistics	“der" “und" “music" “automatischen" “learning" “sheet" “die" “musikverfolgung" “deep" “algorithms"

Table 1.4: Most common words found by Topic SCORE



(a) Point cloud for $K = 4$ from Topic-SCORE



(b) Point cloud for $K = 4$ from our method

Figure 1.8: Comparison of the 3-dimensional point clouds from TTS (right) and Topic-SCORE (left), projected on the first two axes for visualization. Estimated vertices are colored red, and the point clouds are represented by gray dots. Most outlying words in Topic-SCORE's point cloud are thresholded away by TTS, thus contributing to higher point cloud stability for our method.

As a final verification, we analyze the performance of the different methods when used for recovering the ground truth labels for each document. Having estimated A, it is quite natural in light of the pLSI model to perform regression of D against $\hat{A}$ in order to yield an estimator of W. To this end, we use the estimation procedure for W in Ke and Wang [2022], where the problem of estimating W given $\hat{A}$ is reduced to a weighted constrained linear regression problem:

$$\forall i \in [n], \quad \hat{W}_{*i} = \text{argmin}_{\omega \in [0,1]^K} \frac{1}{p} \sum_{j=1}^{p} \frac{1}{M_{jj}} (D_{ji} - \sum_{k=1}^{K} \hat{A}_{jk} \omega_{ki})^2 \tag{1.37}$$

We strongly emphasize that the aim of this experiment is to evaluate the estimation of A,

and we do not claim here that our method provides state-of-the-art results in the estimation of W. Other potentially better estimation procedures are available for W, many of which do not require estimating A first. Rather, as topic labels are available for this dataset, we use this simple estimation procedure for W via $\hat{A}$ as another way of comparing the quality of $\hat{A}$ obtained from TTS, Topic-SCORE and LDA. Since the $\hat{W}$ obtained from (1.37) depends on $\hat{A}$ as input, it stands to reason that a better estimation procedure for A may be reflected in a better agreement between $\hat{W}$ and the provided topic labels for each document, if we use (1.37) to estimate W.

Let $y_{ki} = 1$ if document i is labeled as belonging to topic k (and $y_{ki} = 0$ otherwise). We compute the average l_1 distance $\mathcal{D}(\hat{W}, y)$ and cosine similarity S_k between the permuted matrix $\hat{W}$ and the provided labels y for each topic k as follows:

$$\mathcal{D}(\hat{W}, y) := \min_{\sigma \in \Pi_K} \frac{1}{nK} \sum_{ki} |\hat{W}_{\sigma(k)i} - y_{ki}|, \quad S_k = \max_{\sigma \in \Pi_K} \frac{\sum_{i=1}^{n} \hat{W}_{\sigma(k)i} y_{ki}}{\|\hat{W}_{\sigma(k)*}\|_2 \|y_{k*}\|_2} \tag{1.38}$$

Here, a smaller value of the l_1 distance or a larger value of the cosine similarity score between y and $\hat{W}$ indicate greater alignment with the provided topic labels. The results are displayed in Table 1.5.

Methods	S_{CS}	S_{Phys}	S_{Math}	S_{Stat}	$\bar{S}$	$\mathcal{D}(\hat{W}, y)$
LDA(Blei et al)	0.671	0.576	0.534	0.493	0.569	0.403
TTS(this paper)	0.610	0.748	0.636	0.494	0.622	0.305
Topic SCORE(Ke et al)	0.670	0.545	0.588	0.373	0.544	0.348

Table 1.5: The evaluation of $\hat{W}$ obtained via estimating A first by using the three methods. $\bar{S}$ is the average cosine similarity across all K topics

Table 1.5 indicates that our method improves the estimation of W overall and provides the best topic alignment on average, when using (1.37) to estimate W. This suggests that our procedure yields a more accurate estimator of A.

1.4.2 Single cell analysis (low p, high n, low N)

In this subsection, we consider a different application area for our methodology: the analysis of single-cell data. We revisit the mouse spleen dataset presented by Goltsev et al. [2018]. This dataset consists of a set of images from both healthy and diseased mouse spleens. Each sample undergoes staining with 30 different antibodies via the CODEX process, as detailed in Goltsev et al. [2018]. In Chen et al. [2020], each spleen sample is divided into a set of non-overlapping Voronoi bins, and the count of immune cell types is recorded in each bin. In this framework, each bin can be viewed as a document and cell types correspond to words. It is of interest to determine appropriate groupings of cell types (topics), as this may help one study the interactions between cells.

Since this dataset does not come with ground-truth labels, we sample two disjoint sets of size $n = 10,000$ out of the 100,840 Voronoi tessellations across all spleen samples (where 10,000 is a number chosen to be large enough to ensure a "high n" regime while still allowing all methods to have reasonable computational runtimes). On the contrary, there are only 24 different cell types ($p = 24$), while the average "document" length is $N = 11.2$ with an interquartile range of $(6, 16)$. While Chen et al. [2020] focus on evaluating estimators of the matrix W, here we repurpose the use of this dataset to study our estimator of A. In this dataset, the precise number of topics K is unknown. We thus apply the three methods for different values of K and use the metrics introduced at the beginning of this section (topic resolution, topic coherence and refinement) to compare the three methods. The results are presented in Figures 1.9 and 1.10.

Discussion of the results. Due to the structured nature of this dataset, all methods perform remarkably well, exhibiting an average topic similarity above 0.95. Going into more details, we see that our method outperforms Topic-SCORE in terms of topic resolution. In particular, Topic-SCORE (in red) appears to have more variable performance, as reflected

in its larger interquartile ranges and its jittery resolution as a function of K. Interestingly, in this specific instance, LDA seems to score higher on topic resolution (although we again emphasize that all methods perform very well on this metric). Additionally, Figure 1.10a shows the refinement and coherence of the topics for our method as K increases, in contrast to those of LDA in Figure 1.10b. In this data example, our method seems to provide topics with higher refinement (fewer ancestors per topic) and higher coherence (note in particular the stability of topic 1, 2, and 18) compared to LDA. In Figure 1.10b, it can be observed that topics 1, 2, and 18 are dispersed across different branches within the refinement plot as K varies.

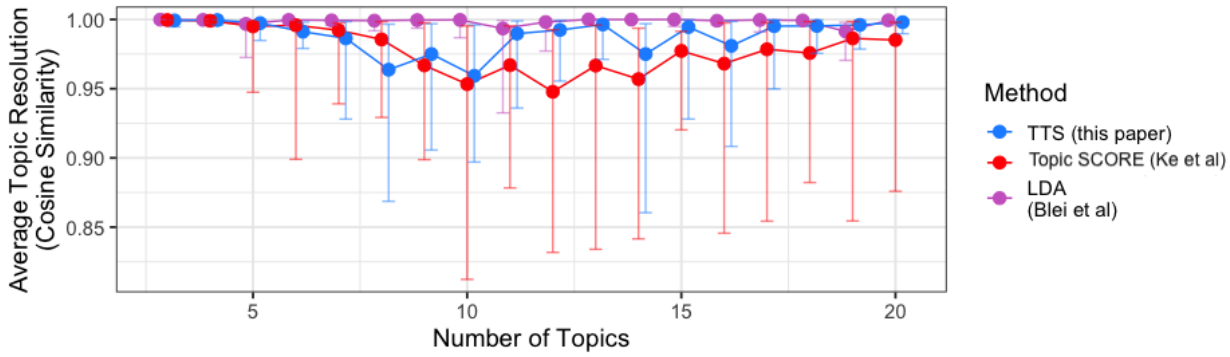
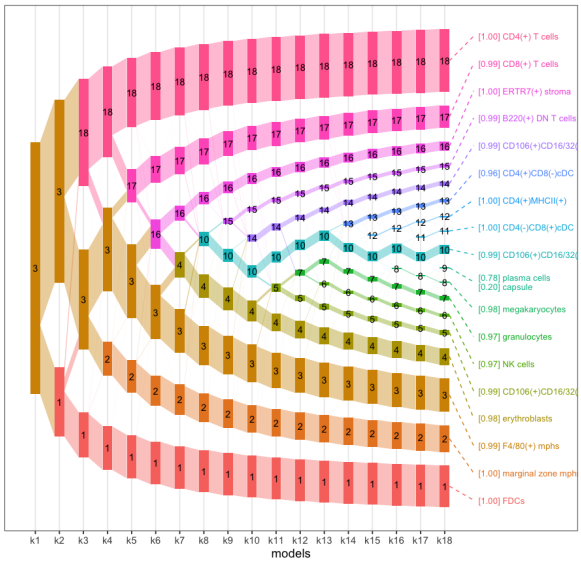


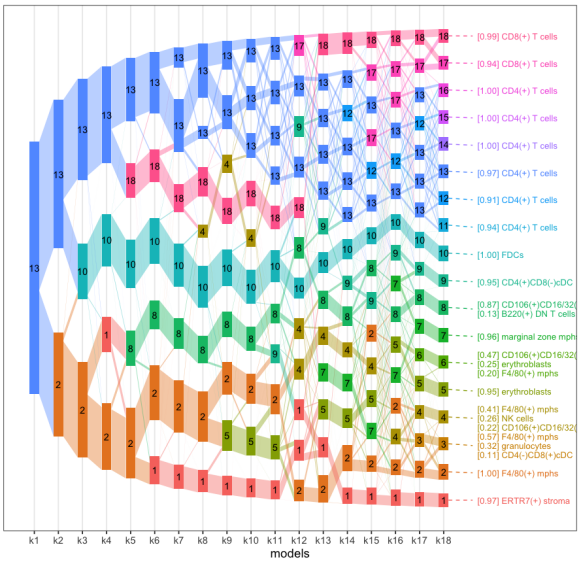
Figure 1.9: Median Topic Resolution as a function of K on the Mouse Spleen Data [Goltsev et al., 2018, Chen et al., 2020]. Vertical error bars represent the interquartile range for the average topic resolution scores over 25 trials.

1.4.3 Microbiome examples (low p, low n, high N)

We finish our discussion with an application of our method to microbiome data analysis. In particular, we reanalyze two datasets that have been previously analyzed through topic modeling: the colon dataset of Yachida et al. [2019] and the vaginal microbiome example of Callahan et al. [2017], which was re-analyzed in Fukuyama et al. [2021] using LDA. Microbiome data are represented in the form of a count matrix. In this matrix, each column corresponds to a different sample, while each row represents various taxa of bacteria. The



(a) Topic refinement for our method as K varies, provided by the package `alto` [Fukuyama et al., 2021].



(b) Topic refinement for LDA [Blei et al., 2003] as K varies, provided by the package `alto` [Fukuyama et al., 2021].

Figure 1.10: Comparison of the refinement and coherence of topics recovered using our method (left) and LDA (right).

entries within the matrix represent the abundance of each bacteria in a given sample. Taking samples to be documents and bacteria as words, topic modeling offers an interesting way of exploring communities of bacteria ("topics") [Sankaran and Holmes, 2019]. For the sake of conciseness, we present the results here for the colon dataset of Yachida et al. [2019], and refere the reader to Section 1.5.8 for the results on the other dataset.

After pre-processing and eliminating species with a relative abundance below 0.001%, this dataset contains microbiome counts for $p = 541$ distinct taxa from $n = 503$ samples. In contrast, the length of each "document" is extremely high, with around $N = 43$ million bacteria per sample. We test all three methods for different values of K and display the average topic resolution in Figure 1.11. On this metric, our method exhibits significantly better results than both LDA and Topic SCORE for up to 15 topics. After 15 topics, LDA outperforms all SCORE-based methods in terms of topic resolutions. However, this comes at a much higher computational cost: while each of the SCORE methods in this example

could be fitted in under a minute, each of the LDA fits took on the order of tens of minutes. Note that LDA's high topic resolution could also be due to the higher weight of the prior in the estimation of the topic-word matrix A, which, due to the relatively small size of the dataset, could have a stabilizing effect on estimation. On the other hand, the performance of Topic-SCORE quickly drops to 0.65 as K increases, before reaching a plateau at around $K \approx 10$. By contrast, for small K, our method exhibits a resolution up to 40% higher than Topic Score (for $K = 10$) before also decreasing as the number of topics increases.

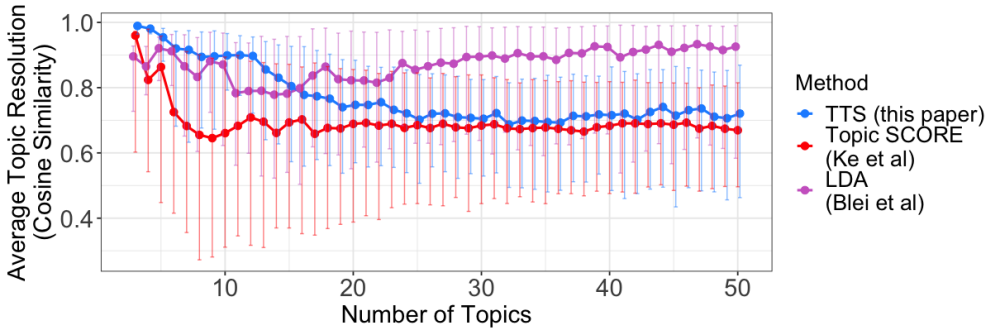


Figure 1.11: Topic resolution (measured by the average cosine similarity between halves of the data) of our method (in blue) and Topic-SCORE (red) on the microbiome dataset of Yachida et al. [2019]. Topic resolution is averaged over 25 random splits of the data.

To understand the gap in performance between Topic-SCORE and our method, we again visualize the point clouds obtained by both methods. The visualization can be found in Figure 1.17 in Section 1.5.8. Similarly to our first example with text analysis, we observe that the point cloud of Topic-SCORE is heavily distorted; in contrast, ours is more compact.

1.5 Proofs and supplementary materials

All proofs make use of notations described in Section 1.1.4. Assumptions 1-4 (which include separability) are assumed from Section 1.5.1 to Section 1.5.4, whereas the sparsity assumption (Assumption 5) is further imposed in Section 1.5.5.

1.5.1 Properties of the set J

Lemma 9 (Weak sparsity of A)**.** *Order the ℓ_2 row norms of A so that*

$$\|A_{(1)*}\|_2 \geq \cdots \geq \|A_{(p)*}\|_2$$

Then the matrix A satisfies $\max_{j\in[p]} j\|A_{(j)*}\|_2 \leq K$.

Proof. Observe that for any $j \in [p]$,

$$j\|A_{(j)*}\|_2 \leq \sum_{l=1}^{p} \|A_{l*}\|_2 \leq \sum_{l=1}^{p} \|A_{l*}\|_1 = K$$

since A contains only non-negative entries and each column sums up to 1. □

Lemma 10. *If $M_0 := diag(n^{-1}D_0\mathbf{1}_n)$ and $h_j := \|A_{j*}\|_1$, then for any $j \in [p]$,*

$$\sigma_K(\Sigma_W)h_j \leq M_0(j,j) \leq h_j$$

Proof. Note that

$$M_0(j,j) = \frac{1}{n}\sum_{i=1}^{n}[D_0]_{ji} = \frac{1}{n}\sum_{i=1}^{n}\sum_{k=1}^{K} A_{jk}W_{ki} = \sum_{k=1}^{K} A_{jk}\left(\frac{1}{n}\sum_{i=1}^{n} W_{ki}\right)$$

and observe that $h_j := \sum_{k=1}^{K} A_{jk}$ and for each $k \in [K]$ (recall $\Sigma_W := \frac{1}{n}WW^T$),

$$\sigma_K(\Sigma_W) \leq \Sigma_W(k,k) = \frac{1}{n}\sum_{i=1}^{n} W_{ki}^2 \leq \frac{1}{n}\sum_{i=1}^{n} W_{ki} \leq 1$$

□

Theorem 11. *Define $p_n := p \vee n$, $\tau_n := \sqrt{\frac{\log p_n}{Nn}}$. Let*

$$J := \{j \in [p] : M(j,j) > \alpha\tau_n\}, \qquad J_\pm := \{j \in [p] : M_0(j,j) > \alpha_\pm\alpha\tau_n\}$$

for some suitably chosen $\alpha > 0$ and $0 < \alpha_+ < 1 < \alpha_-$. The following statements hold:

(a) For a fixed $j \in [p]$, we have

$$\mathbb{P}(|M(j,j) - M_0(j,j)| \geq t) \leq 2\exp\left(-nNt^2/2\right)$$

(b) The event

$$\mathcal{E} := \{J_- \subseteq J \subseteq J_+\}$$

occurs with probability at least $1 - o(p_n^{-1})$. Here, we can select $\alpha = 8, \alpha_+ = \frac{1}{2}, \alpha_- = 2$. Note that this implies that on $\mathcal{E}$, $\min_{j\in J} h_j > \alpha_+\alpha\tau_n$.

(c) We have

$$\|A_{J_-^c *}\|_F^2 \leq \left[2K(\beta\tau_n)^{-1} \wedge p\right](\beta\tau_n)^2 = o(1)$$

where $\beta := \frac{\alpha_-\alpha}{\sigma_K(\Sigma_W)} = \frac{16}{\sigma_K(\Sigma_W)} \leq C$, and the same is true of $\|A_{J^c *}\|_F^2$ on event $\mathcal{E}$.

(d) $|J_+| \leq \frac{K\tau_n^{-1}}{\alpha\alpha_+} \wedge p = \frac{K\tau_n^{-1}}{4} \wedge p$, and the same is true of $|J|$ on event $\mathcal{E}$.

(e) $\sigma_K(A_{J*}) > c\sqrt{K}$ for some absolute constant $c > 0$ on event $\mathcal{E}$. This implies $|J| \geq K$ and $[D_0]_{J*}$ have rank K on $\mathcal{E}$.

(f) There exists an absolute constant $c \in (0,1)$ such that the entries of $A_{J*}^T A_{J*}$ are all greater than c on event $\mathcal{E}$.

Proof. (a) Denote $Z := D - D_0$. We introduce the set of p-dimensional one-hot vectors

$$\{T_{im} : 1 \leq i \leq n, 1 \leq m \leq N\}$$

for each word in the dataset; note that $T_{im} \sim \text{Multinomial}(1, [D_0]_{*i})$ and these one-hot

vectors are mutually independent. It follows that each column of Z satisfies

$$[Z]_{*i} = \frac{1}{N}\sum_{m=1}^{N}(T_{im} - \mathbb{E}[T_{im}]) \tag{1.39}$$

Note that for a given $j \in [p]$:

$$M(j,j) - M_0(j,j) = \frac{1}{n}\sum_{i=1}^{n} Z_{ji} = \frac{1}{nN}\sum_{i=1}^{n}\sum_{m=1}^{N}(T_{im}(j) - \mathbb{E}[T_{im}(j)]) \tag{1.40}$$

and since $|T_{im}(j) - \mathbb{E}[T_{im}(j)]| \leq 1$, we can apply Hoeffding's inequality to conclude

$$\mathbb{P}(|M(j,j) - M_0(j,j)| \geq t) \leq 2\exp\left(-nNt^2/2\right)$$

(b) Note that $\alpha_- > 1$. We have

$$\begin{aligned}
\mathbb{P}(J_- \not\subseteq J) &= \mathbb{P}\left(\cup_{j\in J_-}\{M(j,j) \leq \alpha\tau_n\}\right) \\
&\leq \sum_{j\in J_-}\mathbb{P}\left(M(j,j) \leq \alpha\tau_n\right) \\
&\leq \sum_{j\in J_-}\mathbb{P}\left(M(j,j) - M_0(j,j) \leq \alpha\tau_n - \alpha_-\alpha\tau_n\right) \\
&\leq \sum_{j\in J_-}\mathbb{P}\left(|M(j,j) - M_0(j,j)| \geq (\alpha_- - 1)\alpha\tau_n\right) \\
&\leq \sum_{j\in J_-} 2\exp\left(-Nn(\alpha_- - 1)^2\alpha^2\tau_n^2/2\right) \\
&\leq 2p_n^{1-(\alpha_- - 1)^2\alpha^2/2}
\end{aligned}$$

where in the last step we used $|J_-| \leq p$. We want to choose $1 - \frac{(\alpha_- - 1)^2\alpha^2}{2} < -1$ or equivalently $(\alpha_- - 1)\alpha > 2$.

Note that $0 < \alpha_+ < 1$. We further have

$$\begin{aligned}
\mathbb{P}(J \not\subseteq J_+) &= \mathbb{P}\left(\cup_{j \in J_+^c} \{M(j,j) > \alpha\tau_n\}\right) \\
&\leq \sum_{j \in J_+^c} \mathbb{P}\left(M(j,j) - M_0(j,j) > (\alpha - \alpha\alpha_+)\tau_n\right) \\
&\leq \sum_{j \in J_+^c} \mathbb{P}\left(|M(j,j) - M_0(j,j)| > (1-\alpha_+)\alpha\sqrt{\frac{\log p_n}{nN}}\right) \\
&\leq \sum_{j \in J_+^c} 2\exp\left(-\frac{Nn(1-\alpha_+)^2\alpha^2 \log p_n}{2Nn}\right) \\
&\leq 2p_n^{1-(1-\alpha_+)^2\alpha^2/2}
\end{aligned}$$

Again, we want to choose $1 - \frac{(1-\alpha_+)^2\alpha^2}{2} < -1$ or equivalently $(1-\alpha_+)\alpha > 2$. A suitable choice is $\alpha = 8, \alpha_+ = \frac{1}{2}, \alpha_- = 2$.

(c) From Lemma 10, we have

$$M_0(j,j) \geq \sigma_K(\Sigma_W)\|A_{j*}\|_1 \geq \sigma_K(\Sigma_W)\|A_{j*}\|_2$$

and so if we define

$$L := \{j \in [p] : \|A_{j*}\|_2 > \beta\tau_n\} \quad \text{where } \beta := \frac{\alpha_-\alpha}{\sigma_K(\Sigma_W)} = \frac{16}{\sigma_K(\Sigma_W)}$$

then $L \subseteq J_-$ by definition of J_-, and thus $\|A_{J_-^c *}\|_F \leq \|A_{L^c *}\|_F$. Now, if we order the

ℓ_2 row norms $\|A_{(1)*}\|_2 \geq \cdots \geq \|A_{(p)*}\|_2$ and apply Lemma 9,

$$\begin{aligned}\|A_{L^c*}\|_F^2 &= \sum_{j\notin L} \|A_{j*}\|_2^2 = \sum_{j\notin L} \min(\|A_{j*}\|_2^2, \beta^2\tau_n^2) \\ &\leq \sum_{j=1}^{p} \min(\|A_{(j)*}\|_2^2, \beta^2\tau_n^2) \leq \sum_{j=1}^{p} \min\left(\frac{K^2}{j^2}, \beta^2\tau_n^2\right) \\ &\leq \int_0^\infty \min(\beta^2\tau_n^2, K^2t^{-2})dt\end{aligned}$$

Let t_0 satisfies $\beta^2\tau_n^2 = K^2t^{-2}$ or $t_0 = \frac{K}{\beta\tau_n}$. We continue:

$$\begin{aligned}\|A_{L^c*}\|_F^2 &\leq t_0\beta^2\tau_n^2 + K^2\int_{t_0}^\infty t^{-2}dt \\ &= t_0\beta^2\tau_n^2 + K^2t_0^{-1} = 2t_0\beta^2\tau_n^2 \\ &= 2K\beta\tau_n = 2K\beta\sqrt{\frac{\log p_n}{Nn}} = o(1)\end{aligned}$$

given our assumption that $\sigma_K(\Sigma_W) > c$ for some absolute constant $c > 0$. Moreover, it is also clear from the definition of L that

$$\|A_{L^c*}\|_F^2 \leq p(\beta\tau_n)^2$$

(d) For all $j \in J_+ := \{j \in [p] : M_0(j,j) > \alpha\alpha_+\tau_n\}$, note that $h_j \geq M_0(j,j) > \alpha_+\alpha\tau_n$. Then observe that

$$K = \sum_{j=1}^{p} h_j \geq \sum_{j\in J_+} h_j \geq |J_+|\alpha\alpha_+\tau_n = 4|J_+|\tau_n$$

(e) Here we use the assumption that $\sigma_K(A) > c\sqrt{K}$ for some absolute constant $c > 0$.

Observe that by Weyl's inequality for singular values, on event $\mathcal{E}$ we have

$$\begin{aligned}\sigma_K(A_{J*}) &\geq \sigma_K(A) - \|A_{J^c*}\|_{\text{op}} \\ &\geq \sigma_K(A) - \|A_{J^c*}\|_F \\ &\geq c\sqrt{K} - o(1) \geq c\sqrt{K}/2\end{aligned}$$

when nN is sufficiently large, since in part (c) we have shown $\|A_{J^c*}\|_F \leq C\left(\frac{\log p_n}{Nn}\right)^{1/4}$. Hence, A_{J*} has rank K on $\mathcal{E}$, and by Sylvester's rank inequality,

$$K = \text{rank}(A_{J*}) + \text{rank}(W) - K \leq \text{rank}([D_0]_{J*}) \leq \text{rank}(A_{J*}) = K$$

(f) Here we use the assumption that the entries of A^TA are bounded below by an absolute constant. For any $k, l \in [K]$, since $A^TA = A_{J*}^TA_{J*} + A_{J^c*}^TA_{J^c*}$, on event $\mathcal{E}$ the (k,l)-entry of $A_{J*}^TA_{J*}$ satisfies

$$\begin{aligned}(A_{J*}^TA_{J*})(k,l) &= (A^TA)(k,l) - \sum_{j\notin J} A_{jk}A_{jl} \\ &\geq c - \|A_{J^ck}\|_2\|A_{J^cl}\|_2 \\ &\geq c - \|A_{J^c*}\|_F^2 = c - o(1) \geq c/2\end{aligned}$$

when nN is sufficiently large.

□

1.5.2 Properties of unobserved quantities

Lemma 12. *The following statements are true:*

(a) $\sigma_1(A) \leq \sqrt{K}$ *and* $\sigma_1(\Sigma_W) \leq 1$, *where* $\Sigma_W := \frac{1}{n}WW^T$.

(b) *If* $\Xi \in \mathbb{R}^{|J| \times K}$ *contains the first* K *left singular vectors of* $[D_0]_{J*}$*, then* $\Xi^T A_{J*}$ *is invertible. If* $V := (\Xi^T A_{J*})^{-1} \in \mathbb{R}^{K \times K}$ *then* V *satisfies the following:*

(i) $\Xi = A_{J*}V$

(ii) *The singular values of* V *are the inverses of the singular values of* A_{J*}

(iii) *The columns* $V_1, \ldots, V_K$ *of* V *are eigenvectors of the matrix* $\Theta := \Sigma_W A_{J*}^T A_{J*}$*, associated with the eigenvalues*

$$\lambda_k(\Theta) = \frac{\sigma_k^2([D_0]_{J*})}{n} \quad \text{for } 1 \leq k \leq K$$

(c) *The matrix* $\Theta_0 := \Sigma_W A^T A \in \mathbb{R}^{K \times K}$ *satisfies the following:*

(i) *The entries of* Θ_0 *are all positive and bounded below by an absolute constant* $c_1 > 0$*.*

(ii) *The gap between its first two eigenvalues is bounded below by an absolute constant* $c_2 > 0$*.*

(iii) *The entries of the unit-norm leading positive eigenvector of* Θ_0 *are all bounded below by an absolute constant* $c_3 > 0$*.*

(d) *On event* $\mathcal{E}$*, the results of part (c) also apply to* Θ*, possibly with smaller absolute constants* $c_1, c_2, c_3 > 0$*.*

(e) *There exist absolute constants* $c, C > 0$ *such that on* $\mathcal{E}$*, the entries of the first column of* V *satisfy*

$$\frac{c}{\sqrt{K}} \leq \min_{k \in [K]} V_1(k) \leq \max_{k \in [K]} V_1(k) \leq \frac{C}{\sqrt{K}}$$

and if $\xi_1, \ldots, \xi_K$ *are the columns of* Ξ*, then for any* $j \in J$*, its first column satisfies*

$$\frac{ch_j}{\sqrt{K}} \leq \xi_1(j) \leq \frac{Ch_j}{\sqrt{K}}$$

(f) Let $Q \in \mathbb{R}^{K\times K}$ be defined by $Q^T = [diag(V_1)]^{-1}V$, and note that the entries of the first row of Q are all equal to 1. If $v_1^*, \ldots, v_K^* \in \mathbb{R}^{K-1}$ are defined by the relation

$$Q = \begin{pmatrix} 1 & \ldots & 1 \\ v_1^* & \ldots & v_K^* \end{pmatrix}$$

then we have

$$c \leq \sigma_K(Q) \leq \sigma_1(Q) \leq C$$

Consequently, $v_1^*, \ldots, v_K^*$ are affinely independent (which means the simplex defined by their convex hull is non-degenerate) and $\max_{k\in[K]} \|v_k^*\|_2 \leq C$.

Proof. (a) The k^{th} diagonal entry of A^TA is $\|A_{*k}\|_2^2 \leq \|A_{*k}\|_1 = 1$, so $\text{tr}(A^TA) \leq K$ which implies $\sigma_1(A) \leq \sqrt{K}$. Similarly,

$$\sigma_1(\Sigma_W) = \frac{\sigma_1(W^TW)}{n} \leq \frac{\text{tr}(W^TW)}{n} = \frac{\sum_{i=1}^n \|W_{*i}\|_2^2}{n} \leq \frac{\sum_{i=1}^n \|W_{*i}\|_1}{n} = 1$$

(b) By singular value decomposition, we have

$$[D_0]_{J*} = \Xi\Lambda B^T$$

where $\Lambda = \text{diag}(\sigma_1, \ldots, \sigma_K) \in \mathbb{R}^{K\times K}$ contains the singular values of $[D_0]_{J*}$ and $B \in \mathbb{R}^{n\times K}$ contains its right singular vectors. Here $\Xi^T\Xi = B^TB = I_K$. Then

$$\Xi = \Xi\Lambda B^TB\Lambda^{-1} = [D_0]_{J*}B\Lambda^{-1} = A_{J*}WB\Lambda^{-1}$$

If we let $V = WB\Lambda^{-1} \in \mathbb{R}^{K\times K}$, then $\Xi = A_{J*}V$. Furthermore, since

$$\Xi^T\Xi = \Xi^TA_{J*}V = I_K$$

we can see that V can be defined as the inverse of $\Xi^T A_{J*}$, thus proving (i). Also, since

$$\Xi^T \Xi = V^T A_{J*}^T A_{J*} V = I_K$$

we have $VV^T A_{J*}^T A_{J*} VV^T = VV^T$, which implies $VV^T = (A_{J*}^T A_{J*})^{-1}$ and (ii) follows. Now let $\xi_1, \ldots, \xi_K$ be the columns of Ξ, and let $V_1, \ldots, V_K$ be the columns of V. Note that $\xi_k = A_{J*} V_k$. Since $[D_0]_{J*}[D_0]_{J*}^T \xi_k = \sigma_k^2 \xi_k$ and $\Sigma_W := \frac{1}{n} WW^T$, we have

$$A_{J*}\Sigma_W A_{J*}^T A_{J*} V_k = A_{J*}\Sigma_W A_{J*}^T \xi_k = \frac{1}{n}[D_0]_{J*}[D_0]_{J*}^T \xi_k = \frac{\sigma_k^2}{n}\xi_k = \frac{\sigma_k^2}{n} A_{J*} V_k$$

Multiplying both sides by $(A_{J*}^T A_{J*})^{-1} A_{J*}^T$ on the left, we have

$$\Sigma_W A_{J*}^T A_{J*} V_k = \frac{\sigma_k^2}{n} V_k$$

and so $V_1, \ldots, V_K$ are eigenvectors (not necessarily orthonormal) of $\Theta := \Sigma_W A_{J*}^T A_{J*}$, associated with eigenvalues σ_k^2/n for $k = 1, \ldots, K$. This proves (iii).

(c) For any $1 \le k, l \le K$, (i) follows from our assumptions:

$$\begin{aligned}
\Theta_0(k,l) &= \sum_{s=1}^{K} \Sigma_W(k,s) \cdot (A^T A)(s,l) \\
&\ge \min_{t,u \in [K]} (A^T A)(t,u) \cdot \sum_{s=1}^{K} \Sigma_W(k,s) \\
&\ge \min_{t,u \in [K]} (A^T A)(t,u) \cdot \Sigma_W(k,k) \\
&\ge \min_{t,u \in [K]} (A^T A)(t,u) \cdot \sigma_K(\Sigma_W) > c
\end{aligned}$$

Let $\gamma(\Theta_0) := \lambda_1(\Theta_0) - \lambda_2(\Theta_0) \ge 0$ denote the gap between the first two eigenvalues of Θ_0. The proof of (ii) is an asymptotic argument. If we consider a sequence $\{\Theta_0^{(n)}\}$ that

varies with n as $n \to \infty$, then (ii) follows if we can establish that

$$\liminf_{n\to\infty} \gamma(\Theta_0^{(n)}) > 0$$

Assume to the contrary that $\liminf_{n\to\infty} \gamma(\Theta_0^{(n)}) = 0$. Then there exists a subsequence $\{\Theta_0^{(n_m)}\}_{m=1}^{\infty}$ such that the gap between the first two eigenvalues decays to zero. Since

$$\|\Theta_0^{(n)}\|_{\text{op}} \leq \|\Sigma_W^{(n)}\|_{\text{op}} \|A^{(n)}\|_{\text{op}}^2 \leq K$$

and K is fixed as n varies, there must exist a further subsequence that converges to some matrix $\Theta_0^{(\infty)}$. By part (i), this matrix $\Theta_0^{(\infty)}$ has entries that are bounded below by some absolute constant $c > 0$, and yet its first two eigenvalues are equal (by eigenvalue continuity). By Perron's theorem (see Section 8.2 of Horn and Johnson [2012] for a reference), such a matrix $\Theta_0^{(\infty)}$ cannot exist.

(iii) is also proven in a similar manner. Let $\eta_0^{(n)} \in \mathbb{R}^K$ denote the leading unit-norm positive eigenvector of $\Theta_0^{(n)}$; its entries are all positive by Perron's theorem. Suppose there exists some $k \in [K]$ such that

$$\liminf_{n\to\infty} \eta_0^{(n)}(k) = 0$$

Note that the mapping from a matrix in $\mathbb{R}^{K\times K}$ with strictly positive entries to its leading unit-norm positive eigenvector is continuous (this will be further elaborated in part (d)). Again, this implies that there exists a subsequence $\{\Theta_0^{(n_m)}\}$ that converges to some $\Theta_0^{(\infty)}$ having strictly positive entries, and yet its leading eigenvector contains a zero entry. This contradicts Perron's theorem.

(d) In light of Theorem 11(f) which shows $A_{J*}^T A_{J*}$ has entries bounded below by $c > 0$ on $\mathcal{E}$, (i) is proven similarly as in part (c).

We will first show (iii). Note that we refrain from applying the asymptotic arguments of part (c) directly to Θ since, unlike Θ_0, Θ depends on J which is random. Also, the $\sin\theta$ theorem is not applicable to eigenvectors of Θ and Θ_0 as these matrices are not symmetric. Hence, we opt for the approach presented below.

Define the open domain

$$E = \{\Psi \in \mathbb{R}^{K\times K} : \Psi(k,l) > 0 \text{ for all } k,l \in [K]\}$$

and define $\mathbf{f} : E \to \mathbb{R}^K$ as the function mapping a matrix in E to its leading unit-norm positive eigenvector. Also, fix $\Psi_0 \in E$ and $1 \le k,l \le K$. For any real-valued t in a neighborhood of zero, consider the function

$$f_{kl}^{\Psi_0}(t) := \mathbf{f}(\Psi_0 + t\mathbf{e}_k\mathbf{e}_l^T)$$

where $\mathbf{e}_k$ and $\mathbf{e}_l$ are the k^{th} and l^{th} canonical basis vectors of $\mathbb{R}^k$ respectively.

Since the algebraic multiplicity of the first eigenvalue of any matrix in E is 1 (Perron's theorem), by Theorem 2 of Greenbaum et al. [2020], for any $\Psi_0 \in E$ and any $k,l \in [K]$, the function $f_{kl}^{\Psi_0}(\cdot)$ is continuously differentiable around 0 (more specifically, one can write $f_{kl}^{\Psi_0}(t) = \frac{x(t)}{\|x(t)\|_2}$ for some eigenvector function $x(t)$ that is analytic in a neighborhood of 0). Therefore, the function $\mathbf{f}$ itself is continuously differentiable on E, and we can define its derivative $\mathbf{f}'(\Psi)$ as a matrix in $\mathbb{R}^{K^2\times K}$ containing all the partial derivatives of $\mathbf{f}$ at $\Psi \in E$. Since these partial derivatives are all continuous, $\mathbf{f}' : E \to \mathbb{R}^{K^2\times K}$ is a continuous function.

Now if $c > 0$ is an absolute constant such that all the entries of Θ_0 and Θ are greater than c (the latter on event $\mathcal{E}$), then Θ and Θ_0 belong to the set

$$E' = \{\Psi \in \mathbb{R}^{K\times K} : \Psi(k,l) \ge c \text{ for all } k,l \in [K] \text{ and } \|\Psi\|_{\text{op}} \le K\}$$

which is a compact subset of E. Let η and η_0 be the unit-norm positive first eigenvectors of Θ and Θ_0 respectively. On event $\mathcal{E}$, by Theorem 9.19 of Rudin et al. [1976],

$$\begin{aligned}\|\eta - \eta_0\|_2 = \|\mathbf{f}(\Theta) - \mathbf{f}(\Theta_0)\|_2 &\leq \max_{\Psi \in E'} \|\mathbf{f}'(\Psi)\|_{\text{op}} \|\Theta - \Theta_0\|_F \\ &\leq C\|\Sigma_W\|_{\text{op}} \|A^T A - A_{J*}^T A_{J*}\|_F \\ &= C\|\Sigma_W\|_{\text{op}} \|A_{J^c*}^T A_{J^c*}\|_F \leq C\|A_{J^c*}\|_F^2 = o(1)\end{aligned}$$

where we note that $A^T A = A_{J*}^T A_{J*} + A_{J^c*}^T A_{J^c*}$. Hence, for any $k \in [K]$,

$$\eta(k) \geq \eta_0(k) - \|\eta - \eta_0\|_2 \geq c - o(1) > c/2$$

if nN is sufficiently large. We have shown $\min_k \eta(k) \geq c/2 > 0$ on $\mathcal{E}$.

As for (ii), we have shown in (b)(iii) for Θ (and the proof is similar for Θ_0) that

$$\lambda_k(\Theta) = \frac{\sigma_k^2([D_0]_{J*})}{n}, \quad \lambda_k(\Theta_0) = \frac{\sigma_k^2(D_0)}{n}$$

Note that since $\|A\|_{\text{op}} \leq \sqrt{K}$ and $\|W\|_{\text{op}} \leq \sqrt{n}$,

$$\max[\sigma_k([D_0]_{J*}), \sigma_k(D_0)] \leq \|A\|_{\text{op}} \|W\|_{\text{op}} \leq \sqrt{Kn}$$

and by Weyl's inequality for singular values (which can be applied after appending zero rows to the matrix A_{J*} so as to match the dimension of A),

$$\begin{aligned}|\lambda_k(\Theta) - \lambda_k(\Theta_0)| &\leq \frac{|\sigma_k([D_0]_{J*}) - \sigma_k(D_0)||\sigma_k([D_0]_{J*}) + \sigma_k(D_0)|}{n} \\ &\leq \frac{\|A_{J^c*}\|_{\text{op}} \|W\|_{\text{op}} (2\sqrt{Kn})}{n} \\ &\leq 2\sqrt{K}\|A_{J^c*}\|_{\text{op}} = o(1)\end{aligned}$$

on event $\mathcal{E}$, so

$$\begin{aligned}|\lambda_1(\Theta) - \lambda_2(\Theta)| &\geq |\lambda_1(\Theta_0) - \lambda_2(\Theta_0)| - o(1) \\ &\geq c - o(1) \geq c/2\end{aligned}$$

if nN is sufficiently large, for some absolute constant $c > 0$.

(e) Since we assume $\sigma_K(A) \geq c\sqrt{K}$ for some $c \in (0, 1)$,

$$\max_{k\in[K]} V_1(k) \leq \|V_1\|_2 \leq \sigma_1(V) = \sigma_K^{-1}(A_{J*}) \leq \sigma_K^{-1}(A) \leq \frac{C}{\sqrt{K}}$$

and since

$$\|V_1\|_2 \geq \sigma_K(V) = \sigma_1^{-1}(A_{J*}) \geq \sigma_1^{-1}(A) \geq \frac{1}{\sqrt{K}}$$

and $\frac{1}{\|V_1\|_2}V_1$ is the unit-norm leading positive eigenvector of Θ, on event $\mathcal{E}$ we have

$$\min_{k\in[K]} V_1(k) = \|V_1\|_2 \min_{k\in[K]} \left\{\frac{V_1(k)}{\|V_1\|_2}\right\} \geq \frac{c}{\sqrt{K}}$$

Since $\xi_1 = A_{J*}V_1$, it follows that on event $\mathcal{E}$, for any $j \in J$,

$$\frac{ch_j}{\sqrt{K}} \leq \xi_1(j) \leq \frac{Ch_j}{\sqrt{K}}$$

(f) It can be seen by the definition of Q that

$$\sigma_K(Q) \geq \frac{\sigma_K(V)}{\max_{k\in[K]} V_1(k)} \geq c > 0$$

and

$$\sigma_1(Q) \leq \frac{\sigma_1(V)}{\min_{k\in[K]} V_1(k)} \leq C$$

for some $c, C > 0$. Thus, $\max_{k\in[K]} \|v_k^*\|_2 \leq C$ and Q has independent columns, which implies $v_1^*, \ldots, v_K^*$ are affinely independent.

□

Lemma 13. *Let $A_{J1}, \ldots, A_{JK}$ be the columns of A_{J*}. Under Assumption 4 on the vertex hunting function $\mathcal{V}(\cdot)$, the oracle procedure in Definition (4) returns*

$$\tilde{A}_{J*} = A_{J*} \cdot diag(\|A_{J1}\|_1^{-1}, \ldots, \|A_{JK}\|_1^{-1}) \tag{1.41}$$

on event $\mathcal{E}$.

Proof. Note that from Lemma 12(b)(i), we have $\Xi = A_{J*}V$. Let $\mathbf{1}_J$ be the vector of size $|J|$ with entries all equal to 1. Now, by the definition of R,

$$[\mathbf{1}_J, R] = [\mathrm{diag}(\xi_1)]^{-1}\Xi = [\mathrm{diag}(\xi_1)]^{-1}A_{J*}V$$

Recall from Lemma 12(f) the definition $Q^T := [\mathrm{diag}(V_1)]^{-1}V = \begin{pmatrix} 1 & \ldots & 1 \\ v_1^* & \ldots & v_K^* \end{pmatrix}^T$. Then

$$[\mathbf{1}_J, R] = [\mathrm{diag}(\xi_1)]^{-1}A_{J*} \cdot \mathrm{diag}(V_1)Q^T = \Pi \begin{pmatrix} 1 & \ldots & 1 \\ v_1^* & \ldots & v_K^* \end{pmatrix}^T \tag{1.42}$$

where Π is defined as follows:

$$\Pi := [\mathrm{diag}(\xi_1)]^{-1}A_{J*} \cdot \mathrm{diag}(V_1) \in \mathbb{R}^{|J|\times K} \tag{1.43}$$

From (1.42) and (1.43), we can see that Π contains only non-negative entries and the rows of Π sum up to 1. This means the rows of R (the point cloud) lie inside the convex hull of simplex vertices $\{v_1^*, \ldots, v_K^*\} \subseteq \mathbb{R}^{K-1}$.

By Assumption 3, for each topic there exists at least an anchor word for that topic in the set J on event $\mathcal{E}$. This means that the point cloud contains at least one point on each vertex $v_1^*, \ldots, v_K^*$. By Assumption 4, the vertex hunting procedure $\mathcal{V}(\cdot)$ returns precisely the vertices $v_1^*, \ldots, v_K^*$. Now let $\{\pi_j : j \in J\} \subseteq R^K$ denote the rows of Π. From taking the transpose of (1.42), Π is then estimated correctly by solving

$$\begin{pmatrix} 1 & \ldots & 1 \\ v_1^* & \ldots & v_K^* \end{pmatrix} \pi_j = \begin{pmatrix} 1 \\ r_j \end{pmatrix}$$

Now, by the definition of Π in (1.43),

$$\operatorname{diag}(\xi_1) \cdot \Pi = A_{J*} \cdot \operatorname{diag}(V_1) \tag{1.44}$$

and thus (1.41) follows if we normalize $\operatorname{diag}(\xi_1) \cdot \Pi$ to ensure its columns sum up to 1. □

1.5.3 Concentration inequalities involving $Z = D - D_0$

Remark 10. This section contains all the concentration inequalities necessary for our analysis, and is comparable to Section E in the appendix of Ke and Wang [2022].

Lemma 15 and Lemma 16 are similar to Lemmas E.1 and E.2 of Ke and Wang [2022] in that they are simple applications of Bernstein's inequality. However, it is crucial to note that our results are applicable even when $\min_{j\in[p]} h_j$ is extremely small, as we only restrict our attention to $j \in J_+$ (where J_+ is defined in Section A). In contrast, Lemmas E.1 and E.2 of Ke and Wang [2022] require $\min_{j\in[p]} h_j \geq cK/p$ (or at least $\min_{j\in[p]} h_j \gg (Nn)^{-1} \log n$).

Lemma 17 in our paper is based on standard techniques for deriving concentration inequalities for U-statistics. Our results here can be compared to Lemmas E.3-E.6 of Ke and Wang [2022], which use a truncation argument and the fact that the product of two sub-Gaussian variables is sub-exponential. Our bounds do not depend on p except for log factors

and are applicable to all parameter regimes (in particular when $p \gg n \vee N$), whereas the bounds in Lemmas E.3-E.6 Ke and Wang [2022] depend heavily on p and $\min_{j\in[p]} h_j$.

Lemma 14 (Bernstein's inequality)**.** *Let $X_1, \ldots, X_n$ be independent random variables with $\mathbb{E}(X_i) = 0$ and $Var(X_i) \le \sigma_i^2$ for all i. Let $\sigma^2 := n^{-1}\sum_{i=1}^n \sigma_i^2$. Then for any $t > 0$,*

$$\mathbb{P}\left(n^{-1}\left|\sum_{i=1}^n X_i\right| \ge t\right) \le 2\exp\left(-\frac{nt^2/2}{\sigma^2 + bt/3}\right)$$

Lemma 15. *Denote $\tilde{h}_j := h_j \wedge 1$. With probability at least $1 - o(p_n^{-1})$,*

$$|M(j,j) - M_0(j,j)| \le C^*\sqrt{\frac{\tilde{h}_j \log p_n}{nN}} \quad \text{for all } j \in J_+ \tag{1.45}$$

Proof. Similar to (1.40), for a fixed $j \in J_+$ we have

$$M(j,j) - M_0(j,j) = \frac{1}{n}\sum_{i=1}^n Z_{ji} = \frac{1}{nN}\sum_{i=1}^n\sum_{m=1}^N (T_{im}(j) - \mathbb{E}[T_{im}(j)])$$

Note that since $T_{im}(j) \sim \text{Bernoulli}(D_0(j,i))$, $|T_{im}(j) - \mathbb{E}[T_{im}(j)]| \le 1$ and

$$\text{Var}(T_{im}(j)) \le D_0(j,i) = \sum_{k=1}^K A_{jk}W_{ki} \le \sum_{k=1}^K A_{jk} = h_j \tag{1.46}$$

(and also $\text{Var}(T_{im}(j)) \le 1$). We apply Bernstein's inequality to conclude for any $t > 0$:

$$\mathbb{P}\left(|M(j,j) - M_0(j,j)| \ge t\right) \le 2\exp\left(-\frac{nNt^2/2}{\tilde{h}_j + t/3}\right)$$

One can choose $t = C^*\sqrt{\frac{\tilde{h}_j \log p_n}{nN}}$ or $t = \frac{C^* \log p_n}{nN}$ depending on whether $\tilde{h}_j \ge \frac{\log p_n}{nN}$ holds.

Thus with probability at least $1 - o(p_n^{-2})$,

$$|M(j,j) - M_0(j,j)| \le C^* \max\left(\sqrt{\frac{\tilde{h}_j \log p_n}{nN}}, \frac{\log p_n}{nN}\right)$$
$$\le C^* \sqrt{\frac{\tilde{h}_j \log p_n}{nN}}$$

since if $j \in J_+$, then $\tilde{h}_j > \alpha_+ \alpha \sqrt{\frac{\log p_n}{nN}} \ge \frac{c^* \log p_n}{nN}$ when nN is sufficiently large so that $\frac{\log p_n}{nN} \le 1$. We then take union bound over $j \in J_+$. □

Lemma 16. *Denote $\{Z_j : j \in J_+\} \subseteq \mathbb{R}^n$ as the rows of Z in J_+, and $\{W_k : k \in [K]\} \subseteq \mathbb{R}^n$ as the rows of W. With probability at least $1 - o(p_n^{-1})$,*

$$|Z_j^T W_k| \le C^* \sqrt{\frac{n\tilde{h}_j \log p_n}{N}} \quad \textit{for all } j \in J_+ \textit{ and } k \in [K] \tag{1.47}$$

Proof. Note that

$$Z_{ji} = \frac{1}{N} \sum_{m=1}^{N} (T_{im}(j) - \mathbb{E}[T_{im}(j)]) \tag{1.48}$$

and so for any $j \in J_+$ and $k \in [K]$,

$$Z_j^T W_k = \sum_{i=1}^{n} Z_{ji} W_{ki} = \frac{1}{nN} \sum_{i=1}^{n} \sum_{m=1}^{N} nW_{ki}(T_{im}(j) - \mathbb{E}[T_{im}(j)])$$

We note that $|nW_{ki}(T_{im}(j) - \mathbb{E}[T_{im}(j)])| \le n$ and $\mathrm{Var}[nW_{ki}(T_{im}(j) - \mathbb{E}[T_{im}(j)])] \le n^2 \tilde{h}_j$, so by Bernstein inequality, for any $t > 0$,

$$\mathbb{P}(|Z_j^T W_k| > t) \le 2 \exp\left(-\frac{nNt^2/2}{n^2 \tilde{h}_j + nt/3}\right)$$

We can let $t = C^* \sqrt{\frac{n\tilde{h}_j \log p_n}{N}}$ and again by noting that $\tilde{h}_j \ge \alpha_+ \alpha \sqrt{\frac{\log p_n}{nN}} \ge c^* \frac{\log p_n}{nN}$ if

$j \in J_+$, we obtain (1.47).

□

Lemma 17. *With probability at least* $1 - o(p_n^{-1})$,

$$|Z_j^T Z_l - \mathbb{E}(Z_j^T Z_l)| \leq C^* \sqrt{\frac{n\tilde{h}_j \tilde{h}_l \log p_n}{N}} \quad \textit{for all } j, l \in J_+ \textit{ with } j \neq l \tag{1.49}$$

$$|Z_j^T Z_j - \mathbb{E}(Z_j^T Z_j)| \leq C^* \sqrt{\frac{n\tilde{h}_j^2 \log p_n}{N}} + \frac{C^*}{N} \sqrt{\frac{n\tilde{h}_j \log p_n}{N}} \textit{ for all } j \in J_+ \tag{1.50}$$

Proof. Denote $X_{im}(j) := T_{im}(j) - \mathbb{E}[T_{im}(j)]$. Fix $j, l \in J_+$. By (1.48), note that

$$\begin{aligned} Z_j^T Z_l &= \sum_{i=1}^n Z_{ji} Z_{li} = \frac{1}{N^2} \sum_{i=1}^n \sum_{m=1}^N \sum_{s=1}^N X_{im}(j) X_{is}(l) \\ &= \frac{1}{N^2} \sum_{i=1}^n \sum_{m=1}^N X_{im}(j) X_{im}(l) + \frac{1}{N^2} \sum_{i=1}^n \sum_{\substack{1 \leq m,s \leq N \\ m \neq s}} X_{im}(j) X_{is}(l) \\ &= \frac{n}{N} V_1 + \frac{N-1}{N} V_2 \end{aligned}$$

where we define

$$V_1 := \frac{1}{nN} \sum_{i=1}^n \sum_{m=1}^N X_{im}(j) X_{im}(l)$$

$$V_2 := \frac{1}{N(N-1)} \sum_{i=1}^n \sum_{\substack{1 \leq m,s \leq N \\ m \neq s}} X_{im}(j) X_{is}(l)$$

Note that $\mathbb{E}(V_2) = 0$, and we need an upper bound on $|V_1 - \mathbb{E}(V_1)|$ and $|V_2|$. We will deal with V_2 first. Define S_N as the set of permutations on $\{1, \ldots, N\}$ and $N' := \lfloor N/2 \rfloor$. Also define

$$W_i(X_{i1}, \ldots, X_{iN}) := \frac{1}{N'} \sum_{m=1}^{N'} X_{i,2m-1}(j) X_{i,2m}(l)$$

Then by symmetry (note that the inner sum over m, s in the definition of V_2 has $N(N-1)$

summands),

$$V_2 = \frac{\sum_{i=1}^{n} \sum_{\pi \in S_N} W_i(X_{i,\pi(1)}, \ldots, X_{i,\pi(N)})}{N!}$$

Define, for a given $\pi \in S_N$,

$$Q_\pi := \sum_{i=1}^{n} N' W_i(X_{\pi(1)}, \ldots, X_{\pi(N)})$$

so that $N'V_2 = \frac{1}{N!} \sum_{\pi \in S_N} Q_\pi$. For arbitrary $t, s > 0$, by Markov's inequality and the convexity of the exponential function,

$$\mathbb{P}(N'V_2 \geq t) \leq e^{-st} \mathbb{E}(e^{sN'V_2}) \leq e^{-st} \frac{\sum_{\pi \in S_N} \mathbb{E}(e^{sQ_\pi})}{N!}$$

Also, define $Q = Q_\pi$ for π being the identity permutation. Observe that

$$Q = \sum_{i=1}^{n} \sum_{m=1}^{N'} Q_{im} \quad \text{where } Q_{im} = X_{i,2m-1}(j) X_{i,2m}(l)$$

so Q is a (double) summation of mutually independent variables. We have $|Q_{im}| \leq 1$, $\mathbb{E}(Q_{im}) = 0$ and $\mathbb{E}(Q_{im}^2) \leq \tilde{h}_j \tilde{h}_l$. The rest of the proof for V_2 is similar to the standard proof for the usual Bernstein's inequality and one can skip to the conclusion (1.51).

If we denote $G(x) = \frac{e^x - 1 - x}{x^2}$, observe $G(x)$ is increasing. Hence,

$$\begin{aligned}
\mathbb{E}(e^{sQ_{im}}) &= \mathbb{E}\left(1 + sQ_{im} + \frac{s^2 Q_{im}^2}{2} + \ldots\right) \\
&= \mathbb{E}[1 + s^2 Q_{im}^2 G(sQ_{im})] \\
&\leq \mathbb{E}[1 + s^2 Q_{im}^2 G(s)] \\
&\leq 1 + s^2 \tilde{h}_j \tilde{h}_l G(s) \leq e^{s^2 \tilde{h}_j \tilde{h}_l G(s)}
\end{aligned}$$

Hence,

$$e^{-st}\mathbb{E}(e^{sQ}) = \exp(-st + N'n\tilde{h}_j\tilde{h}_l s^2 G(s))$$

Since this bound is applicable to all the other Q_π and not just π being equal to the identity permutation, we have

$$\mathbb{P}(N'V_2 \geq t) \leq \exp(-st + N'n\tilde{h}_j\tilde{h}_l s^2 G(s)) = \exp\left(-st + N'n\tilde{h}_j\tilde{h}_l(e^s - 1 - s)\right)$$

Now we choose $s = \log\left(1 + \frac{t}{N'n\tilde{h}_j\tilde{h}_l}\right) > 0$. Then

$$\begin{aligned}\mathbb{P}(N'V_2 \geq t) &\leq \exp\left[-t\log\left(1 + \frac{t}{N'n\tilde{h}_j\tilde{h}_l}\right) + N'n\tilde{h}_j\tilde{h}_l\left(\frac{t}{N'n\tilde{h}_j\tilde{h}_l} - \log\left(1 + \frac{t}{N'n\tilde{h}_j\tilde{h}_l}\right)\right)\right] \\ &= \exp\left[-N'n\tilde{h}_j\tilde{h}_l H\left(\frac{t}{N'n\tilde{h}_j\tilde{h}_l}\right)\right]\end{aligned}$$

where we define the function $H(x) = (1+x)\log(1+x) - x$. Note that we have the inequality

$$H(x) \geq \frac{3x^2}{6+2x}$$

for all $x > 0$. Hence,

$$\mathbb{P}\left(N'V_2 \geq t\right) \leq \exp\left(-\frac{t^2/2}{N'n\tilde{h}_j\tilde{h}_l + t/3}\right)$$

or by rescaling,

$$\mathbb{P}(N'V_2 \geq N'nt) \leq \exp\left(-\frac{N'nt^2/2}{\tilde{h}_j\tilde{h}_l + t/3}\right) \tag{1.51}$$

We can choose $t^2 = \frac{C^*\tilde{h}_j\tilde{h}_l}{N'n}\log p_n$ and note that $\tilde{h}_j\tilde{h}_l \geq (\alpha_+\alpha)^2\frac{\log p_n}{nN}$ if $j, l \in J_+$. Hence, with probability $1 - o(p_n^{-1})$ (even after taking union bound over $j, l \in J_+$),

$$|V_2| \leq C^*\sqrt{\frac{n\tilde{h}_j\tilde{h}_l\log p_n}{N}}$$

As for V_1, we can just apply the usual Bernstein's inequality. Let $\mu_{ij} = \mathbb{E}[T_{im}(j)] = [D_0]_{ji}$ and define μ_{il} similarly; note $\mu_{ij} \leq \tilde{h}_j$. Since $X_{im}(j) = T_{im}(j) - \mu_{ij}$,

$$X_{im}(j)X_{im}(l) = T_{im}(j)T_{im}(l) - \mu_{ij}T_{im}(l) - \mu_{il}T_{im}(j) + \mu_{ij}\mu_{il} \tag{1.52}$$

If $j \neq l$ then $T_{im}(j)T_{im}(l) = 0$ and so

$$\begin{aligned}
\mathrm{Var}[X_{im}(j)X_{im}(l)] &= \mathrm{Var}\left[\mu_{ij}T_{im}(l) + \mu_{il}T_{im}(j)\right] \\
&\leq \mathbb{E}[\mu_{ij}T_{im}(l) + \mu_{il}T_{im}(j)]^2 \\
&= \mu_{ij}^2\mu_{il} + \mu_{il}^2\mu_{ij} = \mu_{ij}\mu_{il}(\mu_{ij} + \mu_{il}) \\
&\leq \mu_{ij}\mu_{il} \leq \tilde{h}_j\tilde{h}_l
\end{aligned}$$

since $\mu_{ij} + \mu_{il} \leq 1$. Hence, by Bernstein's inequality,

$$\mathbb{P}\left(|V_1 - \mathbb{E}(V_1)| \geq t\right) \leq 2\exp\left(-\frac{-nNt^2/2}{\tilde{h}_j\tilde{h}_l + t/3}\right)$$

which is similar to (1.51), so we obtain with probability $1 - o(p_n^{-1})$ that

$$\frac{n}{N}|V_1 - \mathbb{E}(V_1)| \leq \frac{C^*}{N}\sqrt{\frac{n\tilde{h}_j\tilde{h}_l\log p_n}{N}} \leq C^*\sqrt{\frac{n\tilde{h}_j\tilde{h}_l\log p_n}{N}}$$

and (1.49) is proven.

If $j = l$ then since $T_{im}^2(j) = T_{im}(j)$, (1.52) leads to

$$X_{im}^2(j) = T_{im}(j)(1 - 2\mu_{ij}) + \mu_{ij}^2$$

and since $|1-2\mu_{ij}| \leq 1$ and $\mathrm{Var}(T_{im}(j)) = \mu_{ij}(1-\mu_{ij})$,

$$\mathrm{Var}[X_{im}^2(j)] \leq \mu_{ij} \leq \tilde{h}_j$$

and so we obtain (1.50) since with probability $1 - o(p_n^{-1})$

$$\frac{n}{N}|V_1 - \mathbb{E}(V_1)| \leq \frac{C^*}{N}\sqrt{\frac{n\tilde{h}_j \log p_n}{N}}$$

□

Corollary 18. *With probability $1 - o(p_n^{-1})$, the following statements hold:*

$$\|[ZW_k]_{J_+}\|_2 \leq C^*\sqrt{\frac{nK\log p_n}{N}} \text{ for all } 1 \leq k \leq K \tag{1.53}$$

$$\|[ZZ^T - \mathbb{E}(ZZ^T)]_{jJ_+}\|_2 \leq C^*\sqrt{\frac{n\tilde{h}_j K\log p_n}{N}} \text{ for all } j \in J_+ \tag{1.54}$$

$$\|[ZZ^T - \mathbb{E}(ZZ^T)]_{J_+J_+}\|_F \leq C^* K\sqrt{\frac{n\log p_n}{N}} \tag{1.55}$$

Proof. This follows from (1.47), (1.49) and (1.50) after squaring the error bounds and summing them up. We note that $\sum_{j\in J_+}\tilde{h}_j \leq \sum_{j=1}^p h_j = K$. □

1.5.4 Estimation errors for singular vectors and the point cloud

We will use the following theorem (a row-wise perturbation bound for eigenvectors) from Ke and Wang [2022].

Lemma 19 (Lemma F.1 of Ke and Wang [2022])**.** *Let B_0 and B be $m \times m$ symmetric matrices with rank$(B_0) = K$, and assume B_0 is positive semi-definite. For $1 \leq k \leq K$, let δ_k^0 and δ_k be the k^{th} largest eigenvalues of B_0 and B respectively, and let u_k^0 and u_k be the k^{th} eigenvectors of B_0 and B. Fix $1 \leq s \leq k \leq K$. If for some $c \in (0,1)$, suppose (by*

default, if $s = 1$ *then* $\delta^0_{s-1} - \delta^0_s = \infty$*)*

$$\min(\delta^0_{s-1} - \delta^0_s, \delta^0_k - \delta^0_{k+1}, \min_{l\in[K]} \delta^0_l) \geq c\|B_0\|_{op}, \quad \|B - B_0\|_{op} \leq (c/3)\|B_0\|_{op}$$

Write $U_0 = [u^0_s, \ldots, u^0_k], U = [u_s, \ldots u_k]$ *and* $\Xi = [u^0_1, \ldots, u^0_K]$*. There exists an orthonormal matrix* O *such that for all* $1 \leq j \leq p$*,*

$$\|(UO - U_0)_{j*}\|_2 \leq \frac{5}{c\|B_0\|_{op}}(\|B - B_0\|_{op}\|\Xi_{j*}\|_2 + \sqrt{K}\|(B - B_0)_{j*}\|_2)$$

If we define

$$G := DD^T - \frac{n}{N}M$$

$$G_0 := \left(1 - \frac{1}{N}\right) D_0 D_0^T$$

then the above lemma can be applied to the submatrices G_{JJ} and $[G_0]_{JJ}$ (see Lemma 25).

Lemma 20. *With probability* $1 - o(p_n^{-1})$*, we have*

$$\|(G - G_0)_{J_+J_+}\|_{op} \leq C^* K\sqrt{\frac{nK\log p_n}{N}}$$

and for any $j \in J_+$*, row* j *of* $(G - G_0)_{J_+J_+}$ *has* ℓ_2 *norm satisfying*

$$\|(G - G_0)_{jJ_+}\|_2 \leq C^* K\sqrt{\frac{nh_j\log p_n}{N}}$$

Proof. From basic properties of the multinomial distribution, we can show that

$$\mathbb{E}(ZZ^T) = \sum_{i=1}^n \mathrm{Cov}(Z_{*i}) = \sum_{i=1}^n \mathrm{Cov}(D_{*i}) = \frac{n}{N}M_0 - \frac{1}{N}D_0D_0^T$$

and therefore

$$\begin{aligned} G - G_0 &= DD^T - \frac{n}{N}M - \left(1 - \frac{1}{N}\right) D_0 D_0^T \\ &= (D_0 + Z)(D_0 + Z)^T - \frac{n}{N}M - \left(1 - \frac{1}{N}\right) D_0 D_0^T \\ &= ZD_0^T + D_0 Z^T + ZZ^T - \frac{n}{N}M + \frac{1}{N} D_0 D_0^T \\ &= ZD_0^T + D_0 Z^T + (ZZ^T - \mathbb{E}[ZZ^T]) + \frac{n}{N}(M_0 - M) \end{aligned}$$

and so we can write $(G - G_0)_{J_+ J_+} = E_1 + E_2 + E_3$ where

$$E_1 := (ZD_0^T + D_0 Z^T)_{J_+ J_+}$$

$$E_2 := (ZZ^T - \mathbb{E}[ZZ^T])_{J_+ J_+}$$

$$E_3 := \frac{n}{N}(M_0 - M)_{J_+ J_+}$$

We can deal with E_3 first. From (1.45), with probability $1 - o(p_n^{-1})$ we have

$$\|E_3\|_{\text{op}} \leq \frac{C^* n}{N} \sqrt{\frac{(\max_{j \in J_+} \tilde{h}_j) \log p_n}{nN}} \leq \frac{C^*}{N} \sqrt{\frac{n \log p_n}{N}}$$

and for any $j \in J_+$,

$$\|[E_3]_{j*}\|_2 = \frac{n}{N} |M(j,j) - M_0(j,j)| \leq \frac{C^*}{N} \sqrt{\frac{n \tilde{h}_j \log p_n}{N}}$$

From (1.54) and (1.55), with probability $1 - o(p_n^{-1})$

$$\|E_2\|_{\text{op}} \leq \|E_2\|_F \leq C^* K \sqrt{\frac{n \log p_n}{N}}$$

$$\|[E_2]_{j*}\|_2 \leq C^* \sqrt{\frac{n\tilde{h}_j K \log p_n}{N}}$$

If we denote $A_1, \ldots, A_K$ as the columns of A and $W_1, \ldots, W_k$ as the rows of W,

$$D_0 Z^T = \sum_{k=1}^{K} A_k (ZW_k)^T$$

and so from (1.53) and the fact that $\sum_{k=1}^{K} \|A_k\|_2 \leq \sum_{k=1}^{K} \|A_k\|_1 \leq K$,

$$\|E_1\|_{\text{op}} \leq 2\|[D_0 Z^T]_{J_+ J_+}\|_{\text{op}} \leq 2 \sum_{k=1}^{K} \|A_k\|_2 \|Z_{J_+ *} W_k\|_2 \leq C^* K \sqrt{\frac{nK \log p_n}{N}}$$

Let $Z_1, \ldots, Z_p$ denote the rows of Z. From (1.47), (1.53) and the fact that $\sum_{k=1}^{K} A_k(j) = h_j$ and $h_j \leq K$, for any $j \in J_+$:

$$\begin{aligned}
\|[E_1]_{j*}\|_2 &\leq \sum_{k=1}^{K} A_k(j) \|Z_{J_+ *} W_k\|_2 + \sum_{k=1}^{K} |Z_j^T W_k| \|A_k\|_2 \\
&\leq C^* h_j \sqrt{\frac{nK \log p_n}{N}} + C^* K \sqrt{\frac{n\tilde{h}_j \log p_n}{N}} \\
&\leq C^* K \sqrt{\frac{nh_j \log p_n}{N}}
\end{aligned}$$

Since the bounds for $\|E_1\|_{\text{op}}$ and $\|[E_1]_{j*}\|_2$ dominate those for E_2 and E_3, our result follows.

□

Lemma 21. *With probability $1 - o(p_n^{-1})$, we also have*

$$\|(G - G_0)_{JJ}\|_{op} \leq C^* K \sqrt{\frac{nK \log p_n}{N}} \tag{1.56}$$

and for any $j \in J$,

$$\|(G - G_0)_{jJ}\|_2 \leq C^* K \sqrt{\frac{nh_j \log p_n}{N}} \tag{1.57}$$

Proof. This is simply a consequence of the previous lemma and the fact that $J \subseteq J_+$ with probability $1-o(p_n^{-1})$, which implies that $(G-G_0)_{JJ}$ is a submatrix of $(G-G_0)_{J_+J_+}$. Note that we refrain from applying the argument of the previous lemma directly to $(G-G_0)_{JJ}$, since J and Z are not independent (whereas J_+ is a non-random index set). □

Corollary 22. *Let g_n be a quantity satisfying*

$$c\sqrt{\frac{nN}{\log p_n}} \geq g_n \geq C^* K\sqrt{K} \tag{1.58}$$

where C^* in (1.58) *is the constant from* (1.56) *and c is another constant to be determined. If*

$$\hat{K} := \max\left\{k : \lambda_k(G_{JJ}) > g_n\sqrt{\frac{n\log p_n}{N}}\right\}$$

then $\hat{K} = K$ with probability $1-o(p_n^{-1})$.

Proof. We have shown in Lemma A.3(e) that $[G_0]_{JJ}$ has rank K on $\mathcal{E}$. By Weyl's inequality,

$$\lambda_{K+1}(G_{JJ}) \leq \|(G-G_0)_{JJ}\|_{\text{op}} \leq C^*K\sqrt{K}\sqrt{\frac{n\log p_n}{N}} \leq g_n\sqrt{\frac{n\log p_n}{N}}$$

This implies $\hat{K} \leq K$. On the other hand, again by Weyl's inequality,

$$|\lambda_K(G_{JJ}) - \lambda_K([G_0]_{JJ})| \leq C^*K\sqrt{K}\sqrt{\frac{n\log p_n}{N}} \leq g_n\sqrt{\frac{n\log p_n}{N}}$$

and since $G_0 := \left(1-\frac{1}{N}\right)D_0D_0^T$, by our assumption that $\sigma_K(A) \geq c\sqrt{K}$ and $\sigma_K(\Sigma_W) \geq c$,

$$\lambda_K([G_0]_{JJ}) \geq \left(1-\frac{1}{N}\right)\sigma_K^2(A_{J*})\sigma_K^2(W) \geq cKn > 2g_n\sqrt{\frac{n\log p_n}{N}}$$

when nN is sufficiently large and c in (1.58) is chosen appropriately. Hence,

$$\lambda_K(G_{JJ}) \geq \lambda_K([G_0]_{JJ}) - |\lambda_K(G_{JJ}) - \lambda_K([G_0]_{JJ})| > g_n\sqrt{\frac{n\log p_n}{N}}$$

and thus $K \leq \hat{K}$ with probability $1 - o(p_n^{-1})$. □

Recall that $\hat{\Xi}$ contains the first K eigenvectors of G_{JJ} and Ξ contains the first K left singular vectors of $[D_0]_{J*}$, or equivalently the first K eigenvectors of $[G_0]_{JJ}$. We will provide a coordinate-wise error bound for $\hat{\Xi}$ in Lemma 25. First we need a few lemmas.

Lemma 23. *For any $j \in J$, $\|\Xi_{j*}\|_2 \leq Ch_j$.*

Proof. Note $\Xi = A_{J*}V$ so $\Xi_{j*} = A_{j*}V$. Hence,

$$\|\Xi_{j*}\|_2 \leq \|V\|_{\text{op}}\|A_{j*}\|_2 \leq \|V\|_{\text{op}}\|A_{j*}\|_1 \leq \sigma_K^{-1}(A)h_j \leq Ch_j$$

since we have shown before that the singular values of V are just the inverses of the singular values of A_{J*}. □

Note that on event $\mathcal{E}$, $[D_0]_{J*}$ and hence $[G_0]_{JJ}$ has rank K.

Lemma 24. *On event $\mathcal{E}$,*

$$cnK \leq \lambda_k([G_0]_{JJ}) \leq nK \text{ for all } k \in [K] \text{ and } \lambda_1([G_0]_{JJ}) \geq cn + \max_{2\leq k\leq K} \lambda_k([G_0]_{JJ})$$

Proof. We note that $[D_0D_0^T]_{JJ} = A_{J*}WW^TA_{J*}^T = nA_{J*}\Sigma_W A_{J*}^T$. Hence,

$$\lambda_1([G_0]_{JJ}) \leq n\|A\|_{\text{op}}^2\|\Sigma_W\|_{\text{op}} \leq nK$$

$$\lambda_K([G_0]_{JJ}) \geq n[\sigma_K(A_{J*})]^2\sigma_K(\Sigma_W) \geq n[\sigma_K(A)]^2\sigma_K(\Sigma_W) \geq cnK$$

We also note that for any two matrices P and Q, the nonzero eigenvalues of PQ are the same as those of QP. Thus the nonzero eigenvalues of $[D_0D_0^T]_{JJ}$ are the same as the nonzero eigenvalues of $WW^TA_{J*}^TA_{J*} =: n\Theta$. We have already shown in Lemma 12(d) that the gap between the first two eigenvalues of Θ are at least an absolute constant on $\mathcal{E}$. Hence, our result follows.

□

Lemma 25 (Row-wise estimation error for $\hat{\Xi}$)**.** *Denote* $\{\Xi_j : j \in J\}$ *as the rows of* Ξ *and* $\{\hat{\Xi}_j : j \in J\}$ *as the rows of* $\hat{\Xi}$*. With probability* $1 - o(p_n^{-1})$*, there exist* $\omega \in \{\pm 1\}$ *and an orthonormal matrix* $\Omega^* \in \mathbb{R}^{(K-1)\times(K-1)}$ *such that, if* $\Omega := diag(\omega, \Omega^*) \in \mathbb{R}^{K\times K}$*, we have*

$$\|\Omega\hat{\Xi}_j - \Xi_j\|_2 \le C\sqrt{\frac{h_j \log p_n}{nN}} \quad \textit{for all } j \in J$$

Proof. Let $\hat{\xi}_1$ and ξ_1 be the first eigenvectors of $[G]_{JJ}$ and $[G_0]_{JJ}$ respectively. The gap between the first two eigenvalues of $[G_0]_{JJ}$ is at least cn, which is much greater than $C^*K\sqrt{\frac{nK\log p_n}{N}}$ (the high-probability bound on $\|(G - G_0)_{JJ}\|_{\text{op}}$). By applying Lemma 19, there exists $\omega \in \{\pm 1\}$ such that with probability $1 - o(p_n^{-1})$, for all $j \in J$,

$$\begin{aligned}
|\omega\hat{\xi}_1(j) - \xi_1(j)| &\le C\frac{h_j\|(G - G_0)_{JJ}\|_{\text{op}} + \sqrt{K}\|(G - G_0)_{jJ}\|_2}{n} \\
&\le C\frac{h_j\sqrt{\frac{n\log p_n}{N}} + \sqrt{\frac{nh_j\log p_n}{N}}}{n} \\
&\le C\sqrt{\frac{h_j\log p_n}{nN}}
\end{aligned}$$

where we applied $h_j \le K$.

Let $\Xi^* = [\xi_2, \ldots, \xi_K]$ contain the other $(K-1)$ eigenvectors of $[G_0]_{JJ}$, and define $\hat{\Xi}^*$ similarly. Again, since the smallest nonzero eigenvalue of $[G_0]_{JJ}$ is at least cnK, there exists an orthonormal matrix $\Omega^* \in \mathbb{R}^{(K-1)\times(K-1)}$ such that for all $j \in J$,

$$\|(\hat{\Xi}^*\Omega^* - \Xi^*)_{j*}\|_2 \le C\sqrt{\frac{h_j\log p_n}{nN}}$$

We then define $\Omega = \text{diag}(\omega, \Omega^*)$ and combine the above results. □

Lemma 26 (Estimation error for the point cloud)**.** *With probability* $1 - o(p_n^{-1})$*, all entries of* $\hat{\xi}_1$ *have the same sign and there exists an orthonormal matrix* $\Omega^* \in \mathbb{R}^{(K-1)\times(K-1)}$ *such*

that

$$\max_{j\in J} \|\Omega^* \hat{r}_j - r_j\|_2 \leq C \left(\frac{\log p_n}{nN}\right)^{1/4}$$

Proof. First, we note that WLOG, we can assume $\omega = 1$. This is because from the previous lemma, for any $j \in J$, since $h_j \geq c\sqrt{\frac{\log p_n}{nN}}$,

$$|\omega\hat{\xi}_1(j) - \xi_1(j)| \leq C\sqrt{\frac{h_j \log p_n}{nN}} \leq Ch_j \left(\frac{\log p_n}{nN}\right)^{1/4}$$

whereas we also know from Lemma 12(e) that

$$\xi_1(j) > ch_j > 0$$

We can see that $|\omega\hat{\xi}_1(j) - \xi_1(j)| \ll \xi_1(j)$ with high probability as nN is sufficiently large, and this implies $\omega\hat{\xi}_1(j) \geq \xi_1(j)/2$. If $\hat{\xi}_1$ is defined such that the majority of its entries are positive (and in fact its entries are all of the same sign with high probability), we can simply assume $\omega = 1$ from now on.

Denote $\{\Xi_j : j \in J\}$ as the rows of Ξ and $\{\hat{\Xi}_j : j \in J\}$ as the rows of $\hat{\Xi}$. Now, since by definition,

$$\begin{pmatrix} 1 \\ r_j \end{pmatrix} = [\xi_1(j)]^{-1}\Xi_j, \quad \begin{pmatrix} 1 \\ \Omega^*\hat{r}_j \end{pmatrix} = [\hat{\xi}_1(j)]^{-1}\Omega\hat{\Xi}_j$$

it follows that

$$\begin{aligned} \|\Omega^*\hat{r}_j - r_j\|_2 &= \left\| \frac{1}{\hat{\xi}_1(j)}\Omega\hat{\Xi}_j - \frac{1}{\xi_1(j)}\Xi_j \right\|_2 \\ &= \left\| \frac{1}{\hat{\xi}_1(j)}(\Omega\hat{\Xi} - \Xi_j) - \frac{\hat{\xi}_1(j) - \xi_1(j)}{\hat{\xi}_1(j)} r_j \right\|_2 \\ &\leq |\hat{\xi}_1(j)|^{-1}(\|\Omega\hat{\Xi}_j - \Xi_j\|_2 + \|r_j\|_2|\hat{\xi}_1(j) - \xi_1(j)|) \end{aligned}$$

We have noted in Lemma 13 that the point cloud $\{r_j : j \in J\}$ lies entirely in the convex hull of $v_1^*, \ldots, v_K^*$, and Lemma 12(f) shows that $\max_{k\in[K]} \|v_k^*\|_2 \le C$, so we also have $\max_{j\in J} \|r_j\|_2 \le C$. We have also noted before that

$$\hat{\xi}_1(j) \ge \frac{\xi_1(j)}{2} > ch_j$$

with high probability. Therefore, with probability $1 - o(p_n^{-1})$, for all $j \in J$:

$$\begin{aligned}\|\Omega^* \hat{r}_j - r_j\|_2 &\le C\sqrt{\frac{\log p_n}{h_j nN}} \\ &\le C\left(\frac{\log p_n}{nN}\right)^{1/4}\end{aligned}$$

since $\min_{j\in J} h_j \ge c\sqrt{\frac{\log p_n}{nN}}$ with probability $1 - o(p_n^{-1})$. □

Lemma 27. *If we denote the rows of $\hat{\Pi}$ from our proposed procedure as $\{\hat{\pi}_j : j \in J\}$ and the rows of Π from the oracle procedure as $\{\pi_j : j \in J\}$, then with probability $1 - o(p_n^{-1})$,*

$$\max_{j\in J} \|\hat{\pi}_j - \pi_j\|_1 \le C\left(\frac{\log p_n}{nN}\right)^{1/4}$$

Proof. Recall that $\hat{\pi}_j^\diamond \in \mathbb{R}^K$ is the unnormalized vector solving

$$\begin{pmatrix} 1 & \ldots & 1 \\ \hat{v}_1^* & \ldots & \hat{v}_K^* \end{pmatrix} \hat{\pi}_j^\diamond = \begin{pmatrix} 1 \\ \hat{r}_j \end{pmatrix} \iff \begin{pmatrix} 1 & \ldots & 1 \\ \Omega^*\hat{v}_1^* & \ldots & \Omega^*\hat{v}_K^* \end{pmatrix} \hat{\pi}_j^\diamond = \begin{pmatrix} 1 \\ \Omega^*\hat{r}_j \end{pmatrix}$$

Therefore,

$$\hat{\pi}_j^\diamond = \hat{Q}^{-1} \begin{pmatrix} 1 \\ \Omega^*\hat{r}_j \end{pmatrix} \quad \text{where} \quad \hat{Q} := \begin{pmatrix} 1 & \ldots & 1 \\ \Omega^*\hat{v}_1^* & \ldots & \Omega^*\hat{v}_K^* \end{pmatrix}$$

We also have

$$\pi_j = Q^{-1}\begin{pmatrix}1\\ r_j\end{pmatrix} \quad \text{where } Q = \begin{pmatrix}1 & \dots & 1\\ v_1^* & \dots & v_K^*\end{pmatrix}$$

Consequently,

$$\|\hat{\pi}_j^{\diamond} - \pi_j\|_2 \le \|\hat{Q}^{-1}\|_{\text{op}}\|\Omega^*\hat{r}_j - r_j\|_2 + \|\hat{Q}^{-1} - Q^{-1}\|_{\text{op}}\sqrt{\|r_j\|_2^2 + 1}$$

Note that $\max_{j\in J}\|r_j\|_2 \le C$ since the r_j's are in the convex hull of $v_1^*, \dots, v_K^*$. Also, since $Q^T = [\text{diag}(V_1)]^{-1}V$, we have $\|Q^{-1}\|_{\text{op}} \le C$ since

$$\max_{k\in[K]} V_1(k) \le \frac{C}{\sqrt{K}} \quad \text{and} \quad \|V^{-1}\|_{\text{op}} = \sigma_1(A) \le \sqrt{K}$$

Now, we note that with probability $1 - o(p_n^{-1})$,

$$\begin{aligned}\|\hat{Q} - Q\|_{\text{op}} \le \|\hat{Q} - Q\|_F &\le \sqrt{K}\max_{k\in[K]}\|\Omega^*\hat{v}_k^* - v_k^*\|_2 \\ &\le \sqrt{K}\max_{j\in J}\|\Omega^*\hat{r}_j - r_j\|_2 \le C\left(\frac{\log p_n}{nN}\right)^{1/4} = o(1)\end{aligned}$$

where we used Assumption 4, since in the oracle procedure the vertex hunting algorithm correctly returns $v_1^*, \dots, v_K^*$. Therefore,

$$\|\hat{Q}^{-1} - Q^{-1}\|_{\text{op}} = \|\hat{Q}^{-1}(Q - \hat{Q})Q^{-1}\|_{\text{op}} \le \|\hat{Q}^{-1}\|_{\text{op}}\|\hat{Q} - Q\|_{\text{op}}\|Q\|_{\text{op}}^{-1} \le C\left(\frac{\log p_n}{nN}\right)^{1/4}$$

Here we note $\sigma_K(\hat{Q}) \ge \sigma_K(Q) - \|\hat{Q} - Q\|_{\text{op}} \ge c - o(1) \ge c/2$ if nN is large enough, so $\|\hat{Q}^{-1}\|_{\text{op}} \le C$. Therefore, we obtain

$$\|\hat{\pi}_j^{\diamond} - \pi_j\|_2 \le C\left(\frac{\log p_n}{nN}\right)^{1/4}$$

Now if we define $\hat{\pi}_j = \frac{\tilde{\pi}_j^\diamond}{\|\tilde{\pi}_j^\diamond\|_1}$ where $\tilde{\pi}_j^\diamond(k) = \max(\hat{\pi}_j^\diamond(k), 0)$, then since $\|\hat{\pi}_j\|_1 = \|\pi_j\|_1 = 1$,

$$\begin{aligned}
\|\hat{\pi}_j - \pi_j\|_1 &\leq \|\hat{\pi}_j - \tilde{\pi}_j^\diamond\|_1 + \|\tilde{\pi}_j^\diamond - \pi_j\|_1 \\
&= |1 - \|\tilde{\pi}_j^\diamond\|_1| \|\hat{\pi}_j\|_1 + \|\tilde{\pi}_j^\diamond - \pi_j\|_1 \\
&= |\|\pi_j|_1 - \|\tilde{\pi}_j^\diamond\|_1| + \|\tilde{\pi}_j^\diamond - \pi_j\|_1 \\
&\leq 2\|\pi_j - \tilde{\pi}_j^\diamond\|_1 \\
&\leq 2\|\pi_j - \hat{\pi}_j^\diamond\|_1 \leq 2\sqrt{K}\|\hat{\pi}_j^\diamond - \pi_j\|_2 \\
&\leq C\left(\frac{\log p_n}{nN}\right)^{1/4}
\end{aligned}$$

□

1.5.5 Estimation error of $\hat{A}$

In this section, we will additionally impose the ℓ_q-sparsity assumption (1.19) for $q \in (0, 1)$.

Lemma 28. *Under Assumption 5, if $\beta := \frac{\alpha_- \alpha}{\sigma_K(\Sigma_W)}$ and $\tau_n := \sqrt{\frac{\log p_n}{nN}}$, on event $\mathcal{E}$*

$$\|A_{J^c *}\|_1 \leq \frac{K}{1-q} s(\beta\tau_n)^{1-q} \tag{1.59}$$

Remark 11. We assume from now on that s does not grow too quickly relative to nN so that the RHS of (1.59) is $o(1)$.

Proof. On event $\mathcal{E}$ we have $J_- \subseteq J$, so $j \notin J$ implies $M_0(j, j) \leq \alpha_- \alpha\tau_n$ where $\tau_n := \sqrt{\frac{\log p_n}{nN}}$. Since $\sigma_K(\Sigma_W)h_j \leq M_0(j, j)$, $j \notin J$ implies $A_{jk} \leq h_j \leq \beta\tau_n$ for any $k \in [K]$ on $\mathcal{E}$. Then

with probability $1 - o(p_n^{-1})$, for any $k \in [K]$,

$$\begin{aligned}\|A_{J^c k}\|_1 &= \sum_{j \notin J} \min(A_{jk}, \beta\tau_n) \leq \sum_{j=1}^{p} \min(A_{(j)k}, \beta\tau_n) \\ &\leq \sum_{j=1}^{p} \min(s^{1/q} j^{-1/q}, \beta\tau_n) \leq \int_0^{\infty} \min(s^{1/q} t^{-1/q}, \beta\tau_n) dt\end{aligned}$$

Now, let $t_0 := s(\beta\tau_n)^{-q}$ so that $s^{1/q} t_0^{-1/q} = \beta\tau_n$. Then continuing from the above display,

$$\begin{aligned}\|A_{J^c k}\|_1 &\leq t_0 \beta\tau_n + s^{1/q} \int_{t_0}^{\infty} t^{-1/q} dt \\ &= t_0 \beta\tau_n + \frac{q}{1-q} s^{1/q} t_0^{1-1/q} = \frac{1}{1-q} t_0 \beta\tau_n \\ &= \frac{1}{1-q} s(\beta\tau_n)^{1-q}\end{aligned}$$

and the result follows by summing up this bound across $k \in [K]$. Note that the assumption $q \in (0, 1)$ ensures the integrals above converge. □

Lemma 29. *On event $\mathcal{E}$, if $\tilde{A}_{J*}$ is defined as in* (1.41)*,*

$$\|\tilde{A}_{J*} - A_{J*}\|_1 = \|A_{J^c *}\|_1 \leq \frac{K}{1-q} s(\beta\tau_n)^{1-q} \tag{1.60}$$

Proof. We note that the columns of $\tilde{A}_{J*}$ sum up to 1, the columns of A sum up to 1, and as a result of the definition of $\tilde{A}_{J*}$ in (1.41),

$$\tilde{A}_{J*} - A_{J*} = \tilde{A}_{J*} \cdot \mathrm{diag}(\|A_{J^c 1}\|_1, \ldots, \|A_{J^c K}\|_1)$$

Then $\|\tilde{A}_{J*} - A_{J*}\|_1 = \|A_{J^c *}\|_1$ and our result follows from the previous lemma. □

Theorem 30. *With probability $1 - o(p_n^{-1})$, for some constant C that may depend on K and*

q, we have

$$\|\hat{A} - A\|_1 \leq C \left[\left(\frac{\log p_n}{nN} \right)^{1/4} + s \left(\frac{\log p_n}{nN} \right)^{\frac{1-q}{2}} \right]$$

Proof. Consider the unnormalized matrices

$$\hat{A}^{\diamond}_{J*} := \mathrm{diag}(\hat{\xi}_1)\hat{\Pi} \quad \text{and} \quad \tilde{A}^{\diamond}_{J*} := \mathrm{diag}(\xi_1)\Pi$$

Then with probability $1 - o(p_n^{-1})$, for any $j \in J$,

$$\begin{aligned}
\|(\hat{A}^{\diamond} - \tilde{A}^{\diamond})_{j*}\|_1 &= \|\hat{\xi}_1(j)\hat{\pi}_j - \xi_1(j)\pi_j\|_1 \\
&\leq |\hat{\xi}_1(j)|\|\hat{\pi}_j - \pi_j\|_1 + |\hat{\xi}_1(j) - \xi_1(j)|\|\pi_j\|_1 \\
&\leq C \left[h_j \left(\frac{\log p_n}{nN} \right)^{1/4} + \sqrt{\frac{h_j \log p_n}{nN}} \right] \\
&\leq C h_j \left(\frac{\log p_n}{nN} \right)^{1/4} \qquad (1.61)
\end{aligned}$$

where again we note that on event $\mathcal{E}$, $h_j > \alpha_+ \alpha \sqrt{\frac{\log p_n}{nN}}$ if $j \in J$. Since $\sum_{j=1}^{p} h_j = K$, with probability $1 - o(p_n^{-1})$,

$$\|\hat{A}^{\diamond}_{J*} - \tilde{A}^{\diamond}_{J*}\|_1 \leq C \left(\frac{\log p_n}{nN} \right)^{1/4} = o(1) \qquad (1.62)$$

Now, $\hat{A}_{J*}$ and $\tilde{A}_{J*}$ are defined by normalizing the columns of $\hat{A}^{\diamond}_{J*}$ and $\tilde{A}^{\diamond}_{J*}$, so we have for each $j \in J$ and $k \in [K]$

$$\hat{A}_{jk} = \frac{\hat{A}^{\diamond}_{jk}}{\|\hat{A}^{\diamond}_{Jk}\|_1} \quad \text{and} \quad \tilde{A}_{jk} = \frac{\tilde{A}^{\diamond}_{jk}}{\|\tilde{A}^{\diamond}_{Jk}\|_1} = \frac{A_{jk}}{\|A_{Jk}\|_1}$$

Therefore, for each $j \in J$ and $k \in [K]$,

$$\begin{aligned}
|\hat{A}_{jk} - \tilde{A}_{jk}| &= \left| \frac{\hat{A}^{\diamond}_{jk}}{\|\hat{A}^{\diamond}_{Jk}\|_1} - \frac{\tilde{A}^{\diamond}_{jk}}{\|\tilde{A}^{\diamond}_{Jk}\|_1} \right| \\
&\leq \frac{|\hat{A}^{\diamond}_{jk} - \tilde{A}^{\diamond}_{jk}|}{\|\hat{A}^{\diamond}_{Jk}\|_1} + \tilde{A}^{\diamond}_{jk} \left| \frac{1}{\|\hat{A}^{\diamond}_{Jk}\|_1} - \frac{1}{\|\tilde{A}^{\diamond}_{Jk}\|_1} \right| \\
&\leq \frac{|\hat{A}^{\diamond}_{jk} - \tilde{A}^{\diamond}_{jk}| + \tilde{A}_{jk}\|\hat{A}^{\diamond}_{Jk} - \tilde{A}^{\diamond}_{Jk}\|_1}{\|\hat{A}^{\diamond}_{Jk}\|_1} \\
&= \frac{|\hat{A}^{\diamond}_{jk} - \tilde{A}^{\diamond}_{jk}|}{\|\hat{A}^{\diamond}_{Jk}\|_1} + \frac{A_{jk}\|\hat{A}_{Jk} - \tilde{A}^{\diamond}_{Jk}\|_1}{\|A_{Jk}\|_1 \|\hat{A}^{\diamond}_{Jk}\|_1}
\end{aligned} \tag{1.63}$$

Now,

$$\|A_{Jk}\|_1 = 1 - \|A_{J^c k}\|_1 \geq 1 - \frac{1}{1-q} s(\beta\tau_n)^{1-q} \geq c$$

for some absolute constant $c \in (0, 1)$ as nN becomes sufficiently large. Furthermore, since by definition of Π we have $\tilde{A}^{\diamond}_{Jk} = \mathrm{diag}(\xi_1)\Pi = A_{J*} \cdot \mathrm{diag}(V_1)$ and $\min_{k\in[K]} V_1(k) \geq \frac{c}{\sqrt{K}}$, so

$$\|\tilde{A}^{\diamond}_{Jk}\|_1 = V_1(k)\|A_{Jk}\|_1 \geq c$$

and thus

$$\|\hat{A}^{\diamond}_{Jk}\|_1 \geq \|\tilde{A}^{\diamond}_{Jk}\|_1 - \|\tilde{A}^{\diamond}_{Jk} - \hat{A}^{\diamond}_{Jk}\|_1 \geq c - C\left(\frac{\log p_n}{nN}\right)^{1/4} \geq c/2$$

as nN becomes sufficiently large. Hence, we have from (1.63), (1.61) and (1.62) that

$$|\hat{A}_{jk} - \tilde{A}_{jk}| \leq C h_j \left(\frac{\log p_n}{nN}\right)^{1/4}$$

and so for any $j \in J$,

$$\|\hat{A}_{j*} - \tilde{A}_{j*}\|_1 \leq C h_j \left(\frac{\log p_n}{nN}\right)^{1/4}$$

which, since $\sum_{j=1}^{p} h_j = K$, implies

$$\|\hat{A}_{J*} - \tilde{A}_{J*}\|_1 \leq C \left(\frac{\log p_n}{nN}\right)^{1/4} \tag{1.64}$$

We combine (1.59), (1.60) and (1.64) to obtain what we need to prove. □

1.5.6 Results on Archetype Analysis [Javadi and Montanari, 2020]

To facilitate our discussion on relaxing the separability assumption, we summarize the results of Javadi and Montanari [2020] in this section.

We first introduce the notations in this paper. For a point $u \in \mathbb{R}^d$ and a matrix $V \in \mathbb{R}^{m\times d}$, let

$$\mathscr{D}(u; V) := \min\{\|u - V^T\pi\|_2^2 : \pi \in \Delta^m\}, \text{ where}$$

$$\Delta^m := \{x \in \mathbb{R}^m : x^T\mathbf{1}_m = 1 \text{ and } x_j \geq 0 \text{ for all } j \in [m]\}$$

In words, $\mathscr{D}(u; V)$ is the square of the distance between u and $\text{conv}(V)$, where $\text{conv}(V)$ denotes the convex hull of the rows of V. If $U \in \mathbb{R}^{p\times d}$ is a matrix with rows $u_1, \ldots, u_p \in \mathbb{R}^d$, we generalized the above definition by letting

$$\mathscr{D}(U; V) := \sum_{l=1}^{p} \mathscr{D}(u_l; V) \tag{1.65}$$

Now, consider a factorization of the form $X_0 = W_0H_0$, where the rows of $X_0 \in \mathbb{R}^{m\times(K-1)}$ form a point cloud, $W_0 \in \mathbb{R}^{m\times K}$ is a matrix of weights whose rows are in Δ^K, and the rows of $H_0 \in \mathbb{R}^{K\times(K-1)}$ are the K simplex vertices.

Definition 6 (α-uniqueness)**.** *We say that the point cloud $X_0 = W_0H_0$ satisfies uniqueness with parameter $\alpha > 0$ (or α-uniqueness) if for all $H \in \mathbb{R}^{K\times(K-1)}$ with conv$(X_0) \subseteq$ conv(H),*

we have

$$\mathscr{D}(H; X_0)^{1/2} \geq \mathscr{D}(H_0; X_0)^{1/2} + \alpha[\mathscr{D}(H; H_0)^{1/2} + \mathscr{D}(H_0; H)^{1/2}] \tag{1.66}$$

The motivation behind this assumption is quite clear. Any H with $\text{conv}(X_0) \subseteq \text{conv}(H)$ is a plausible explanation of the data. For H_0 to be identifiable, we want $\mathscr{D}(H; X_0) > \mathscr{D}(H_0; X_0)$ if $H \neq H_0$, and so (1.66) is a quantitative formulation of this requirement. Note that if $X_0 = W_0 H_0$ is a separable factorization, then it always satisfies uniqueness with $\alpha = 1$. Indeed, whenever $\text{conv}(H_0) = \text{conv}(X_0)$, one has $\mathscr{D}(H; X_0) = \mathscr{D}(H; H_0)$ and $\mathscr{D}(H_0; X_0) = \mathscr{D}(H_0; H) = 0$.

The vertex hunting procedure considered in Javadi and Montanari [2020] is as follows. Suppose we observe X which is a noisy version of X_0:

$$X = X_0 + Z = W_0 H_0 + Z \tag{1.67}$$

Let $x_1, \ldots, x_m$ be the rows of X. We can obtain an estimator $\hat{H}$ of H_0 by solving the following optimization problem (Archetype Analysis):

$$\text{minimize } \mathscr{D}(H; X) \text{ s.t. } \mathscr{D}(x_i; H) \leq \delta^2 \text{ for all } i \in [m] \tag{1.68}$$

where $\delta \geq \max_{i \in [m]} \|Z_{i*}\|_2$. In light of Corollary 5, we want to choose $\delta \geq C \left(\frac{\log p_n}{nN}\right)^{1/4}$ in our context, where C is the constant in (1.21) (replace X_0 in (1.67) with the point cloud matrix R from out oracle procedure, and X with the point cloud matrix $\hat{R}$ from Definition 5).

The main theoretical result of Javadi and Montanari [2020] is that their vertex hunting procedure is robust to noise in the point cloud.

Theorem 31 (Theorem 1 of Javadi and Montanari [2020])**.** *Suppose X_0 satisfies the α-uniqueness assumption, and conv(X_0) contains a $(K-1)$-dimensional ball of radius $\mu > 0$.*

Consider the vertex hunting procedure defined by (1.68)*, with* $\delta = \max_{i\in[m]} \|Z_{i*}\|_2$*. If*

$$\max_{i\in[m]} \|Z_{i*}\|_2 \leq \frac{\alpha\mu}{30K^{3/2}}$$

then

$$\|\hat{H} - H_0\|_F^2 \leq \frac{C^2K^5}{\alpha^2}\delta^2 \tag{1.69}$$

Here, the constant C *may depend on* μ *and the maximum/minimum singular values of* H_0*, and we ignore the vertex label permutation (by redefining* $\hat{H}$ *if necessary).*

Using similar proof techniques as in the above theorem, we can also show the following robustness result for Archetype Analysis without using the α-uniqueness condition (the proof is omitted for brevity). In (1.70), we do not need to assume separability (in which case one has $\mathscr{D}(H_0; X_0) = 0$), but we want the distance from the vertices in H_0 to the convex hull of the point cloud X_0 to be no larger than δ. Again, $\delta \asymp \left(\frac{\log p_n}{nN}\right)^{1/4}$ when applied to our topic modeling setup.

Theorem 32. *Using the same assumptions as in Theorem 31 except the* α*-uniqueness condition, if* $\max_{i\in[m]} \|Z_{i*}\|_2 \leq \delta \leq \frac{\mu}{2K+2}$*, the vertex hunting procedure* (1.68) *satisfies for some constants* $C_1, C_2 > 0$*:*

$$\|\hat{H} - H_0\|_F^2 \leq C_1\mathscr{D}(H_0; X_0) + C_2\delta^2 \tag{1.70}$$

In practice, the vertex hunting procedure defined (1.68) is difficult to use. When applied on real dataset, one may prefer to work with the Lagrangian form of (1.68):

$$\hat{H}_\lambda = \arg\min_H [\mathscr{D}(X; H) + \lambda\mathscr{D}(H; X)] \tag{1.71}$$

Algorithms to solve this non-convex optimization problem are available in Section 4 of Javadi and Montanari [2020].

1.5.7 Further details on experiments with synthetic data

1.5.7.1 Zipf's law: illustrations and comparisons

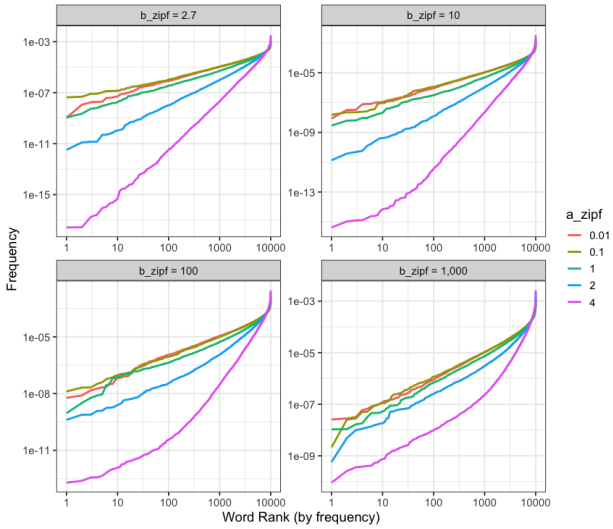
Most of the synthetic experiments we use in this paper rely on the generation of documents where word frequencies follow a Zipf's law distribution (see Equation 1.35). Figure 1.12a illustrates instances of such frequency distributions for a dictionary of size $p = 10,000$ words as we vary the parameters of this distribution (namely, the values of α_{zipf} and β_{zipf}). Figure 1.12b compares the frequency heterogeneity resulting from sampling frequencies $f_{(j)}$'s from a Zipf's law distribution (Equation 1.35) to frequencies sampled from a Uniform distribution:

$$f_{(j)} \propto \text{Uniform}(0, 1).$$

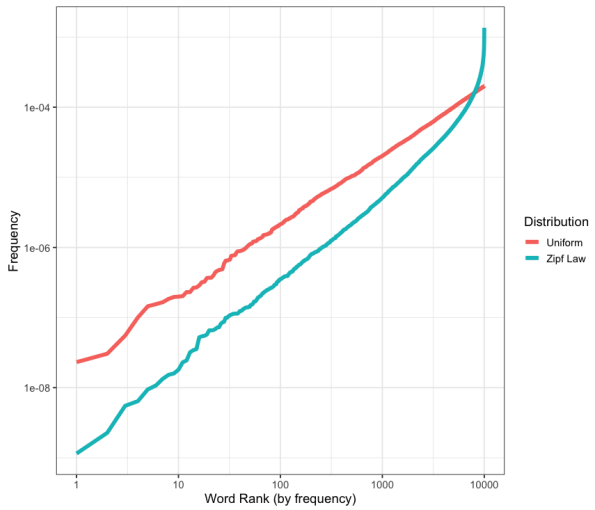
These two figures illustrate the fast decay in word frequencies under Zipf's law. With $a_{\text{zipf}} = 1$ and $b_{\text{zipf}} = 2.7$ (a choice of parameters empirically observed to fit the behavior of real text data), only 10% of words have frequencies above 0.001. The rate of decay increases rapidly as the parameter α_{zipf} increases (Figure 1.12a). By comparison, under the uniform distribution often assumed in other papers, all word frequencies are of the same order of magnitude. Our weak sparsity assumption, imposed on the row sums of the topic matrix A, is well-aligned with the empirically observed Zipf's law.

1.5.7.2 Synthetic data from the uniform distribution of non-anchor words

For experiments involving the uniform distribution, the data generation mechanism is as follows:



(a) Illustrations of word frequency distributions under Zip's law. As α_{zipf} increases, the frequencies decay faster.

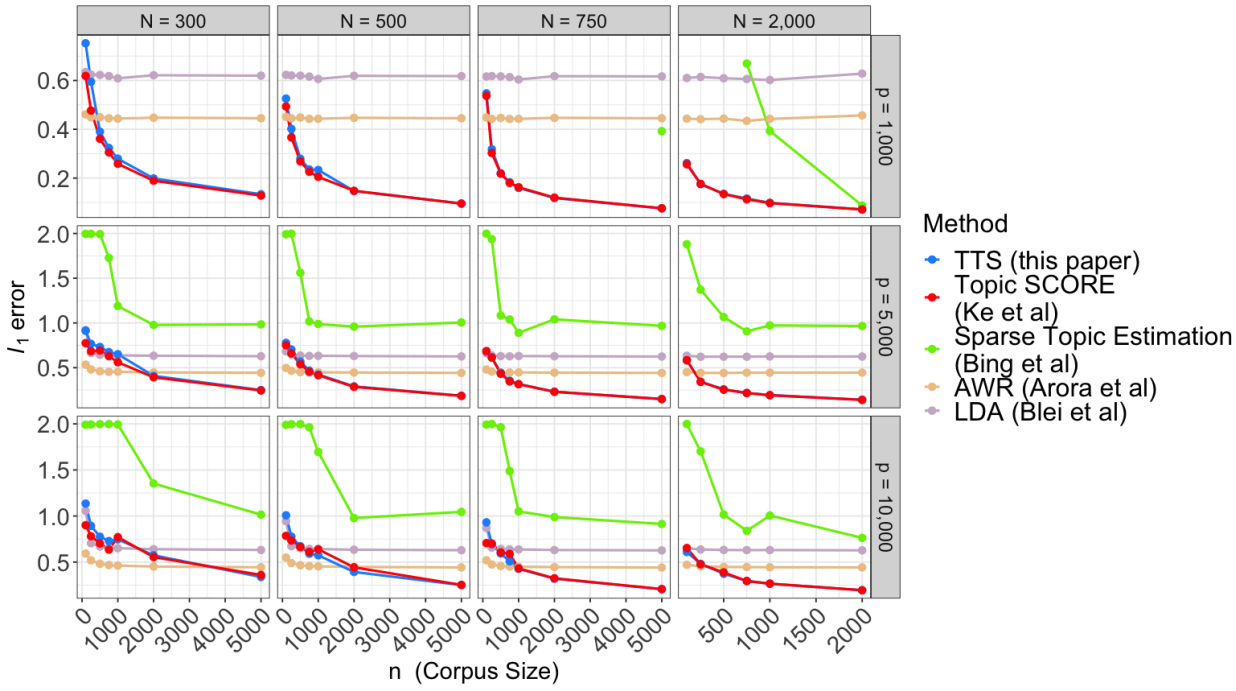


(b) Word frequencies generated under the Uniform distribution (red) and the Zipf's law distribution (blue) with $a_{\text{zipf}} = 1$ and $b_{\text{zipf}} = 2.7$.

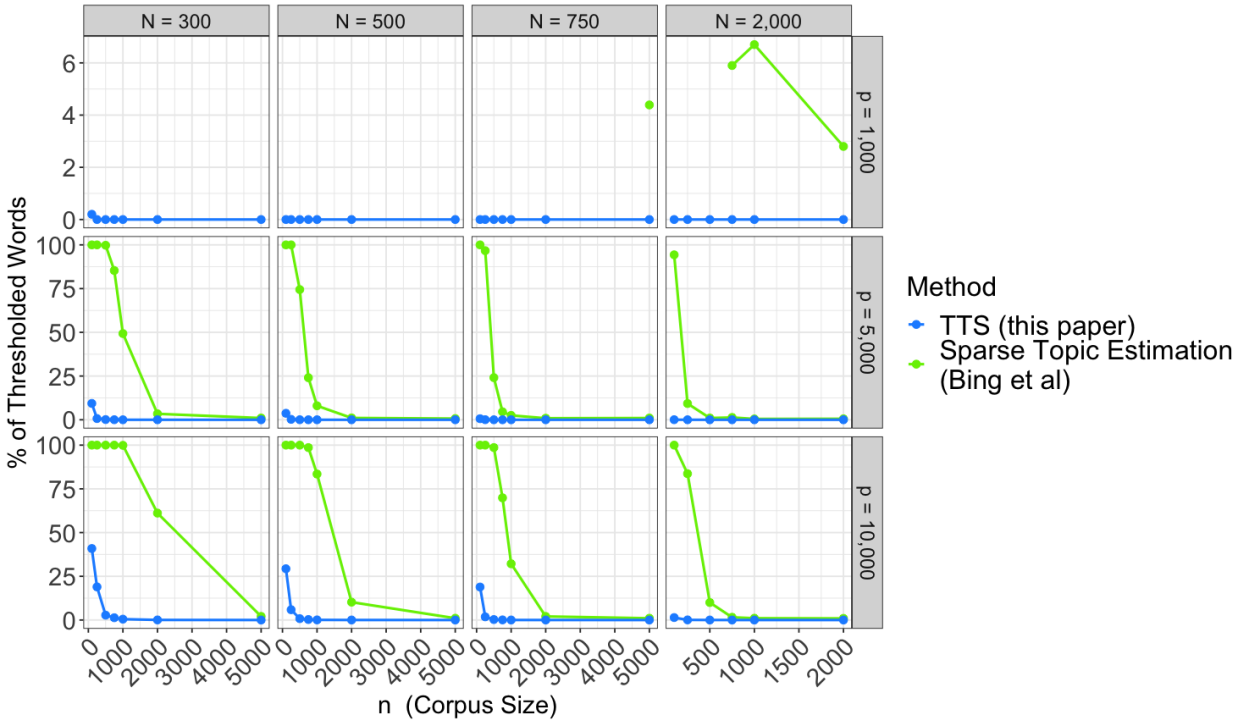
Figure 1.12: Comparisons between word frequencies from different generation mechanisms. Both axes are log-scaled.

$$\begin{aligned} \forall i \in [n], \quad & W_{*i} \sim \text{Dirichlet}(\mathbf{1}_K) \\ \forall j \in \{5K+1, \cdots p\} \text{ and } k \in [K], \quad & A_{jk} \sim \text{Uniform}(0,1) \end{aligned} \tag{1.72}$$

and for $j \in [5K]$, each topic $k \in [K]$ has 5 anchor words, with $A_{jk} = \delta_{\text{anchor}}$ if word j is an anchor word of topic k and $A_{jk} = 0$ otherwise. This setup is identical to the main text, except for the uniform distribution used to generate word frequencies. As noted before, the assumption that word frequencies all have roughly the same amplitude does not align with our weak sparsity assumption (Assumption 5) or the behavior of word frequencies in real text data [Corral et al., 2015]. Nonetheless, we also perform experiments in this setting and report the results in Figure 1.13a. Results are averaged over 50 experiments. We use $K = 5$, $\alpha_{\text{dirichlet}} = 1$, and 5 anchor words with intensity $\delta_{\text{anchor}} = 0.001$.



(a) Median ℓ_1 error $\mathcal{L}_1(\hat{A}, A) = \min_{\Pi \in \mathcal{P}} \frac{1}{K} \|\hat{A}\Pi - A\|_1$ for different methods.



(b) Percentage of thresholded words as a function of n, N and p.

Figure 1.13: Performance of the different methods under the uniform frequency generation mechanism detailed in Equation 1.72. For small vocabulary size p, the method of Bing et al. [2020b] does not appear as the number of topics it estimated was less than the true value $K = 5$; therefore, we were unable to evaluate its performance.

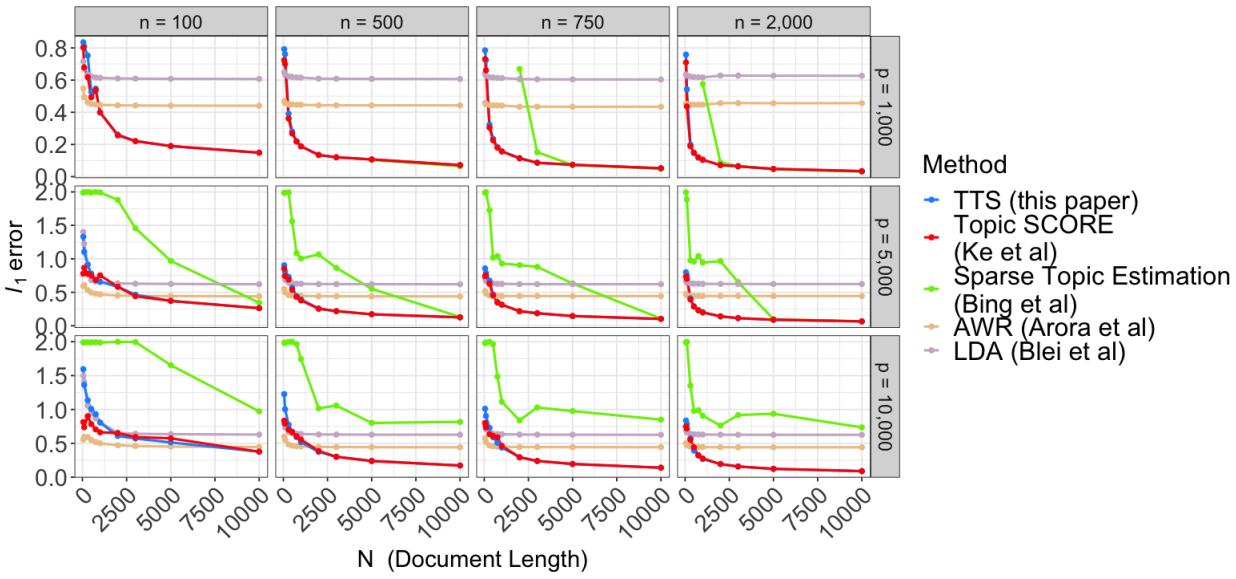


Figure 1.14: Same experiments as those in Figure 1.13. Results here are plotted as a function of N

We note that Topic-SCORE [Ke and Wang, 2022] and our method perform similarly under the uniform generating distribution. As shown in Figure 1.13b, our method does not threshold much in this regime, which is expected since all word frequencies are roughly of the same order. While our method is mainly designed to leverage the weak sparsity of the matrix A, experiments suggest that it performs well even if the weak sparsity assumption is violated.

1.5.7.3 Varying additional parameters

We also examine the effects of the anchor word frequency δ_{anchor} and the Zipf's law parameter α_{zipf} on the performance of our estimator.

As observed in Figure 1.15, the frequency of the anchor words does not appear to have a great impact on the results of the SCORE-based methods; this suggests SCORE-based methods are less dependent on the presence of anchor words. Increasing the frequency of anchor words seems to improve the performance of LDA [Blei et al., 2003] and the method of Bing et al. [2020a].

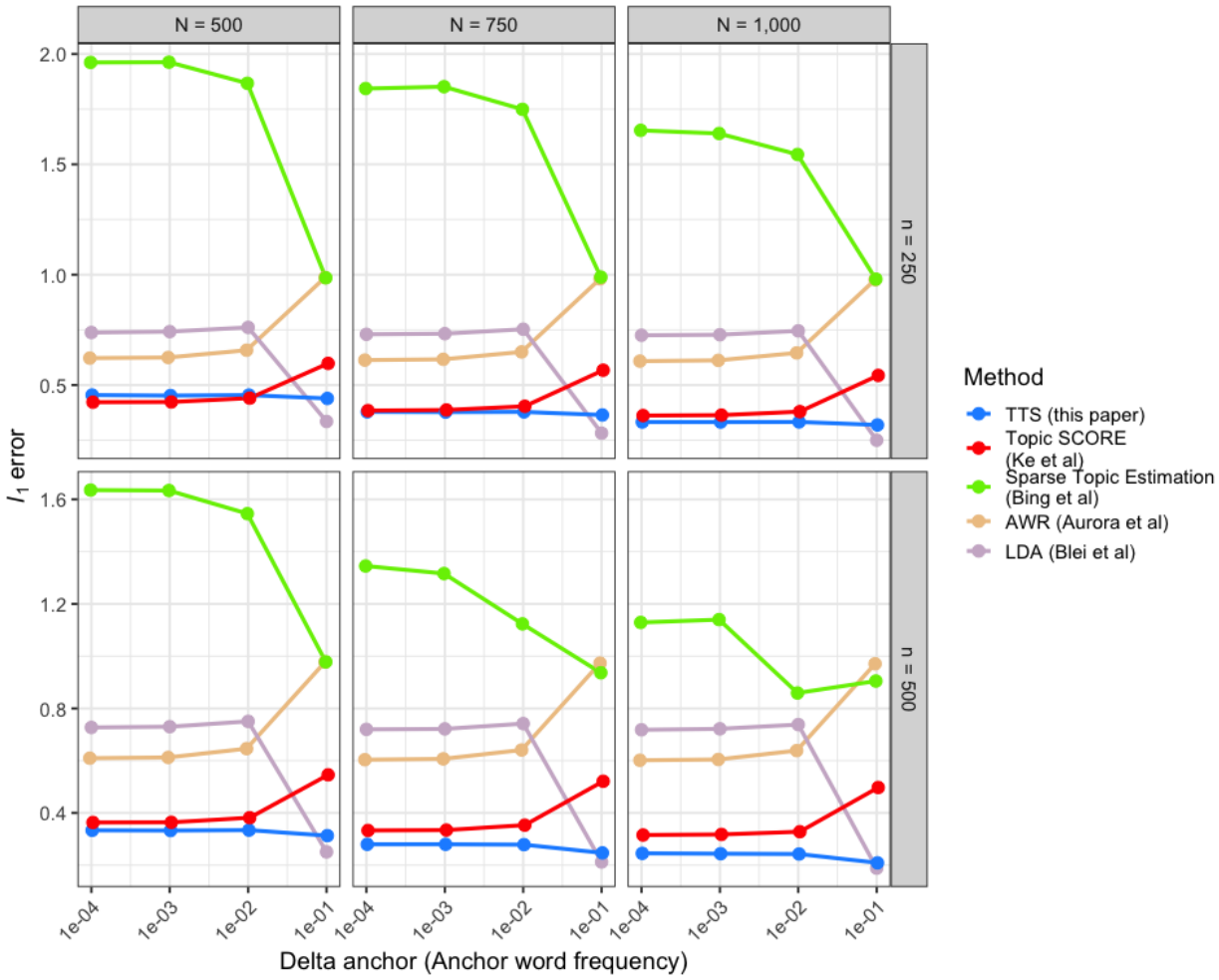


Figure 1.15: Performance of various estimators as δ_{anchor} varies. Here the number of topics is fixed at $K = 3$ and the dictionary has size $p = 5,000$. 5 anchor words are used per topic.

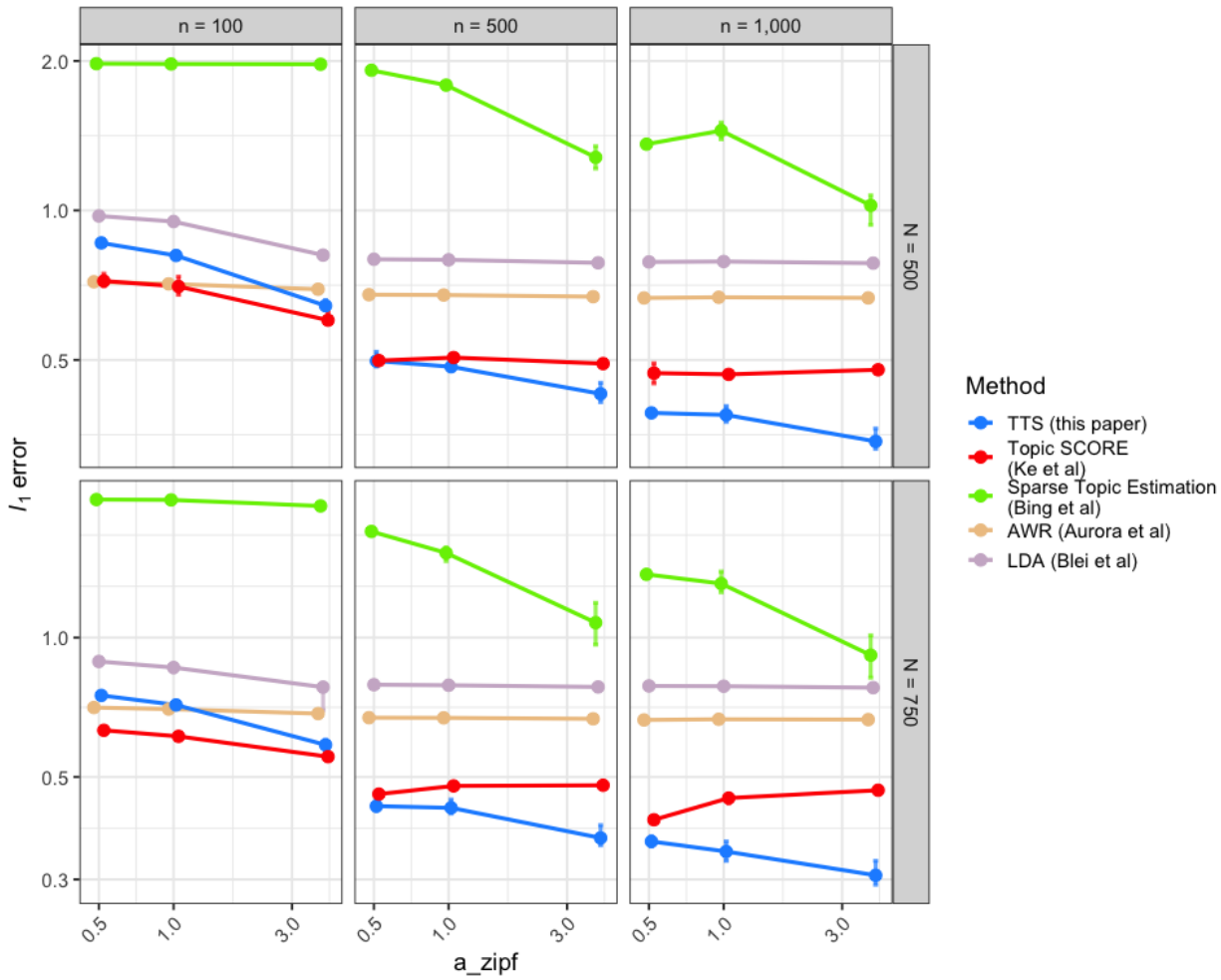


Figure 1.16: Performance of various estimators as α_{zipf} varies. The number of topics is fixed to $K = 5$ and the dictionary has size $p = 10,000$. Each topic has 5 anchor words with frequency $\delta_{\text{anchor}} = 0.001$.

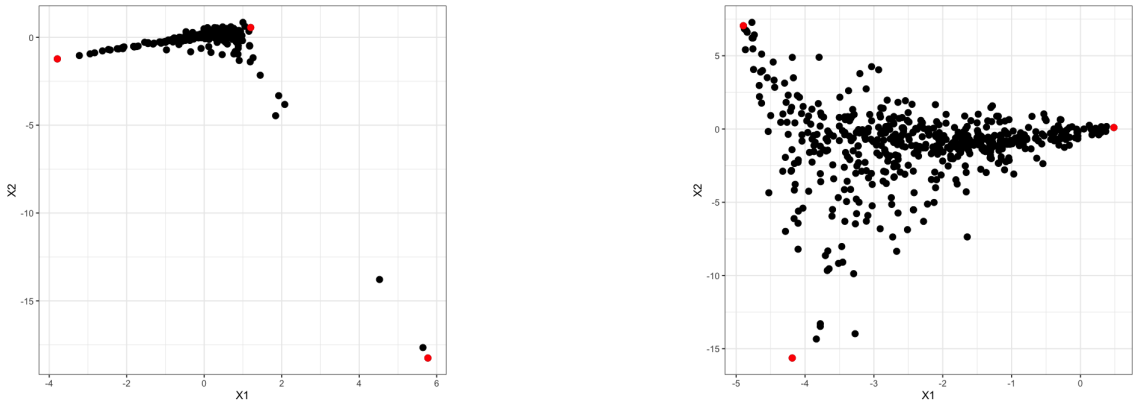
As observed in Figure 1.16, our method offers significant improvement over others when the word frequency decay rate is higher. In comparison, Topic-SCORE's performance does not vary much when α_{zipf} increases. This suggests that our method is able to leverage the sparsity structure to improve estimation.

1.5.8 Further details on real-data experiments

In this section, we provide additional discussions of how our method can be applied to microbiome data analysis. We focus on two microbiome datasets; one is the colon dataset of Yachida et al. [2019] from before (with low p, low n, high N) and the other is the vaginal microbiome data of Callahan et al. [2017] (with medium p, medium n, high N).

1.5.8.1 Microbiome dataset from Yachida et al. [2019]

As depicted in Figure 1.11, our approach consistently produces a higher average topic resolution score compared to Topic SCORE [Ke and Wang, 2022]. To understand the reasons for the gap in performance between Topic-SCORE and our method, we compare the point clouds produced by both (illustrated in Figures 1.17a and 1.17b). It is evident that low-frequency words (words with small h_j) heavily distort the point cloud obtain from Topic-SCORE, thus leading to higher errors in the subsequent vertex hunting step. In comparison, the point cloud obtained from our method is much less distorted. This suggests that the thresholding step in our method, which is not present in the Topic-SCORE algorithm, is helpful in improving the signal-to-noise ratio of the point cloud.



(a) Point cloud produced by Topic-SCORE (b) Point cloud from our method

Figure 1.17: Comparison of the point clouds obtained by our method (right) and Topic-SCORE (left), with $K = 3$. Simplex vertices are colored red. Note that the point cloud from Topic-SCORE is heavily distorted by a few outliers.

1.5.8.2 Microbiome dataset from Callahan et al. [2017]

We also revisit the dataset of Callahan et al. [2017], which serves as an example in Fukuyama et al. [2021] to justify their topic refinement procedure. This dataset comprises amplicon sequence variant (ASV) counts for 2,699 different bacterial species from 2,179 longitudinal samples collected throughout pregnancy in 135 individuals [Callahan et al., 2017]. In this case, the average sample length is around $N = 157,500$. In Fukuyama et al. [2021], based on the refinement results of the LDA, the authors conclude that the topic analysis should be done using $K = 7$ topics, or with up to $K = 12$ if one allows for the possibility of spurious topics. We thus fit up to 12 topics and plot the average resolution (Figure 1.18) and refinement from various methods in Figure 1.19. We find that our method performs better than Topic-SCORE and similarly to LDA in terms of average resolution. For a small number of topics ($K \leq 7$), our method seems preferable to LDA in terms of topic resolution, achieving better resolution at a lower computational cost.

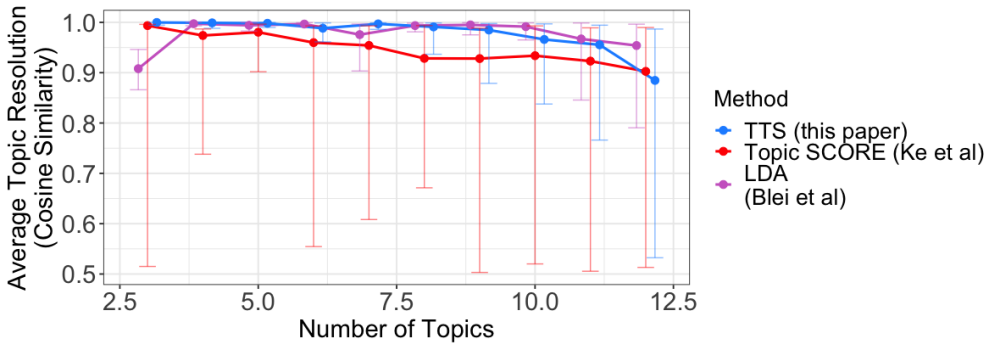


Figure 1.18: Topic resolution scores for our method, Topic-SCORE, and LDA as K varies.

LDA seems to perform better with topic coherence at $K = 7$ (the recommended choice of K by Fukuyama et al. [2021]), although for small K our method also compares favorably. For $K \geq 8$, when compared with LDA, our method seems to yield topics that are often recombined from one hierarchy level to the next (lower topic coherence). This suggests that the choice $K = 7$ by the authors is appropriate, and also suggests that for large K and datasets of moderate size, LDA seems to be a preferable choice.

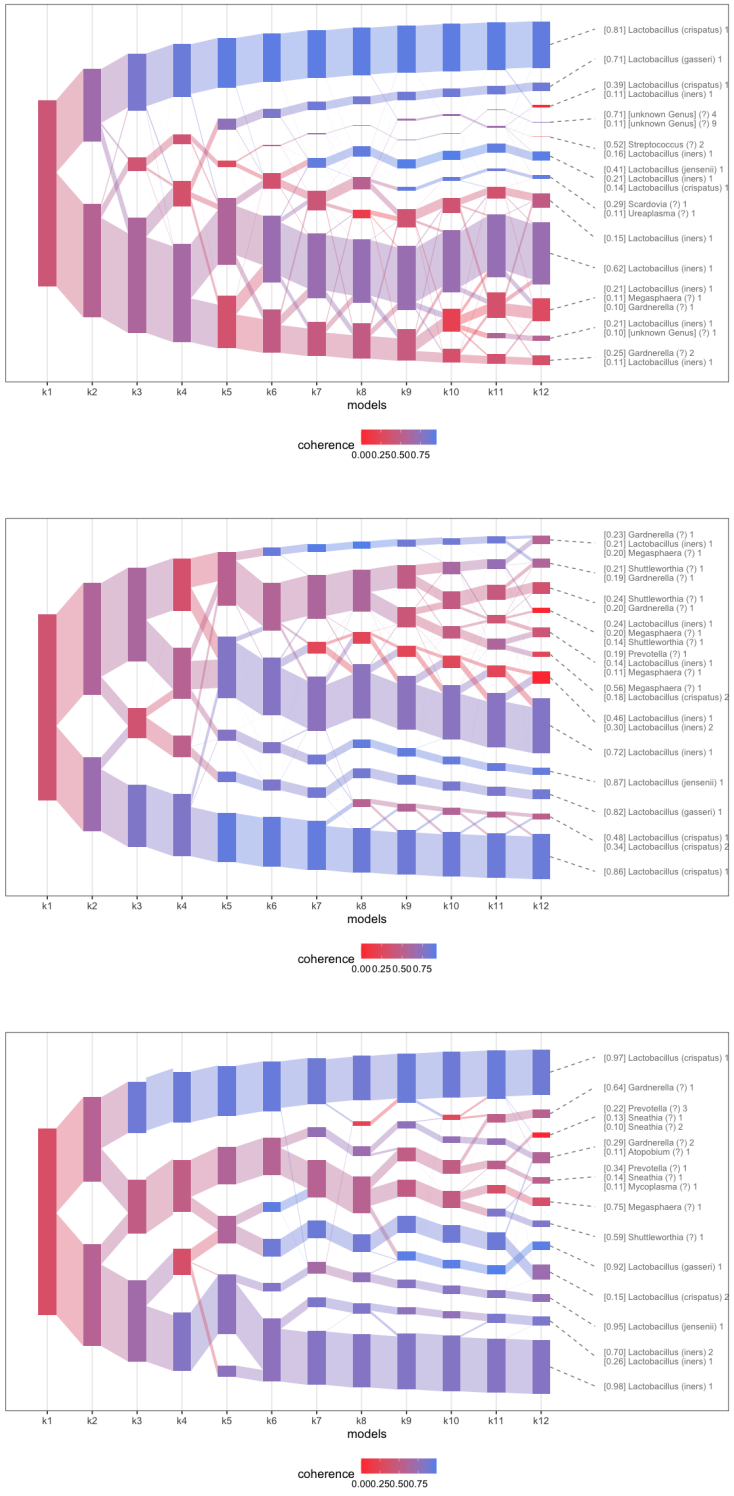


Figure 1.19: Topic coherence and refinement (computed by the method of Fukuyama et al. [2021]) from Topic-SCORE, our method and LDA (in that order) for the vaginal microbiome data of Callahan et al. [2017]. Topics are colored by coherence.

CHAPTER 2

THE GENERALIZED ELASTIC NET PENALTY

2.1 Introduction

In this project, we propose a novel $\ell_1 + \ell_2$-penalty, which we refer to as the *Generalized Elastic Net*, for regression problems where the feature vectors are indexed by vertices of a given graph and the true signal is believed to be smooth or piecewise constant with respect to this graph. Under the assumption of correlated Gaussian design, we derive upper bounds for the prediction and estimation errors, which are graph-dependent and consist of a parametric rate for the unpenalized portion of the regression vector and another term that depends on our network alignment assumption. We also provide a coordinate descent procedure based on the Lagrange dual objective to compute this estimator for large-scale problems. Finally, we compare our proposed estimator to existing regularized estimators on a number of real and synthetic datasets and discuss its potential limitations.

2.1.1 Problem formulation and the proposed penalty

Consider the usual linear regression model

$$Y = X\beta^* + \epsilon \tag{2.1}$$

where the design matrix $X \in \mathbb{R}^{n \times p}$ is random with independent and identically distributed (i.i.d.) rows, $\beta^* \in \mathbb{R}^p$ is the unknown true parameter, and $\epsilon = (\epsilon_1, \ldots, \epsilon_n)^T \in \mathbb{R}^n$ are i.i.d. zero-mean Gaussian variables with (unknown) variance σ^2 and are independent of the design matrix X. In addition to observing the responses $y = (y_1, \ldots, y_n)^T \in \mathbb{R}^n$, we also observe an undirected simple graph $G = (V, E)$ with p vertices and m edges. Here, the p vertices index the entries of β^* as well as the columns of X (which we can think of as feature

vectors). This situation typically entails significant correlation between feature vectors, thus leading to an ill-conditioned design matrix. For simplicity, we assume throughout this paper that the rows of X are i.i.d. $N(0, \Sigma)$-distributed. In our setting, the minimum eigenvalue of $\Sigma \in \mathbb{R}^{p \times p}$ may be very small and β^* might be nearly unidentifiable. Although the addition of an unpenalized intercept should present no difficulty, we assume no intercept in our model to simplify the theoretical analysis.

We further assume that β^* is structured with respect to the graph G so that prediction and estimation can be done with small error, even in the high-dimensional setting where $p \gg n$. As the entries of β^* are indexed by the vertices of G, a natural assumption is that β_i^* and β_j^* should be similar if i and j are adjacent vertices on the graph G. This assumption is related to the notion of *network cohesion* as discussed in Chapter 4 of Kolaczyk [2009]: vertices may display similar characteristics because they are connected (contagion), or they may be connected because they have similar characteristics (homophily). Note that, however, many prior works such as Li et al. [2019] discuss network cohesion in the context where observations (the responses $y_1, \ldots, y_n$ and the rows $x_1, \ldots, x_n$ of X) are indexed by a graph's vertices and thus may no longer be i.i.d., whereas we focus on the case where the features (the columns of X) are indexed by the graph's vertices. Following Li et al. [2019], we also use the term *network cohesion* to cover both homophily and contagion, without distinguishing the difference in causal direction between them.

More specifically, the notion of network cohesion encourages us to assume either that the number of edges $(i, j) \in E$ where $\beta_i \neq \beta_j$ is small (sparse signal jumps), or that β^* is smooth enough so that $\Gamma\beta^*$ lies in an ℓ_q-ball, where Γ is the edge-incidence matrix of the graph G and $0 < q \leq 1$ (note that when $q \in (0, 1)$, an ℓ_q-ball is not convex - see Figure 7.1 of Wainwright [2019] for an illustration of what this "ball" looks like). Mathematically, in our theoretical analysis we assume either

$$\|\Gamma\beta^*\|_0 \leq s \tag{2.2}$$

or

$$\sum_{j=1}^{m} |(\Gamma\beta^*)_j|^q \leq R_q \tag{2.3}$$

for some $s \ll m$ or some $R_q > 0$, respectively (see Section 2.1.4 for the precise definition of any mathematical symbol). Assumption (2.2) means that the number of edges with nonzero signal jumps is small and the true signal has several piecewise constant regions on the graph, whereas Assumption (2.3) means the signal is smooth over the graph in the ℓ_q-sense. We use the term *network alignment* to refer to either Assumption (2.2) or Assumption (2.3). In our experiments, we sometimes also consider the notion that β^* is smooth over the graph in the sense that $\|\Gamma\beta^*\|_\infty$ is small. We emphasize that we allow for the possibility that the entries β^* are all nonzero, as long as β^* satisfies (or can be well approximated by an oracle β that satisfies) either Assumption (2.2) or Assumption (2.3).

Under Model (2.1) and either Assumption (2.2) or (2.3), we study the prediction and estimation errors of the following estimator

$$\hat{\beta} := \arg\min_{\beta\in\mathbb{R}^p} \frac{1}{n}\|Y - X\beta\|_2^2 + \lambda_1\|\Gamma\beta\|_1 + \lambda_2\|\Gamma\beta\|_2^2 \tag{2.4}$$

which can also be rewritten as

$$\hat{\beta} := \arg\min_{\beta\in\mathbb{R}^p} \frac{1}{n}\|Y - X\beta\|_2^2 + \lambda_1 \sum_{(i,j)\in E} |\beta_i - \beta_j| + \lambda_2 \sum_{(i,j)\in E} (\beta_i - \beta_j)^2 \tag{2.5}$$

Note that our focus is mainly on the penalty $\lambda_1\|\Gamma\beta\|_1 + \lambda_2\|\Gamma\beta\|_2^2$ where Γ is the incidence matrix of a general graph. Following the naming conventions in Zou and Hastie [2005] and Tibshirani and Taylor [2011], we refer to this penalty as the *Generalized Elastic Net (GEN)* penalty. The estimator (2.4) can be easily extended to the generalized linear model (GLM) setting, by replacing the term $\frac{1}{n}\|Y - X\beta\|_2^2$ with another negative log-likelihood function from an exponential family distribution (see Chapter 9 of Wainwright [2019] for more examples).

For instance, if we have binary responses that can be modeled with the logistic GLM, then using the logistic log-likelihood function gives

$$\hat{\beta}_{\text{logistic}} := \frac{1}{n}\sum_{i=1}^{n}\log(1+e^{\langle x_i,\beta\rangle}) - \left\langle \frac{1}{n}\sum_{i=1}^{n} y_i x_i, \beta \right\rangle + \lambda_1\|\Gamma\beta\|_1 + \lambda_2\|\Gamma\beta\|_2^2 \tag{2.6}$$

where $x_1, \ldots, x_n$ are the rows of X and $y_1, \ldots, y_n$ are the entries of Y which are binary. For simplicity, we only focus on analyzing the estimator (2.4) under Model (2.1), but analogous theoretical results for the GLM setting should follow by adapting the theoretical framework of Chapter 6 of Bühlmann and van de Geer [2011].

2.1.2 Motivating applications

As network-linked features are quite common and we do not restrict our attention to any particular type of graph, our proposed penalty is potentially applicable to a wide variety of settings. We provide below a non-exhaustive list of concrete examples where our penalty may be relevant.

Example 1: Structural MRI analysis. We consider the use of structural magnetic resonance images (sMRI) of the brain in diagnosing Alzheimer's disease, as in Xin et al. [2014]. In this case, the rows $x_1, \ldots, x_n$ of X might represent sMRI features of n human subjects and the responses $y_1, \ldots, y_n$ are binary variables indicating each subject's disease status. The estimator (2.6) can thus be applied using a 3D grid graph representing contiguous brain voxels. In Xin et al. [2014], the Generalized Fused Lasso penalty $\lambda_L\|\beta\|_1 + \lambda_1\|\Gamma\beta\|_1$ is used, and this penalty leads to a solution that is both sparse and smooth. However, it may be more reasonable to assume only that the true signal aligns with the graph, in which case the estimator (2.6) may fare better for the purpose of predicting Alzheimer's disease.

Example 2: Microarray analysis with prior information. Following Segal et al. [2003], we can also consider a microarray dataset with $X = [x_{ij}]$ where x_{ij} is the expression level of the j^{th} gene for the i^{th} test subject, and y_i is an outcome measure for subject i which can be continuous or discrete. Often, we have prior knowledge from previous biomedical research in the form of gene regulatory pathways which can serve as our graph G (see Li and Li [2008] for specific examples). We can incorporate this prior information using our GEN penalty. In Li and Li [2008], the penalty $\lambda_L\|\beta\|_1 + \lambda_2\beta^T\tilde{L}\beta$ is used instead, where $\tilde{L}$ is the normalized Laplacian matrix. Assuming the vast majority of genes has no effect on the outcome may make it easier to interpret the estimated parameters (in terms of which genes may be responsible for the outcome). However, if many of these genes can be grouped into clusters with small (but nonzero) baseline effects on the outcome, using our penalty may lead to better predictions.

Example 3: Microarray analysis without prior information. In the previous example, without any prior information about gene regulatory pathways, we can take G to be the complete graph in our GEN penalty. The penalty $\lambda_L\|\beta\|_1+\lambda_1\|\Gamma\beta\|_1$, where Γ is the incidence matrix of a complete graph, has been studied in She [2008] under the name *Clustered LASSO*.

Example 4: Temporal data. Given a time series $\{X_t\}_{t\in\mathbb{N}}$, we consider fitting an autoregressive model of the form $X_t = \sum_{j=1}^{p}\beta_j X_{t-j} + \epsilon_t$. If the time points t are sampled sufficiently far apart such that our data points $(X_t, X_{t-1}, \ldots, X_{t-p})$ can be considered independent across t, it may be reasonable to apply our method with G being a p-vertex chain graph.

2.1.3 Comparison with related works

The standalone ℓ_1 penalty $\lambda_1\|\Gamma\beta\|_1$, which is often known as the *total variation penalty* on graphs, has been studied extensively in the context of the graph trend filtering problem where the design matrix is the identity. More precisely, given the model $Y = \beta^* + \epsilon$, the trend filtering estimator for β^* is

$$\hat{\beta}_{\text{tf}} := \arg\min_{\beta\in\mathbb{R}^n} \frac{1}{n}\|Y-\beta\|_2^2 + \lambda_1\|\Gamma\beta\|_1 \tag{2.7}$$

This estimator is also known as the *analysis* estimator, in the terminology of Elad et al. [2007]; see Hütter and Rigollet [2016], Wang et al. [2015], Ortelli and van de Geer [2021] and Guntuboyina et al. [2020] for results on prediction error bounds for the estimator (2.7) and its constrained form when $\Gamma\beta^*$ is sparse. The graph considered in the trend filtering problem is usually a chain or grid graph due to applications such as image denoising, but results for other types of graphs such as trees and star graphs are also available in the literature. The analysis matrix Γ in (2.7) can be generalized to higher order total variation operators (as defined in Wang et al. [2015]). In comparison, we focus solely on the case where Γ is the incidence matrix defined in (2.9), and our design matrix X is random with i.i.d. rows rather than a pre-specified matrix consisting of fixed vectors from some dictionary.

When the design matrix is general, the estimator

$$\hat{\beta}_{\text{GL}} := \arg\min_{\beta\in\mathbb{R}^p} \frac{1}{n}\|Y-X\beta\|_2^2 + \lambda_1\|\Gamma\beta\|_1 \tag{2.8}$$

has been proposed by Tibshirani and Taylor [2011] (under the name *Generalized LASSO* estimator) as well as Land and Friedman [1997] (where the penalty is called *variable fusion*). These works mainly address computational techniques for the estimator (2.8), rather than theoretical guarantees when β^* aligns with the graph. The idea of working with the dual objective to derive our algorithms comes from Kim et al. [2009] and Tibshirani and Taylor

[2011]. Our analysis of the prediction and estimation errors for the estimator (2.4) is also applicable to (2.8), and to our knowledge no similar analysis with random design is available in the literature. However, the error bounds for our estimator (2.4) are better due to the improved minimum eigenvalue in the denominator of the bounds in Theorem 33.

Two previously introduced penalties which involve the Lasso penalty to induce sparsity but are closely related to GEN have also been studied in the context where the design matrix can be non-identity; they serve as the main benchmarks in both our theoretical results and our experiments. The Smooth Lasso penalty $\lambda_L\|\beta\|_1+\lambda_2\|\Gamma\beta\|_2^2$ was first proposed by Hebiri and van de Geer [2011], in which the theoretical analysis assumes fixed design and thus relies on a restricted eigenvalue assumption (Assumption $B(\Theta)$ in Hebiri and van de Geer [2011]) on the expanded Gram matrix $n^{-1}\tilde{X}^T\tilde{X}$ (see Section 2.1.4 for definition of $\tilde{X}$). The Fused Lasso penalty $\lambda_L\|\beta\|_1+\lambda_1\|\Gamma\beta\|_1$ was first proposed by Tibshirani et al. [2005] for the chain graph. These two methods implicitly assume that the true signal is both sparse and aligned with the graph. Such an assumption can be overly restrictive, and sparsity of β^* may not always be a natural assumption in the general graph setting. When $\|\beta^*\|_0=p$, error bounds proven for these estimators usually involve the term $\frac{p\log p}{n}$. In comparison, our penalty only assumes network alignment and should also work well in the sparse-and-smooth case when the zero entries of β^* form large contiguous blocks on the graph. The Fused Lasso and the Smooth Lasso should only perform better than ours when sparsity holds but the network alignment assumption is significantly violated. Empirically, when β^* aligns with the graph but is not sparse, choosing the tuning parameters by cross-validation often results in λ_L being set to almost zero for both the Fused Lasso and the Smooth Lasso.

In Li et al. [2018], the penalty $\lambda_2\|\hat{\Gamma}\beta\|_2^2+\lambda_1\|\hat{\Gamma}\beta\|_1+\lambda_L\|\beta\|_1$ is introduced and referred to as the Graph Total Variation (GTV) method, which involves three hyperparameters that require tuning. Unlike our penalty, the incidence matrix $\hat{\Gamma}$ is obtained by first estimating Σ with $\hat{\Sigma}$ (which can depend on the design X or side information) and then treating $\hat{\Sigma}$ as the

adjacency matrix of a graph $\hat{G}$ with weighted edges. Note that this is a two-step process, and the graph $\hat{G}$ here also differs from our setting in that we do not consider non-binary edge weights, since in many applications only a graph structure is provided. Computationally, since we need to use 3D grid search for hyperparameter tuning and the matrix $\hat{\Gamma}$ is very dense, the estimator introduced in Li et al. [2018] does not scale well. Furthermore, even when we use the true covariance Σ to form $\hat{\Gamma}$, the performance of GTV in most of our synthetic experiments does not compare favorably with that of our method, Fused Lasso or Smooth Lasso. The theoretical analysis in Li et al. [2018] does not account for the error in estimating Σ with $\hat{\Sigma}$, which we believe cannot be overlooked.

2.1.4 Notations and definitions

For any positive integer n, we denote $[n]$ as the set $\{1, \ldots, n\}$. For any matrix A, we denote by $A^\dagger$ the Moore-Penrose inverse of A. For any vector v, $\|v\|_0$ refers to the number of nonzero entries of v, and $\|v\|_p$ for $1 \leq p \leq \infty$ refers to the usual ℓ_p-norm of v. We write $\mathbf{1}(\cdot)$ for the indicator function. For a vector $v \in \mathbb{R}^k$ and any set $S \subseteq [k]$, we denote by $v_S \in \mathbb{R}^k$ to be the vector with the j^{th} coordinate given by $(x_S)_j = x_j \mathbf{1}(j \in S)$. For any vector $\theta \in \mathbb{R}^m$, we write S_θ to refer to the support $\{j \in [m] : \theta_j \neq 0\}$ of θ. We use s to denote $\|\Gamma\beta^*\|_0$. For any positive semi-definite matrix M, let $\gamma_{\max}(M)$ and $\gamma_{\min}(M)$ denote its maximum and minimum eigenvalues respectively, and $\ker(M)$ the null space of M. I_k denotes the identity matrix of size k-by-k.

The notation $\lesssim$ means that the left-hand side (LHS) is bounded by the right-hand side (RHS) multiplied by an absolute constant (not dependent on any parameter of interest) that is omitted. The notation $\gtrsim$ is similarly defined. The notation $\asymp$ means that both $\lesssim$ and $\gtrsim$ hold. The constants C, c, c_1, c_2 are absolute constants which are allowed to change line by line.

Throughout this project, the graph $G = (V, E)$ we consider is undirected and has no

self-loops. We identify the set of vertices V with $[p]$ and the set of edges E with $[m]$; note that $m \leq p^2$, and $p \lesssim m$ if the graph has no isolated vertices. We also denote the maximum degree of the graph G by d and the number of connected components of G by n_c (which is also the dimension of the null space of Γ). The edge-vertex incidence matrix of the graph G is denoted by $\Gamma \in \{-1, 0, 1\}^{m \times p}$, which is defined as follows: each edge $e = (i, j) \in E$ is represented by a row $\Gamma_{e,\cdot} \in \{-1, 0, 1\}^p$ of Γ whose k^{th} entry is given by

$$\Gamma_{e,k} = \begin{cases} 1 & \text{if } k = \min(i, j) \\ -1 & \text{if } k = \max(i, j) \\ 0 & \text{otherwise.} \end{cases} \tag{2.9}$$

The unnormalized Laplacian matrix of the graph G (see Chung and Graham [1997]) is then defined by $L := \Gamma^T \Gamma$. We denote by $\Pi \in \mathbb{R}^{p \times p}$ the projection matrix onto the kernel of Γ. Note that we will use the facts $\Pi = \Pi^T$, $\Pi^2 = \Pi$ and $\Pi + \Gamma^\dagger \Gamma = I_p$ throughout the proofs.

In our theoretical analysis, we frequently make use of some definitions and conventions from Hütter and Rigollet [2016]. We denote $s_1, \ldots, s_m$ to be the columns of $\Gamma^\dagger \in \mathbb{R}^{p \times m}$. The *inverse scaling factor* of Γ is defined as

$$\rho(\Gamma) := \max_{j \in [m]} \|s_j\|_2 \tag{2.10}$$

while the *compatibility factor* of Γ for a nonempty set $S \subseteq [m]$ is defined as

$$k_S := \inf_{\beta \in \mathbb{R}^p} \frac{\sqrt{|S|}\|\beta\|_2}{\|(\Gamma\beta)_S\|_1} \tag{2.11}$$

Following Hebiri and van de Geer [2011], we also employ the notations

$$\tilde{Y} := \begin{pmatrix} Y \\ 0 \end{pmatrix}, \quad \tilde{X} := \begin{pmatrix} X \\ \sqrt{\lambda_2 n}\Gamma \end{pmatrix}, \quad \tilde{\epsilon} := \begin{pmatrix} \epsilon \\ -\sqrt{\lambda_2 n}\Gamma\beta^* \end{pmatrix} \tag{2.12}$$

Note that $\tilde{Y} = \tilde{X}\beta^* + \tilde{\epsilon}$ and we can write our estimator as

$$\hat{\beta} = \arg\min_{\beta \in \mathbb{R}^p} \frac{1}{n}\|\tilde{Y} - \tilde{X}\beta\|_2^2 + \lambda_1 \|\Gamma\beta\|_1 \tag{2.13}$$

2.2 Theoretical results

In this section, we aim to provide non-asymptotic bounds showing that the estimator (2.4) is consistent in prediction and estimation under a network alignment assumption, even in the high-dimensional setting where $p \gg n$. We also show that the ℓ_2 component of the penalty helps alleviate the effects of an ill-conditioned covariance matrix Σ. Note that the tuning parameters λ_1 and λ_2 in our theoretical analysis are dependent on unobserved quantities β^*, Σ and σ; therefore, we cannot use the theoretical values for λ_1 and λ_2 in practice and must in general rely on cross-validation. We do not attempt to optimize the constants in our bounds, as our focus is on understanding how the performance of our estimator depends on the quantities n, p, s (or R_q), Σ and the graph G.

2.2.1 Main theorem

We begin by introducing bounds for the prediction and estimation errors that are applicable to all graphs. However, these bounds may not be optimal for some graphs, especially the p-vertex chain graph as in that case $\rho(\Gamma) = \sqrt{p}$. The proof of Theorem 33 relies on the projection argument used in Hütter and Rigollet [2016] to derive error bounds for the trend filtering estimator (2.7). For simplicity, in the discussion of our theoretical results, we assume that $\gamma_{\max}(\Sigma)$, n_c and σ^2 are of constant order as n goes to infinity. Recall that n_c is the dimension of $\ker(\Gamma)$, d is the maximum degree of all vertices of G, $L := \Gamma^T\Gamma$, and k_S is defined in (2.11).

Theorem 33 (Main theorem)**.** *Fix $\delta > 0$ and choose $\lambda_1 = 32\sigma\rho(\Gamma)\sqrt{\frac{\gamma_{max}(\Sigma)\log p}{n}}$, $\lambda_2 \leq$*

$\frac{\lambda_1}{8\|\Gamma\beta^*\|_\infty}$. *Given any set* S *satisfying both*

$$\frac{144\gamma_{max}(\Sigma)(\sqrt{n_c}+\delta)^2}{n}+\frac{36\lambda_1^2|S|k_S^{-2}}{\sigma^2}\leq\frac{1}{2}\gamma_{min}\left(\frac{1}{64}\Sigma+\lambda_2 L\right) \tag{2.14}$$

and

$$\lambda_1\|(\Gamma\beta^*)_{-S}\|_1\leq\frac{\sigma^2}{18} \tag{2.15}$$

with probability at least $1-c_1\exp(-nc_2)-\frac{2}{m}-e^{-\delta^2/2}$ *we have*

$$\|\Sigma^{1/2}(\hat{\beta}-\beta^*)\|_2^2\lesssim\frac{\sigma^2\gamma_{max}(\Sigma)}{\gamma_{min}\left(\frac{1}{64}\Sigma+\lambda_2 L\right)}\frac{n_c+\delta^2}{n}+\frac{\lambda_1^2|S|k_S^{-2}}{\gamma_{min}\left(\frac{1}{64}\Sigma+\lambda_2 L\right)}+\lambda_1\|(\Gamma\beta^*)_{-S}\|_1 \tag{2.16}$$

$$\|\hat{\beta}-\beta^*\|_2^2\lesssim\frac{\sigma^2\gamma_{max}(\Sigma)}{\gamma_{min}^2\left(\frac{1}{64}\Sigma+\lambda_2 L\right)}\frac{n_c+\delta^2}{n}+\frac{\lambda_1^2|S|k_S^{-2}}{\gamma_{min}^2\left(\frac{1}{64}\Sigma+\lambda_2 L\right)}+\frac{\lambda_1\|(\Gamma\beta^*)_{-S}\|_1}{\gamma_{min}\left(\frac{1}{64}\Sigma+\lambda_2 L\right)} \tag{2.17}$$

Note that Theorem 33 is actually valid for any matrix Γ. However, if Γ is the incidence matrix of the graph G, we can further bound k_S^{-2} by applying Lemma 3 of Hütter and Rigollet [2016], which states that $k_S^{-2}\lesssim\min(d,|S|)$.

Denote $\beta^*=\beta_1^*+\beta_2^*$, where $\beta_1^*\in\ker(\Gamma)$ and $\beta_2^*\in\ker(\Gamma)^\perp$. Note that the first term in the RHS of (2.17) represents the error from estimating β_1^*, which is the unpenalized component of β^*. The latter two terms represent the error from estimating the penalized component β_2^*, and given a particular graph G we need to further bound $\rho(\Gamma)$ and k_S^{-2} for that graph.

The estimation error bound (2.17) is only different from the prediction error bound (2.16) by a factor of $\gamma_{\min}\left(\frac{1}{64}\Sigma+\lambda_2 L\right)$ in the denominator. This means we have to make a stronger assumption about how fast $\gamma_{\min}\left(\frac{1}{64}\Sigma+\lambda_2 L\right)$ may decay to zero in order to ensure the estimation error, rather than just the prediction error, is also small. For example, when we specialize our bounds for the 3D grid with p vertices, the prediction error bound (2.30) only requires $\gamma_{\min}\left(\frac{1}{64}\Sigma+\lambda_2 L\right)\gg\frac{s\log p}{n}$ but the estimation error bound for this graph

requires $\gamma_{\min}\left(\frac{1}{64}\Sigma + \lambda_2 L\right) \gg \sqrt{\frac{s\log p}{n}}$.

The conditions (2.14) and (2.15) on S are the result of using Lemma 34. They are equivalent to requiring that the RHS of (2.16) is sufficiently small (smaller than $C\sigma^2$ for some absolute constant $C > 0$). Assuming $\gamma_{\min}\left(\frac{1}{64}\Sigma + \lambda_2 L\right)$ is not too small, it is reasonable to expect the prediction error to converge to zero as n become sufficiently large.

Theorem 2.1 is applicable to the estimator $\hat{\beta}_{\mathrm{GL}}$ in (1.50) (which corresponds to setting $\lambda_2 = 0$). However, when $\gamma_{\min}(\Sigma)$ is small, we may not have a meaningful error bound for $\hat{\beta}_{\mathrm{GL}}$. Generally, we want λ_2 to be as large as possible to improve the minimum eigenvalue term without introducing additional errors, and thus the choice $\lambda_2 = \frac{\lambda_1}{8\|\Gamma\beta^*\|_\infty}$ is appropriate. When $\|\Gamma\beta^*\|_\infty$, the maximum signal difference between adjacent vertices, is small (which is reasonable under the assumption of network cohesion on β^*) and $\gamma_{\min}(\Sigma)$ is very close to zero, the improvement of the minimum eigenvalue term can be significant. In contrast, in Theorem 3 of Hebiri and van de Geer [2011] and Theorem 1 of Li et al. [2018], similar proof ideas are used but the dependence between the ℓ_2 and ℓ_1 tuning parameters is such that $\lambda_2 \propto \frac{\lambda_1}{\|L\beta^*\|_\infty}$. Since L is the second-order graph difference operator (see Wang et al. [2015] for the definitions of higher-order total variation operators), the quantity $\|L\beta^*\|_\infty = \|\Gamma^T\Gamma\beta^*\|_\infty$ is not as related to Assumption (2.2) or (2.3) and can be much larger than $\|\Gamma\beta^*\|_\infty$ for graphs with some high-degree nodes. For example, for the star graph with p nodes where the entries of β^* are 0 at the central node and 1 at the leaves, $\|\Gamma\beta^*\|_\infty = 1$ but $\|L\beta^*\|_\infty$ is of order p. The choice of λ_2 in Theorem 33, however, suggests that the regularization effects of the ℓ_2 component of the penalty may be diminished if $\|\Gamma\beta^*\|_\infty$ is large. This is consistent with what we observe in our synthetic experiments: when $\|\Gamma\beta^*\|_\infty$ is large, cross-validation often yields $\lambda_2 \approx 0$.

The proof of Theorem 33 relies on the following lemma to relate the empirical quadratic form $\frac{1}{n}\|Xv\|_2^2$ to the corresponding theoretical quantity $\|\Sigma^{1/2}v\|_2^2$, uniformly for all $v \in \mathbb{R}^p$. This lemma is an extension of the main result in Raskutti et al. [2010] for our setting and

may be of independent interest.

Lemma 34 (Restricted eigenvalue property for random Gaussian design)**.** *If $X \in \mathbb{R}^{n\times p}$ has i.i.d. $N(0,\Sigma)$ rows and $m \geq 2$, $n \geq 10$, then the event*

$$\left\{\forall v \in \mathbb{R}^p : \frac{\|Xv\|_2}{\sqrt{n}} \geq \frac{1}{4}\|\Sigma^{1/2}v\|_2 - 3\sqrt{\frac{\gamma_{max}(\Sigma)n_c}{n}}\|v\|_2 - 6\sqrt{2}\rho(\Gamma)\sqrt{\frac{\gamma_{max}(\Sigma)\log p}{n}}\|\Gamma v\|_1\right\}$$

holds with probability at least $1 - c_1\exp(-nc_2)$, for some universal constants $c_1, c_2 > 0$.

By setting $S = S_{\Gamma\beta^*}$ and applying $k_S^{-2} \lesssim \min(d, |S|)$, we obtain the following bounds which are applicable when β^* is piecewise constant on the graph G. When $\rho(\Gamma) \gtrsim 1$, the second term in (2.18) and (2.19) should dominate.

Corollary 35. *If $\|\Gamma\beta^*\|_0 = s$, with probability at least $1 - c_1\exp(-nc_2) - \frac{2}{m} - e^{-\delta^2/2}$ we have*

$$\|\Sigma^{1/2}(\hat{\beta} - \beta^*)\|_2^2 \lesssim \frac{\sigma^2\gamma_{max}(\Sigma)}{\gamma_{min}\left(\frac{1}{64}\Sigma + \lambda_2 L\right)}\left(\frac{n_c + \delta^2}{n} + \rho^2(\Gamma)\min(d,s)\frac{s\log p}{n}\right) \tag{2.18}$$

$$\|\hat{\beta} - \beta^*\|_2^2 \lesssim \frac{\sigma^2\gamma_{max}(\Sigma)}{\gamma_{min}^2\left(\frac{1}{64}\Sigma + \lambda_2 L\right)}\left(\frac{n_c + \delta^2}{n} + \rho^2(\Gamma)\min(d,s)\frac{s\log p}{n}\right) \tag{2.19}$$

provided that the RHS of (2.18) *is smaller than $C\sigma^2$.*

On the other hand, if we set $S = \emptyset$, we obtain the following bounds that are applicable when $\|\Gamma\beta^*\|_1$ is small. When β^* is smoothly varying over G and $\|\Gamma\beta^*\|_0$ is large, these bounds are more helpful in explaining our estimator's good performance.

Corollary 36. *With probability at least $1 - c_1\exp(-nc_2) - \frac{2}{m} - e^{-\delta^2/2}$,*

$$\|\Sigma^{1/2}(\hat{\beta} - \beta^*)\|_2^2 \lesssim \frac{\sigma^2\gamma_{max}(\Sigma)}{\gamma_{min}\left(\frac{1}{64}\Sigma + \lambda_2 L\right)}\frac{n_c + \delta^2}{n} + \sigma\rho(\Gamma)\sqrt{\frac{\gamma_{max}(\Sigma)\log p}{n}}\|\Gamma\beta^*\|_1 \tag{2.20}$$

$$\|\hat{\beta}-\beta^*\|_2^2 \lesssim \frac{\sigma^2\gamma_{max}(\Sigma)}{\gamma_{min}^2\left(\frac{1}{64}\Sigma+\lambda_2 L\right)}\frac{n_c+\delta^2}{n}+\frac{\sigma\rho(\Gamma)}{\gamma_{min}\left(\frac{1}{64}\Sigma+\lambda_2 L\right)}\sqrt{\frac{\gamma_{max}(\Sigma)\log p}{n}}\|\Gamma\beta^*\|_1 \quad (2.21)$$

provided that the RHS of (2.20) *is smaller than* $C\sigma^2$.

We can also consider the notion that $\Gamma\beta^*$ is ℓ_q-sparse ($0 < q < 1$), in the sense that $\sum_{j=1}^m |(\Gamma\beta^*)_j|^q \le R_q$ (Assumption (2.3)). This notion of weak sparsity has been considered in Raskutti et al. [2011] (where β^* is assumed to lie in an ℓ_q-ball) and Cai and Zhou [2012] (where, in the context of covariance estimation, the columns of the covariance matrix are assumed to lie in an ℓ_q-ball). In contrast, Hebiri and van de Geer [2011] defines the smoothness of the true signal using ℓ_2-norm, in the sense that $\sum_{j=1}^m |(\Gamma\beta^*)_j|^2 \le R_2$ for some $R_2 > 0$. If there exists an edge with a large signal difference, R_2 can be very large. For smaller values of q, we can more easily accommodate the occasional large signal jump with a reasonably small R_q, which appears in the bound (2.22).

By choosing S to trade off the last two terms in the RHS of (2.16), we obtain the following bound for the prediction error. The proof is routine and is thus omitted.

Corollary 37. *With probability at least* $1-c_1\exp(-nc_2)-\frac{2}{m}-e^{-\delta^2/2}$, *if Assumption* (2.3) *holds for some* $q \in (0,1)$, *we have*

$$\begin{aligned}\|\Sigma^{1/2}(\hat{\beta}-\beta^*)\|_2^2 &\lesssim \frac{\sigma^2\gamma_{max}(\Sigma)}{\gamma_{min}\left(\frac{1}{64}\Sigma+\lambda_2 L\right)}\frac{n_c+\delta^2}{n} \\ &+ \min\left(\frac{[\sigma\rho(\Gamma)]^{2-q}\left(\frac{\gamma_{max}(\Sigma)\log p}{n}\right)^{1-q/2}R_q d^{1-q}}{[\gamma_{min}\left(\frac{1}{64}\Sigma+\lambda_2 L\right)]^{1-q}}, \frac{[\sigma\rho(\Gamma)]^{\frac{2}{1+q}}\left(\frac{\gamma_{max}(\Sigma)\log p}{n}\right)^{\frac{1}{1+q}}R_q^{\frac{2}{1+q}}}{\left[\gamma_{min}\left(\frac{1}{64}\Sigma+\lambda_2 L\right)\right]^{\frac{1-q}{1+q}}}\right)\end{aligned} \quad (2.22)$$

provided that the RHS of (2.22) *is smaller than* $C\sigma^2$.

2.2.2 *Discussion of the quantity* $\gamma_{min}\left(\frac{1}{64}\Sigma + \lambda_2 L\right)$

When $\Gamma = I_p$, our penalty is just the original Elastic Net penalty. In that case, since $\rho(\Gamma) = 1$ and $k_S^{-2} \leq 1$, the corresponding estimator $\hat{\beta}_{\text{EN}}$ satisfies with high probability

$$\|\Sigma^{1/2}(\hat{\beta}_{\text{EN}} - \beta^*)\|_2^2 \lesssim \frac{\sigma^2 \gamma_{\max}(\Sigma)}{\gamma_{\min}\left(\frac{1}{64}\Sigma + \lambda_2 I_p\right)} \frac{\|\beta^*\|_0 \log p}{n} \tag{2.23}$$

Here, it is clear the minimum eigenvalue term is bounded below by λ_2. When Γ is an incidence matrix of a graph, however, Γ has a nontrivial kernel and so the behavior of the quantity $\gamma_{\min}\left(\frac{1}{64}\Sigma + \lambda_2 L\right)$ is less clear.

We conjecture that under reasonable assumptions about (Σ, L), $\gamma_{\min}\left(\frac{1}{64}\Sigma + \lambda_2 L\right)$ is bounded below by $c\lambda_2$ for some absolute constant c, at least when λ_2 is in a neighborhood of zero. *We emphasize that the proof of Theorem 33 makes no assumption about how* Σ *is related to the graph* G *or its Laplacian* L. When we only have $\gamma_{\min}\left(\frac{1}{64}\Sigma + \lambda_2 L\right) \geq c\lambda_2$, (2.18) yields

$$\|\Sigma^{1/2}(\hat{\beta} - \beta^*)\|_2^2 \lesssim \sigma\sqrt{\gamma_{\max}(\Sigma)}\|\Gamma\beta^*\|_\infty \left(\frac{n_c + \delta^2}{\rho(\Gamma)\sqrt{n \log p}} + \rho(\Gamma)\min(d, s)s\sqrt{\frac{\log p}{n}}\right) \tag{2.24}$$

but we fail to obtain any theoretical guarantee of consistency in estimation when $\Gamma\beta^*$ is sparse. If we can assume $\gamma_{\min}\left(\frac{1}{64}\Sigma + \lambda_2 L\right) \geq c\sqrt{\lambda_2}$, however, we obtain from (2.19) that

$$\|\hat{\beta} - \beta^*\|_2^2 \lesssim \sigma\sqrt{\gamma_{\max}(\Sigma)}\|\Gamma\beta^*\|_\infty \left(\frac{n_c + \delta^2}{\rho(\Gamma)\sqrt{n \log p}} + \rho(\Gamma)\min(d, s)s\sqrt{\frac{\log p}{n}}\right) \tag{2.25}$$

The bounds (2.24) and (2.25) may be more applicable when Σ is ill-conditioned and $\gamma_{\min}(\Sigma)$ cannot be assumed to be bounded away from zero. Unfortunately, characterizing the spectrum of the sum of two symmetric matrices in terms of the spectra of the summands is known to be a difficult problem, and we leave as an open problem the question

of identifying a reasonable assumption on (Σ, L) (which may both have nontrivial kernels) under which $\gamma_{\min}\left(\frac{1}{64}\Sigma + \lambda_2 L\right) \geq c\lambda_2$ holds. In comparison to our work, Corollary 1 of Hebiri and van de Geer [2011] (which assumes fixed design) assumes that its restricted eigenvalue constant ϕ_{μ_n}, defined with respect to the matrix $\tilde{X}^T\tilde{X}/n$, may be greater than μ_n or $\sqrt{\mu_n}$ without further justification; here μ_n plays a similar role as our λ_2. In order to prove $\gamma_{\min}\left(\frac{1}{64}\Sigma + \lambda_2 L\right) \geq c\lambda_2$, Lemma 1 of Li et al. [2018] assumes that, for some absolute constant $c_l > 0$, $\min_{j\in[p]} \sum_{k=1}^{p} |\Sigma_{jk}| \geq c_l$ and $\max_{j\in[p]} \sum_{k=1}^{p} |\hat{\Sigma}_{jk} - \Sigma_{jk}| \leq c_l/4$; again, note that $\hat{\Sigma}$ acts as the adjacency matrix of the graph considered in Li et al. [2018]. Such assumptions may be too restrictive as the same absolute constant c_l is used in both assumptions.

In Section 2.4.2, we provide empirical evidence to show that, in many situations where the true covariance matrix Σ reflects the structure of the graph G (that is, features indexed by adjacent or nearby nodes are more correlated) and Σ is degenerate, the improvement of the minimum eigenvalue term is significant and can be better than $c\lambda_2$ (or even $c\sqrt{\lambda_2}$).

2.2.3 Error bounds for specific types of graphs

In this section, we apply our results to some specific graph structures that are also explored in Hütter and Rigollet [2016]. Throughout this section, s denotes $\|\Gamma\beta^*\|_0$ and R_q denotes the bound on $\sum_{j=1}^{m} |(\Gamma\beta^*)_j|^q$. We only present prediction error bounds here as the estimation error bounds are different only by a factor of $\gamma_{\min}\left(\frac{1}{64}\Sigma + \lambda_2 L\right)$ in the denominator. We mainly assume $\frac{\sigma^2\gamma_{\max}(\Sigma)}{\gamma_{\min}(\frac{1}{64}\Sigma+\lambda_2 L)}$ is of constant order, but we also specialize the bound (2.24) assuming $\gamma_{\min}\left(\frac{1}{64}\Sigma + \lambda_2 L\right) \gtrsim \lambda_2$ for the case when $\Gamma\beta^*$ is sparse to illustrate the effects of the ℓ_2 component in our penalty when $\gamma_{\min}(\Sigma)$ is very small. In that situation, the bounds for the standalone ℓ_1 penalty provide no control on the errors.

The 2D grid. From Proposition 4 of Hütter and Rigollet [2016] as well as our lower bound

result on $\rho(\Gamma)$ for the 2D grid (proven in the Appendix), we have the following lemma.

Lemma 38. *If Γ is the incidence matrix of the 2D grid with p vertices, then*

$$1 \lesssim \rho(\Gamma) \lesssim \sqrt{\log p}$$

We therefore obtain the following corollary for the 2D grid.

Corollary 39. *Let Γ be the incidence matrix of the 2D grid with p vertices. With the same choice of δ, λ_1 and λ_2 as in Theorem 33, with high probability we have*

$$\|\Sigma^{1/2}(\hat{\beta} - \beta^*)\|_2^2 \lesssim \frac{\sigma^2 \gamma_{max}(\Sigma)}{\gamma_{min}\left(\frac{1}{64}\Sigma + \lambda_2 L\right)} \left(\frac{1+\delta^2}{n} + \frac{s(\log p)^2}{n}\right) \tag{2.26}$$

$$\|\Sigma^{1/2}(\hat{\beta} - \beta^*)\|_2^2 \lesssim \frac{\sigma^2 \gamma_{max}(\Sigma)}{\gamma_{min}\left(\frac{1}{64}\Sigma + \lambda_2 L\right)} \frac{1+\delta^2}{n} + \sigma\sqrt{\frac{\gamma_{max}(\Sigma)(\log p)^2}{n}} \|\Gamma\beta^*\|_1 \tag{2.27}$$

$$\|\Sigma^{1/2}(\hat{\beta} - \beta^*)\|_2^2 \lesssim \frac{\sigma^2 \gamma_{max}(\Sigma)}{\gamma_{min}\left(\frac{1}{64}\Sigma + \lambda_2 L\right)} \frac{1+\delta^2}{n} + \frac{\sigma^{2-q} R_q \left(\frac{\gamma_{max}(\Sigma)(\log p)^2}{n}\right)^{1-q/2}}{[\gamma_{min}\left(\frac{1}{64}\Sigma + \lambda_2 L\right)]^{1-q}} \tag{2.28}$$

provided that the RHS of the bounds above are smaller than $C\sigma^2$. If $\gamma_{min}\left(\frac{1}{64}\Sigma + \lambda_2 L\right) \gtrsim \lambda_2$, we also have

$$\|\Sigma^{1/2}(\hat{\beta} - \beta^*)\|_2^2 \lesssim \sigma\sqrt{\gamma_{max}(\Sigma)} \|\Gamma\beta^*\|_\infty \left(\frac{1+\delta^2}{\sqrt{n \log p}} + \frac{s \log p}{\sqrt{n}}\right) \tag{2.29}$$

The rates obtained in (2.26) and (2.29) are good if s is of small order relative to n. For example, if there is a small island of size k-by-k where β^* attains a value distinct from its background value outside that island (this situation can correspond to finding abnormal spots on an MRI scan), then (2.26) gives us the rate $\frac{k(\log p)^2}{n}$, provided that $\frac{\sigma^2 \gamma_{\max}(\Sigma)}{\gamma_{\min}\left(\frac{1}{64}\Sigma + \lambda_2 L\right)}$ is of constant order. This can be compared with the rate obtained by the Lasso estimator, which is $\frac{k^2 \log p}{n}$ if the background value outside the island is zero (but it fails to achieve this

rate if the background value is nonzero). However, in the situation where the 2D grid can be divided in the middle into a left island and a right island and β^* is constant on each of these islands, then $s \asymp \sqrt{p}$ and our rates are meaningful only in the $p \ll n$ setting.

The r-dimensional grid ($r \geq 3$). From Proposition 6 of Hütter and Rigollet [2016] as well as our lower bound result on $\rho(\Gamma)$ for the r-dimensional grid, we can conclude that $\rho(\Gamma)$ in this case is of constant order, assuming r is fixed.

Lemma 40. *If Γ is the incidence matrix of the r-dimensional grid with p vertices and $r \geq 3$, then*

$$c(r) \leq \rho(\Gamma) \leq C(r)$$

for some constants $c(r), C(r)$ that only depend on r.

We obtain the following corollary for the r-dimensional grid.

Corollary 41. *Let Γ be the incidence matrix of the r-dimensional grid with p vertices, where $r \geq 3$ is fixed. With the same choice of δ, λ_1 and λ_2 as in Theorem 33, with high probability we have*

$$\|\Sigma^{1/2}(\hat{\beta} - \beta^*)\|_2^2 \lesssim \frac{\sigma^2 \gamma_{max}(\Sigma)}{\gamma_{min}\left(\frac{1}{64}\Sigma + \lambda_2 L\right)} \left(\frac{1+\delta^2}{n} + \frac{s \log p}{n}\right) \tag{2.30}$$

$$\|\Sigma^{1/2}(\hat{\beta} - \beta^*)\|_2^2 \lesssim \frac{\sigma^2 \gamma_{max}(\Sigma)}{\gamma_{min}\left(\frac{1}{64}\Sigma + \lambda_2 L\right)} \frac{1+\delta^2}{n} + \sigma \sqrt{\frac{\gamma_{max}(\Sigma) \log p}{n}} \|\Gamma \beta^*\|_1 \tag{2.31}$$

$$\|\Sigma^{1/2}(\hat{\beta} - \beta^*)\|_2^2 \lesssim \frac{\sigma^2 \gamma_{max}(\Sigma)}{\gamma_{min}\left(\frac{1}{64}\Sigma + \lambda_2 L\right)} \frac{1+\delta^2}{n} + \frac{\sigma^{2-q} R_q \left(\frac{\gamma_{max}(\Sigma) \log p}{n}\right)^{1-q/2}}{[\gamma_{min}\left(\frac{1}{64}\Sigma + \lambda_2 L\right)]^{1-q}} \tag{2.32}$$

provided the RHS of the bounds above are smaller than $C\sigma^2$. If $\gamma_{min}\left(\frac{1}{64}\Sigma + \lambda_2 L\right) \gtrsim \lambda_2$, we also have

$$\|\Sigma^{1/2}(\hat{\beta} - \beta^*)\|_2^2 \lesssim \sigma \sqrt{\gamma_{max}(\Sigma)} \|\Gamma \beta^*\|_\infty \left(\frac{1+\delta^2}{\sqrt{n \log p}} + s\sqrt{\frac{\log p}{n}}\right) \tag{2.33}$$

If we consider $r = 3$ and there is a small island of size k-by-k-by-k where β^* attains a value distinct from its background value outside that island, then (2.30) gives us the rate $\frac{k^2 \log p}{n}$, whereas the Lasso gives us the rate $\frac{k^3 \log p}{n}$ if we further assume the background value is zero. This suggests that if the signal is both sparse and smooth over the graph, in some situations using our estimator is preferable to using the Lasso. More generally, if the island is not cubic but rather has an arbitrary shape, $\|\Gamma\beta^*\|_0$ should be the island's surface area, whereas $\|\beta^*\|_0$ should be the island's volume.

The complete graph. As previously mentioned, we can consider regularization with the complete graph when there is no prior structural information available.

Lemma 42. *If Γ is the incidence matrix of the complete graph with p vertices, $\rho(\Gamma) \asymp \frac{1}{p}$.*

Proof. In Proposition 10 of Hütter and Rigollet [2016], replace any '$\leq$' sign with '$=$'. □

If we replace the term $\min(d, s)$ by p (since $d \asymp p$), we obtain the following corollary:

Corollary 43. *Let Γ be the incidence matrix of the complete graph with p vertices. With the same choice of δ, λ_1 and λ_2 as in Theorem 33, with high probability we have*

$$\|\Sigma^{1/2}(\hat{\beta} - \beta^*)\|_2^2 \lesssim \frac{\sigma^2 \gamma_{max}(\Sigma)}{\gamma_{min}\left(\frac{1}{64}\Sigma + \lambda_2 L\right)} \left(\frac{1+\delta^2}{n} + \frac{s \log p}{pn}\right) \tag{2.34}$$

$$\|\Sigma^{1/2}(\hat{\beta} - \beta^*)\|_2^2 \lesssim \frac{\sigma^2 \gamma_{max}(\Sigma)}{\gamma_{min}\left(\frac{1}{64}\Sigma + \lambda_2 L\right)} \frac{1+\delta^2}{n} + \frac{\sigma}{p}\sqrt{\frac{\gamma_{max}(\Sigma) \log p}{n}} \|\Gamma\beta^*\|_1 \tag{2.35}$$

$$\|\Sigma^{1/2}(\hat{\beta} - \beta^*)\|_2^2 \lesssim \frac{\sigma^2 \gamma_{max}(\Sigma)}{\gamma_{min}\left(\frac{1}{64}\Sigma + \lambda_2 L\right)} \frac{1+\delta^2}{n} + \frac{\sigma^{2-q} \frac{R_q}{p} \left(\frac{\gamma_{max}(\Sigma) \log p}{n}\right)^{1-q/2}}{[\gamma_{min}\left(\frac{1}{64}\Sigma + \lambda_2 L\right)]^{1-q}} \tag{2.36}$$

provided that the RHS of the above bounds are smaller than $C\sigma^2$. If $\gamma_{min}\left(\frac{1}{64}\Sigma + \lambda_2 L\right) \gtrsim \lambda_2$,

we also have

$$\|\Sigma^{1/2}(\hat{\beta}-\beta^*)\|_2^2 \lesssim \sigma\sqrt{\gamma_{max}(\Sigma)}\|\Gamma\beta^*\|_\infty\left(\frac{p(1+\delta^2)}{\sqrt{n\log p}}+s\sqrt{\frac{\log p}{n}}\right) \tag{2.37}$$

In the case when the signal takes $k \ll p$ different values, with $k-1$ of those attained on small islands of size $l \ll p$, s is of order klp, and (2.34) yields the rate $\frac{kl\log p}{n}$, provided that $\frac{\sigma^2\gamma_{\max}(\Sigma)}{\gamma_{\min}\left(\frac{1}{64}\Sigma+\lambda_2 L\right)}$ is of constant order. This is the same as the rate we obtain for the Lasso if the complement of the small islands has value zero. However, if there are two large components with two different values, s is of order p^2 and so (2.34) only guarantees some control when $p \ll n$. If $\gamma_{\min}\left(\frac{1}{64}\Sigma+\lambda_2 L\right)$ is of order λ_2, then (2.37) only gives us a meaningful bound when $p \ll \sqrt{n}$, provided that $\|\Gamma\beta^*\|_\infty$ is of constant order.

The star graph. Here, we consider the graph with p nodes and with one center node connected to $p-1$ leaves. A similar penalty has been considered by Ollier and Viallon [2017] to model stratified data, and this penalty is useful particularly when most outer nodes share the same value as the central node.

Lemma 44. *If Γ is the incidence matrix of the star graph with p vertices, then $\rho(\Gamma) \asymp 1$.*

Proof. From Proposition 12 in Hütter and Rigollet [2016], any column s_j of $\Gamma^\dagger$ has $\|s_j\|_2^2 = 1-\frac{1}{p}$. □

Corollary 45. *Let Γ be the incidence matrix of the star graph with p vertices. With the same choice of δ, λ_1 and λ_2 as in Theorem 33, with high probability we have*

$$\|\Sigma^{1/2}(\hat{\beta}-\beta^*)\|_2^2 \lesssim \frac{\sigma^2\gamma_{max}(\Sigma)}{\gamma_{min}\left(\frac{1}{64}\Sigma+\lambda_2 L\right)}\left(\frac{1+\delta^2}{n}+\frac{s^2\log p}{n}\right) \tag{2.38}$$

$$\|\Sigma^{1/2}(\hat{\beta}-\beta^*)\|_2^2 \lesssim \frac{\sigma^2\gamma_{max}(\Sigma)}{\gamma_{min}\left(\frac{1}{64}\Sigma+\lambda_2 L\right)}\frac{1+\delta^2}{n}+\sigma\sqrt{\frac{\gamma_{max}(\Sigma)\log p}{n}}\|\Gamma\beta^*\|_1 \tag{2.39}$$

$$\|\Sigma^{1/2}(\hat{\beta}-\beta^*)\|_2^2 \lesssim \frac{\sigma^2\gamma_{max}(\Sigma)}{\gamma_{min}\left(\frac{1}{64}\Sigma+\lambda_2 L\right)}\frac{1+\delta^2}{n}+\frac{\left(\frac{\sigma^2\gamma_{max}(\Sigma)R_q^2\log p}{n}\right)^{\frac{1}{1+q}}}{[\gamma_{min}\left(\frac{1}{64}\Sigma+\lambda_2 L\right)]^{\frac{1-q}{1+q}}} \tag{2.40}$$

provided that the RHS of the above bounds are smaller than $C\sigma^2$. *If* $\gamma_{min}\left(\frac{1}{64}\Sigma+\lambda_2 L\right) \gtrsim \lambda_2$, *we also have*

$$\|\Sigma^{1/2}(\hat{\beta}-\beta^*)\|_2^2 \lesssim \sigma\sqrt{\gamma_{max}(\Sigma)}\|\Gamma\beta^*\|_\infty\left(\frac{1+\delta^2}{\sqrt{n\log p}}+s^2\sqrt{\frac{\log p}{n}}\right) \tag{2.41}$$

For the star graph, we obtain meaningful bounds for the prediction error, even in the high-dimensional setting where $p \gg n$.

The chain graph. When Γ is the p-vertex chain graph (1D grid graph), $\rho(\Gamma)=\sqrt{p}$ and Theorem 33 does not yield an error bound that is meaningful in the $p \gg n$ setting. We modify the proof of Theorem 33 using an idea in Theorem 6 of Wang et al. [2015] to obtain the following bound when $\|\Gamma\beta^*\|_1$ is small.

Theorem 46. *Let* Γ *be the incidence matrix of the* p*-vertex chain graph, and fix* $\delta > 0$. *With an appropriate choice of* λ_1 *and* $\lambda_2 \le \frac{\lambda_1}{8\|\Gamma\beta^*\|_\infty}$, *with high probability we have*

$$\|\Sigma^{1/2}(\hat{\beta}-\beta^*)\|_2^2 \lesssim \frac{\sigma^2\gamma_{max}(\Sigma)}{\gamma_{min}(\frac{1}{64}\Sigma+\lambda_2 L)}\frac{1+\delta^2}{n}+\frac{(\sigma^2\gamma_{max}(\Sigma)\|\Gamma\beta^*\|_1)^{2/3}}{\gamma_{min}^{1/3}\left(\frac{1}{64}\Sigma+\lambda_2 L\right)}\sqrt[3]{\frac{p\log p}{n^2}} \tag{2.42}$$

provided that the bound above is smaller than $C\sigma^2$.

The bound above is meaningful when $n \gg \sqrt{p\log p}$ and thus sufficient to justify the use of our estimator when Γ is the chain graph. Optimal error bounds under the assumption of hard sparsity on $\Gamma\beta^*$ are available in the literature if X is identity (see for example Ortelli and van de Geer [2021] and Guntuboyina et al. [2020]). However, such bounds are often derived under a "minimum length" condition, which requires that the distances between jumps for

the true signal are roughly of the same order. The bound (2.42), on the other hand, requires minimal assumptions. We leave open for future work the analysis of our estimator (2.4) under the assumption of hard sparsity on $\Gamma\beta^*$.

2.3 Computation

In this section, we describe our coordinate descent procedure to compute the estimator (2.4). For convenience, we will work with the following definition of $\hat{\beta}$ where we replace the loss $\frac{1}{n}\|Y - X\beta\|_2^2$ in (2.4) by $\frac{1}{2}\|Y - X\beta\|_2^2$. Note that this simply corresponds to a different scaling of λ_1 and λ_2.

$$\hat{\beta} := \arg\min_{\beta \in \mathbb{R}^p} \frac{1}{2}\|Y - X\beta\|_2^2 + \lambda_1\|\Gamma\beta\|_1 + \lambda_2\|\Gamma\beta\|_2^2 \tag{2.43}$$

Again, let $\tilde{Y} := \begin{pmatrix} Y \\ 0 \end{pmatrix} \in \mathbb{R}^{n+m}$, $\tilde{X} := \begin{pmatrix} X \\ \sqrt{2\lambda_2}\Gamma \end{pmatrix} \in \mathbb{R}^{(n+m)\times p}$ so that we can write

$$\hat{\beta} := \arg\min_{\beta \in \mathbb{R}^p} \frac{1}{2}\|\tilde{Y} - \tilde{X}\beta\|_2^2 + \lambda_1\|\Gamma\beta\|_1 \tag{2.44}$$

If we fix λ_2, the solution path in terms of λ_1 for (2.44) as well as its dual objective is piecewise linear, and a path-finding algorithm for the dual objective yielding the entire solution path in terms of λ_1 has been proposed in Tibshirani and Taylor [2011]. However, for the purpose of selecting tuning parameters, this is of limited usefulness since λ_2 needs to be fixed. The solution path in terms of λ_2 is not piecewise linear, and so we cannot use a LARS-like algorithm to get the entire path in terms of both (λ_1, λ_2). Also, as mentioned in Tibshirani and Taylor [2011], the set of knots in the solution path becomes very large as the problem size increases, and at each knot we must solve a large least squares problem (especially at the regularized end of the path, which is typically the region of interest in this

paper) in order to compute the whole path. If we want to compute (2.44) for a small set of candidate (λ_1, λ_2) values, then the path algorithm in Tibshirani and Taylor [2011] is unlikely to be the most efficient.

2.3.1 Coordinate descent on the dual objective

Our coordinate descent algorithm builds upon the dual problem derived in Tibshirani and Taylor [2011] for the Generalized Lasso. Tibshirani and Taylor [2011] suggests, without explicit derivations, that we can use coordinate descent on the dual problem to compute the solution of (2.44) for a fixed value of (λ_1, λ_2). Coordinate descent cannot be directly applied to the primal objective (2.44) as the ℓ_1-penalty here is not separable in terms of β; in such a situation, coordinate descent does not necessarily converge. However, the dual objective (2.47) has a non-smooth component that is separable, and thus convergence is guaranteed (since conditions (A1), (B1)-(B3) and (C2) from Tseng [2001] hold). For completeness, we fully derive this coordinate descent algorithm on the dual and provide experiments to convince the reader that our estimator can be efficiently computed.

Define $\check{Y} := \tilde{X}\tilde{X}^\dagger\tilde{Y} \in \mathbb{R}^{m+n}$, $\check{\Gamma} := \Gamma\tilde{X}^\dagger \in \mathbb{R}^{m\times(m+n)}$. From Equation (36) of Tibshirani and Taylor [2011], the dual problem is:

$$\hat{u} = \arg\min_{u\in\mathbb{R}^m} \frac{1}{2}\|\check{Y} - \check{\Gamma}^T u\|_2^2 \quad \text{subject to } \|u\|_\infty \le \lambda_1, \Gamma^T u \in \text{row}(\tilde{X}) \tag{2.45}$$

and the primal-dual relation, as in Equation (37) of Tibshirani and Taylor [2011], is:

$$\hat{\beta} = \tilde{X}^\dagger(\check{Y} - \check{\Gamma}^T\hat{u}) + z \tag{2.46}$$

where $z \in \ker(\tilde{X})$. In most situations, the augmented matrix $\tilde{X} := \begin{pmatrix} X \\ \sqrt{2\lambda_2}\Gamma \end{pmatrix}$ has a trivial kernel, in which case $\text{row}(\tilde{X}) = \mathbb{R}^p$ and we can ignore z as well as the constraint

$\Gamma^T u \in \text{row}(\tilde{X})$. Now if we let $Q := \check{\Gamma}\check{\Gamma}^T \in \mathbb{R}^{m \times m}$ and $b := \check{\Gamma}\check{Y} \in \mathbb{R}^m$, then we can write the dual objective as:

$$\hat{u} = \arg\min_{u \in \mathbb{R}^m} \frac{1}{2} u^T Q u - b^T u \quad \text{subject to } \|u\|_\infty \leq \lambda_1 \tag{2.47}$$

We denote the projection map from $\mathbb{R}$ onto $[-\lambda, \lambda]$ by $T_\lambda(\cdot)$:

$$T_\lambda(x) := \begin{cases} \lambda & \text{if } x > \lambda \\ x & \text{if } -\lambda \leq x \leq \lambda \\ -\lambda & \text{if } x < -\lambda \end{cases} \tag{2.48}$$

Our coordinate descent algorithm is presented below.

Algorithm 1: Coordinate descent on the dual objective

Input: $\lambda_1, \lambda_2, \Gamma, Y, X$, tolerance ϵ

Output: $\hat{\beta}$ as defined in (2.43)

Compute $Q = \check{\Gamma}\check{\Gamma}^T = (\Gamma \tilde{X}^\dagger)(\Gamma \tilde{X}^\dagger)^T$ and $b = \check{\Gamma}\check{Y} = \Gamma \tilde{X}^\dagger \tilde{Y}$

Initialize $\hat{u}_i^{(0)} \leftarrow 0$ for all $i \in [m]$

while $\|\hat{u}^{(k)} - \hat{u}^{(k-1)}\|_2 > \epsilon$ **do**

  $\hat{u}_i^{(k+1)} \leftarrow T_{\lambda_1}\left(\frac{b_i - \sum_{j<i} Q_{ij} \hat{u}_j^{(k+1)} - \sum_{j>i} Q_{ij} \hat{u}_j^{(k)}}{Q_{ii}} \right)$

Compute $\hat{\beta} \leftarrow \tilde{X}^\dagger (\check{Y} - \check{\Gamma}^T \hat{u})$

Return $\hat{\beta}$

For general GLM loss functions, we can also derive the dual problem with a separable non-smooth constraint; however, we may not be able to write the coordinate descent updates in closed form (we can only do so in Algorithm 1 because the dual objective (2.47) is quadratic). In this case, we can use *coordinate proximal gradient descent*, in which we apply the projection operator to the gradient descent update for each coordinate.

In the Appendix, we also provide an alternative algorithm to compute (2.43), based on the interior point method applied to the dual objective (as in Kim et al. [2009]). This

algorithm will be denoted as IP in the following section.

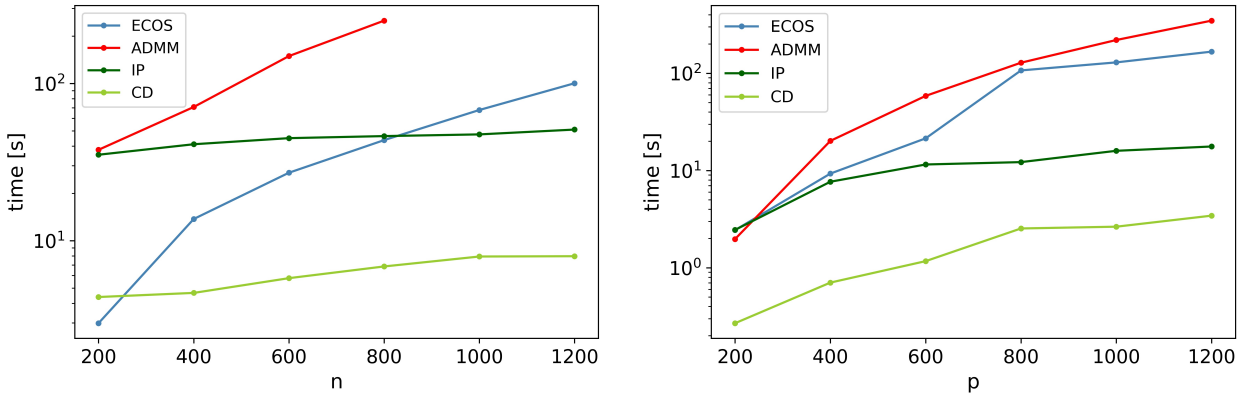
2.3.2 Runtime comparisons

We compare the runtimes for computing the estimator (2.4) using Algorithm 1 (CD), IP, ADMM, and the Embedded Conic Solver (ECOS) from Domahidi et al. [2013] applied to the primal objective. ECOS is a generic solver for second-order cone programs (SOCP) that performs well for small or medium-sized problems. We use the highly optimized ECOS implementation in the Python package CVXPY to serve as a benchmark for comparing the runtimes of our algorithms. Figure 2.1 shows the growth of empirical runtimes as n or p increases for signals over the chain graph (where $m = p - 1$) with $\|\Gamma\beta^*\|_\infty = 0.3$ fixed; here, the hyperparameters λ_1, λ_2 are chosen according to our theory so as to satisfy $\lambda_2 = \frac{\lambda_1}{8\|\Gamma\beta^*\|_\infty}$.

As we can see from Figure 2.1, our coordinate descent algorithm scales well as n and p increase, and its runtime does not exceed 10 seconds if n and p are both smaller than 1,000. More generally, when λ_2 is not too close to zero, the matrix $Q = \check{\Gamma}\check{\Gamma}^T$ is not ill-conditioned and our coordinate descent algorithm performs quite well. We note that this is the setting where our estimator (2.4) should be preferred over the Generalized Lasso estimator $\hat{\beta}_{\mathrm{GL}}$ in (1.50), whose accuracy is impeded by the ill-conditioned nature of the matrix Q when λ_2 is equal to zero. While our estimator requires a two-dimensional grid search to choose (λ_1, λ_2), Algorithm 1 can significantly reduce the time it takes to perform hyperparameter tuning, even for large-scale problems where p and n are both in the thousands. Note that when both n and p are not too large, the generic SOCP solver ECOS can also be competitive.

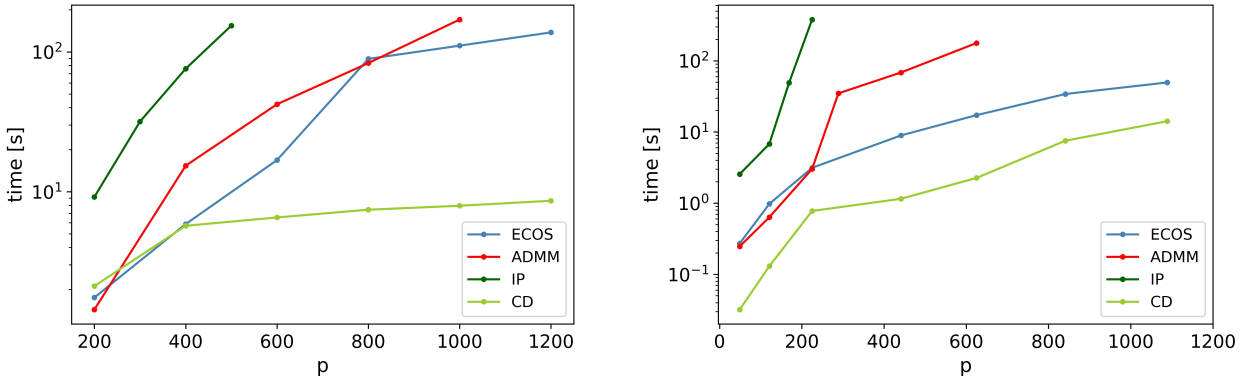
As for our interior point method, the main computational bottleneck is the cost of solving a linear equation involving the Hessian matrix; in other words, we need to solve the problem $Ax = b$ for each iteration, where A is an m-by-m matrix. Solving it requires $O(m^3)$ operations, and thus IP can do well only if the number of iterations required is small. Figure 2.1(b) shows that in the case of the chain graph, when we fix n and increase p, IP still

performs better than the generic solver ECOS and scales well with p.



(a) $p = 2000$ (b) $n = 2000$

Figure 2.1: Runtimes of different algorithms (reported on the log scale) when (a) p is fixed but n increases, or (b) n is fixed but p increases. The tolerance levels for IP, CD, and ECOS are set at 10^{-4}. The tolerance level for ADMM is 10^{-3}. Signals are defined on a 1D chain graph with p vertices. In both situations, CD has the best runtime scaling, and IP scales better than ECOS.



(a) $n = 2000$, 2D grid (b) $n = 2000$, star graph

Figure 2.2: Runtimes of different algorithms (reported on the log scale) when n is fixed but p increases. (a) Signals are defined on a p-vertex 2D grid graph ($m = 2p - 2\sqrt{p}$) with $\|\Gamma\beta^*\|_\infty = 0.66$. (b) Signals are defined on a p-vertex star graph ($m = p - 1$) with $\|\Gamma\beta^*\|_\infty = 0.5$. The tolerance levels for IP, CD, and ECOS are set at 10^{-4}. The tolerance level for ADMM is 10^{-3}. As before, (λ_1, λ_2) are chosen according to theory. In both situations, CD has the best runtime scaling.

We also examine the runtimes for the 2D grid as well as the star graph when n is fixed but p increases. For these graphs, IP no longer scales well with p whereas CD still has the

best scaling, and ECOS is also competitive for small problem sizes.

2.4 Experiments

In this section, we present the empirical performance of our penalty $\lambda_1\|\Gamma\beta\|_1 + \lambda_2\|\Gamma\beta\|_2^2$ and compare with some existing penalized M-estimators in the literature, under several synthetic settings where we vary the true signal structure and graph topology. Particularly, we focus on the case where β^* is not sparse but aligns with the graph G. The design matrix is also allowed to be correlated in a way such that two vertices have more correlated feature vectors if they are adjacent or nearby on the graph G. Such a covariance structure is natural for node-indexed feature vectors and is in line with the notion of *network cohesion* discussed in Section 2.1. We list the methods to which we compare our estimator below.

Graph-independent methods that do not take into account the graph provided. These methods usually do not perform well in the setting we describe above, and they mainly serve as benchmarks for comparison.

1. The *ordinary least squares* (OLS) estimator, which is a standard method in the setting when $p < n$ and the underlying signal is dense. It often does not perform well when we are in the high-dimensional setting ($p > n$) or the design is highly correlated and $\gamma_{\min}(\Sigma)$ is close to zero.

2. The *Lasso* (L) penalty $\lambda_L\|\beta\|_1$ from Tibshirani [1996], which can perform well in the $p \gg n$ setting if the true signal is known to be sparse. In the $p > n$ case, however, it has been shown to select at most n variables before it saturates. As discussed in Zou and Hastie [2005], the Lasso lacks the ability to select groups of correlated variables, and it is empirically observed to suffer from unstable selections in the presence of high correlation between features.

3. The *Elastic Net* (EN) penalty $\lambda_L\|\beta\|_1 + \lambda_E\|\beta\|_2^2$, which was developed in Zou and Hastie [2005] to deal with highly correlated predictors. The Elastic Net tends to encourage strongly correlated predictors to be in or out of the model together while also preserving sparsity of representation like the Lasso. It is a suitable candidate in our setting due to our assumption of highly correlated design.

Graph-based methods that utilizes information from the given graph G (except for possibly the GTV method). We have described these methods in Section 2.1.3.

4. The *Fused Lasso* (FL) penalty $\lambda_1\|\Gamma\beta\|_1 + \lambda_L\|\beta\|_1$ proposed in Tibshirani et al. [2005] encourages the resulting estimate to be both sparse and piecewise constant with respect to G. This penalty may be suitable if we believe the true signal is sparse and also forms clusters on G (that is, in each cluster the true signal attains a single value). When the true signal is not sparse, the tuning parameter λ_L is often set to zero if we use cross-validation (CV) for hyperparameter selection, and FL degenerates into our GEN penalty with $\lambda_2 = 0$.

5. The *Smooth Lasso* (SL) penalty $\lambda_2\|\Gamma\beta\|_2^2 + \lambda_L\|\beta\|_1$ in Hebiri and van de Geer [2011] results in an estimate that is smooth, in the sense that $\|\Gamma\hat{\beta}_{\text{SL}}\|_\infty$ is small. It is useful when β^* is sparse and we also believe $\|\Gamma\beta^*\|_\infty$ is small. When the true signal is not sparse and we use CV for hyperparameter selection, λ_L for SL is often set to zero, in which case SL also degenerates into our GEN penalty with $\lambda_1 = 0$.

6. The *Graph Total Variation* (GTV) penalty $\lambda_1\|\hat{\Gamma}\beta\|_1 + \lambda_2\|\hat{\Gamma}\beta\|_2^2 + \lambda_L\|\beta\|_1$ in Li et al. [2018] estimates Σ with some covariance estimator $\hat{\Sigma}$ and then treats $\hat{\Sigma}$ as the weighted adjacency matrix of some graph $\hat{G}$ with incidence matrix $\hat{\Gamma}$. In our experiments, as suggested by Li et al. [2018], the estimator $\hat{\Sigma}$ is obtained by hard-thresholding the sample covariance matrix (see Bickel and Levina [2008] for details). This choice of $\hat{\Sigma}$ means that we also need to tune the covariance threshold in addition to the 3

hyperparameters that appear in the GTV penalty. In general, however, $\hat{\Sigma}$ can be any covariance estimator and can also incorporate side information such as the graph G provided in our setting.

7. We also denote by *GTV-oracle* the GTV penalty based on using the unobserved covariance matrix Σ (rather than $\hat{\Sigma}$) to construct the corresponding incidence matrix $\hat{\Gamma}_{\text{oracle}}$. Using the true covariance matrix should eliminate any error from covariance estimation. However, if all entries of Σ are nonzero, computation of the GTV-oracle estimator can be especially challenging, since the graph used in the GTV penalty here is a weighted complete graph.

2.4.1 Experiments on synthetic data

We repeatedly generate training and testing data from the model $y = X\beta^* + \epsilon$, where the rows of X are generated i.i.d. from $N(0, \Sigma)$ and independent of ϵ which is generated from $N(0, \sigma^2 I_n)$. Hyperparameter selection via CV is performed using a separate validation data set. We report the estimation error $\|\hat{\beta} - \beta^*\|_2$, as well as the prediction error $\frac{1}{n}\|X_{\text{test}}(\hat{\beta} - \beta^*)\|_2^2$ computed using the testing data.

2.4.1.1 Choices of Σ and the graph G

We consider the chain graph, the 2D grid graph and the barbell graph in our experiments. The first two graphs allow for easier visualization of the true and estimated signals defined on them. The barbell graph, which consists of two non-overlapping cliques connected by a single path that has an endpoint in each clique, allows us to test the performance of our method on a denser graph with a less homogenous degree distribution.

As previously mentioned, Σ is constructed so that nearby nodes have more correlated feature vectors. For the chain graph, we use the Toeplitz covariance structure with $\Sigma_{ij} = \rho^{|i-j|}$ where, if not stated otherwise, we typically choose $\rho = 0.5$ (moderate correlation). For

the 2D grid and barbell graph, we construct Σ by inverting the matrix $L + 0.5I_p$ (recall that L denotes the Laplacian of the graph G) and then normalize Σ so that all covariates have unit variance. The resulting covariance matrices obtained from this process are illustrated in Figure 2.3.

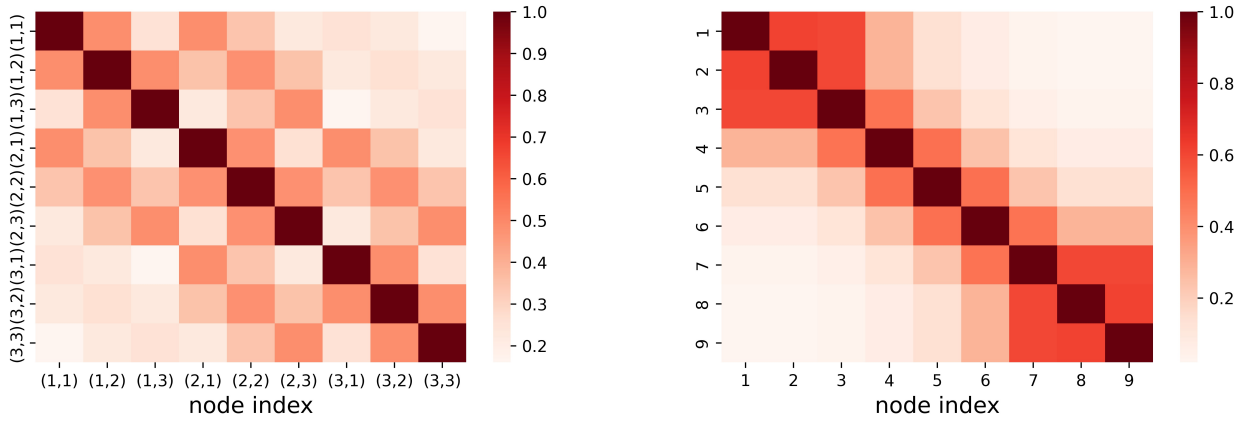


Figure 2.3: Left: the covariance matrix obtained for a 2D grid graph with $p = 3 \times 3$ vertices. Right: the covariance matrix obtained for a barbell graph with two cliques $\{1, 2, 3\}$ and $\{7, 8, 9\}$ connected by the path $\{3, 4, 5, 6, 7\}$. Note that correlation is higher for adjacent or nearby vertices.

2.4.1.2 Hyperparameter selection and tuning time

We select hyperparameters based on 5-fold CV using a fine grid search, where each hyperparameter is chosen from a list of at least 20 values. The scorer for CV is the negative mean squared error (MSE) $-\frac{1}{n}\|Y - X\hat{\beta}\|_2^2$, which tends to select for hyperparameters with better prediction performance.

Hyperparameter tuning for GEN is computationally manageable. When G is a chain graph, the tuning time for GEN is comparable to that of other methods with two hyperparameters, namely the Elastic Net, the Fused Lasso and the Smooth Lasso. If we disregard the covariance thresholding parameter, the GTV penalty still involves three hyperparameters, and the graph $\hat{\Gamma}$ used in the GTV penalty, computed using the covariance estimate $\hat{\Sigma}$, has more nonzero weighted edges compared to the given graph Γ. These factors contribute to

longer tuning time for the GTV penalty. The tuning time is much worse for the GTV-oracle penalty since the true covariance matrix Σ is denser than $\hat{\Sigma}$, and so $\hat{\Gamma}_{\text{oracle}}$ has many more nonzero weighted edges than $\hat{\Gamma}$ does. We present the tuning times for a toy example where each hyperparameter is selected from a small grid search in Table 2.1. Here, since n and p are both not too large, we use the SOCP solver ECOS (see Section 2.3.2) for all methods.

Table 2.1: Tuning times with ECOS when G is the chain graph. $p = 110, m = 109, n = 210, \sigma = 1$, and Σ is constructed as in Section 2.4.1.1. The GTV penalty is based on $\hat{\Gamma}$ which has around 200 nonzero weighted edges. The GTV-oracle penalty is based on $\hat{\Gamma}_{\text{oracle}}$ which has almost 6000 nonzero weighted edges. 5-fold CV is performed for each method on a small grid with 5 candidate values $[0, 0.1, 1, 10, 100]$ for each hyperparameter.

	L	**EN**	**FL**	**SL**	**GTV**	**GTV-oracle**	**GEN**
# **hyperparameters**	1	2	2	2	3	3	2
time [seconds]	0.45	3.22	2.85	2.30	14.31	134.08	2.46

When the graph G contains more edges, we can expect the tuning time for GEN to increase relative to other two-hypterparameter methods, as both the ℓ_1 and ℓ_2 components of the GEN penalty depend on Γ. Table 2.2 repeats the above experiment but with G being the barbell graph and Σ reflecting the structure of this graph. The tuning time with ECOS for GEN is roughly double that of FL or SL, whose penalties contain only one component depending on Γ.

Table 2.2: Tuning times with ECOS when G is the barbell graph. $p = 110, m = 2461, n = 210, \sigma = 1$, and Σ is constructed as in Section 2.4.1.1. The GTV penalty is based on $\hat{\Gamma}$ which has around 2500 nonzero weighted edges. As Σ for the barbell graph is denser than Σ in Table 2.1, $\hat{\Sigma}$ here is also denser than $\hat{\Sigma}$ in Table 2.1.

	L	**EN**	**FL**	**SL**	**GTV**	**GTV-oracle**	**GEN**
# **hyperparameters**	1	2	2	2	3	3	2
time [seconds]	0.44	2.51	4.62	4.57	52.47	131.91	9.68

2.4.1.3 Comparisons between GEN, FL and SL when β^* is dense but aligns with G

In this section, we focus on the case when β^* is not sparse but $\Gamma\beta^*$ is sparse or β^* is otherwise smooth with respect to G. As all parts of the true signals constructed in this section are far from zero, the component $\lambda_L\|\beta\|_1$ in the FL and SL penalties is of little use, and setting $\lambda_L > 0$ worsens both prediction and estimation errors in this setting. Consequently, CV yields λ_L values that are almost identically zero for both FL and SL. Essentially, in this section, FL refers to the standalone $\lambda_1\|\Gamma\beta\|_1$ penalty, whereas SL refers to the standalone $\lambda_2\|\Gamma\beta\|_2^2$ penalty.

We observe that FL performs well when β^* has few signal jumps on G, regardless of whether there exists large jumps (i.e. $\|\Gamma\beta^*\|_\infty$ is large). SL, on the other hand, tends to perform well when the signal is smooth with respect to G, in the sense that $\|\Gamma\beta^*\|_\infty$ is small, even if the number of signal jumps $\|\Gamma\beta^*\|_0$ might be large. To demonstrate these observations, we construct signals with varying smoothness ($\|\Gamma\beta^*\|_\infty$) and numbers of jumps ($\|\Gamma\beta^*\|_0$). Figure 2.4 illustrates the true signals on the 1D chain graph, whereas Figure 2.5 illustrates the true signals on the 2D grid graph (note that p is fixed for these graphs). For the barbell graph, we let the signal values be constant (at 5 and 20 respectively) on each clique. The lengths of the path connecting the two cliques are chosen from $\{1, 4, 7, 10, 13, 16\}$ (and so p has to vary), and we let the signal decrease from 20 to 5 gradually on the connecting path, so that $\|\Gamma\beta^*\|_\infty$ decreases from 15 to 1.46 while $\|\Gamma\beta^*\|_0$ increases from 1 to 16.

Figure 2.6 illustrates the performances of FL, SL and GEN in terms of estimation and prediction errors. When $\|\Gamma\beta^*\|_\infty$ is small, CV yields λ_2 values that are larger relative to λ_1, which is consistent with our theory in that λ_2 can be chosen up to $\frac{C\lambda_1}{\|\Gamma\beta^*\|_\infty}$ without incurring additional errors. As can be seen from Figure 2.6, our GEN penalty adapts well to true signals of various smoothness levels, thus demonstrating the importance of having both the $\lambda_1\|\Gamma\beta\|_1$ and $\lambda_2\|\Gamma\beta\|_2^2$ components in our penalty. From the performances of FL and

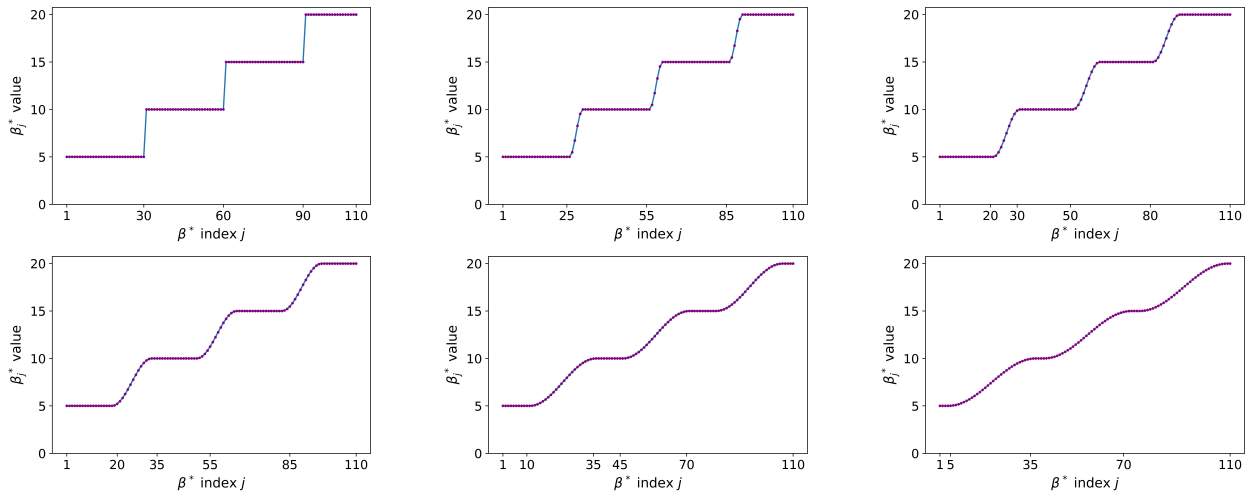


Figure 2.4: True signals defined on the chain graph with $p = 110$. The top left signal is piecewise constant and has the smallest $\|\Gamma\beta^*\|_0 = 3$ but the largest $\|\Gamma\beta^*\|_\infty = 5$. The bottom right signal is the smoothest with the largest $\|\Gamma\beta^*\|_0 = 99$ and the smallest $\|\Gamma\beta^*\|_\infty = 0.24$. The intermediate signals are constructed such that $\|\Gamma\beta^*\|_0$ decreases but $\|\Gamma\beta^*\|_\infty$ increases gradually. All 6 signals have $\|\Gamma\beta^*\|_1 = 15$.

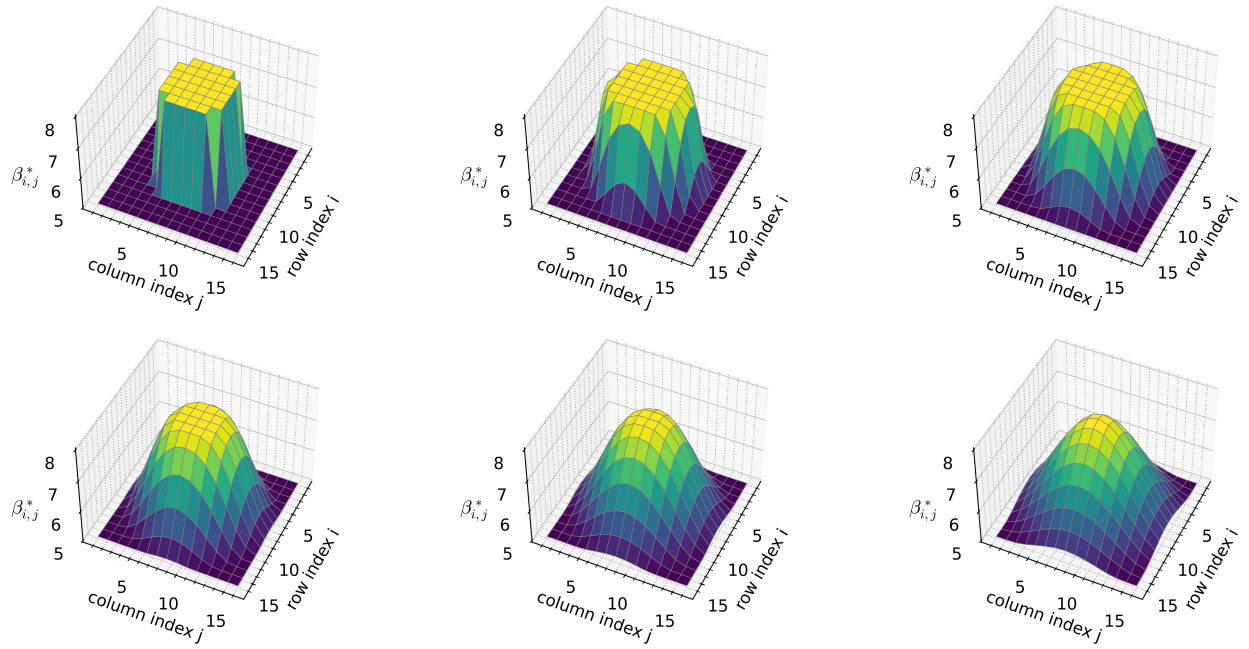
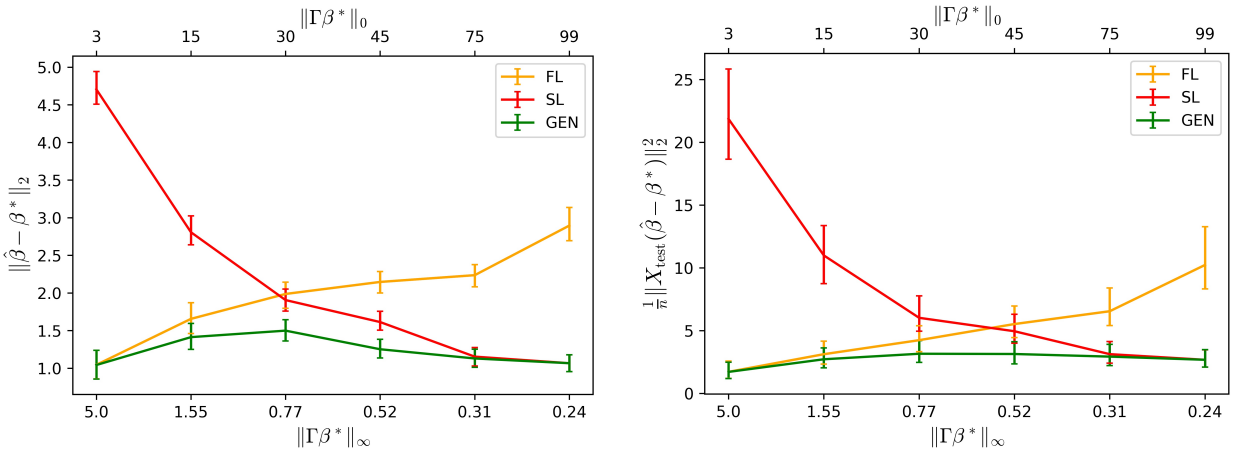
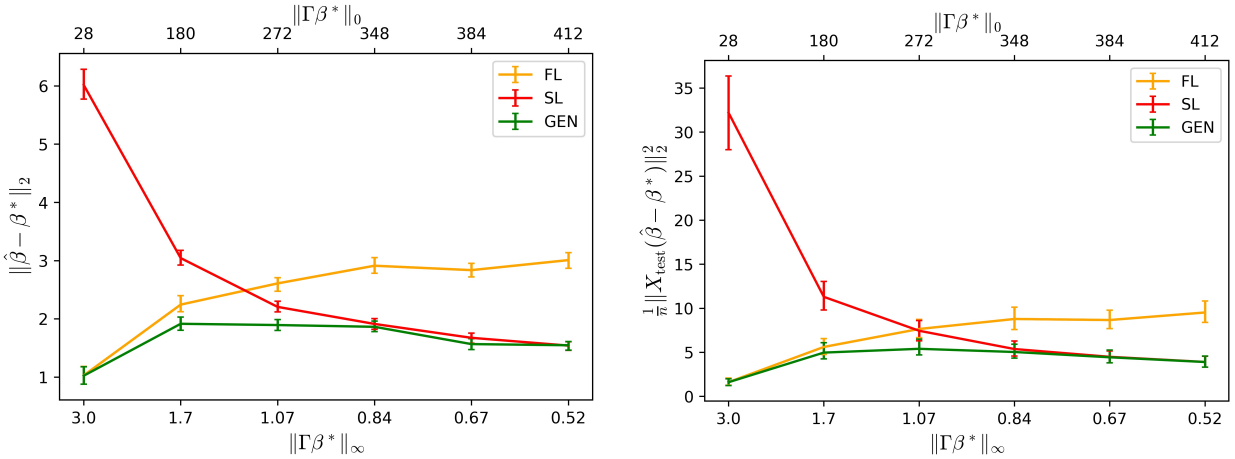


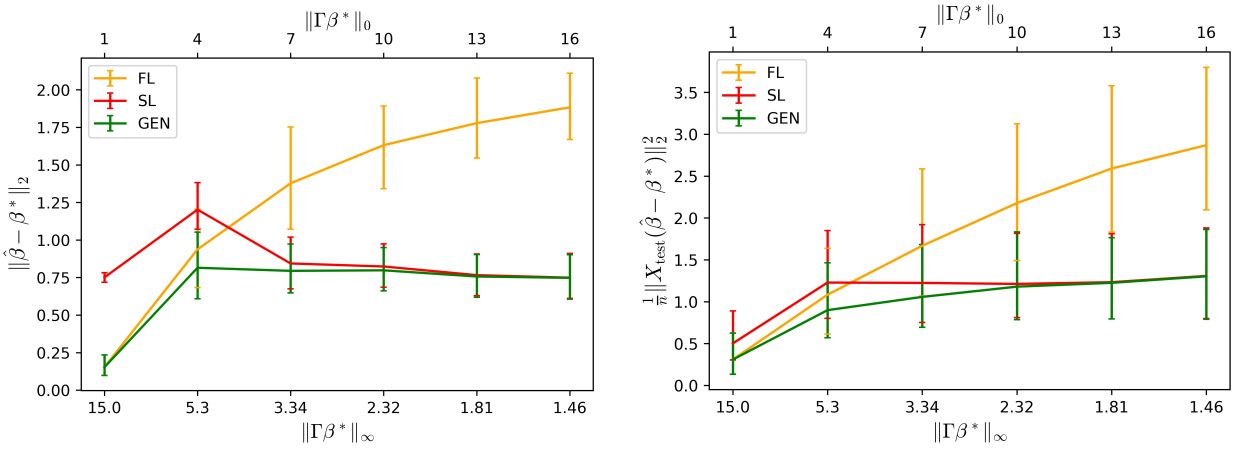
Figure 2.5: True signals defined on the 2D grid with $p = 15 \times 15$. The top left signal is piecewise constant and has the smallest $\|\Gamma\beta^*\|_0 = 28$ but the largest $\|\Gamma\beta^*\|_\infty = 3$. The bottom right signal is the smoothest with the largest $\|\Gamma\beta^*\|_0 = 412$ and the smallest $\|\Gamma\beta^*\|_\infty = 0.24$. All 6 signals have $\|\Gamma\beta^*\|_1$ between 84 and 120.



(a) 1D chain



(b) 2D grid



(c) Barbell

Figure 2.6: Prediction and estimation errors for three graphs as $\|\Gamma\beta^*\|_\infty$ and $\|\Gamma\beta^*\|_0$ vary. Results are based on 500 resamplings. Vertical bars for each true signal connect the 25^{th} and 75^{th} percentiles. The lines labeled by FL, SL and GEN connect the medians of errors.

SL, we can see that the $\lambda_1\|\Gamma\beta\|_1$ penalty ensures good performance when the true signal is piecewise constant, whereas the $\lambda_2\|\Gamma\beta\|_2^2$ penalty ensures good performance when the true signal is smooth over G.

Note again that FL and SL in this setting correspond to the GEN penalty with λ_2 or λ_1 set to zero, respectively. Therefore, GEN's superior performance over FL and SL in terms of prediction error is not surprising, given that the scorer used for CV $-\frac{1}{n}\|Y - X\hat{\beta}\|_2^2$ selects for hyperparameters with stronger prediction performance. However, GEN is also consistently better than FL or SL in terms of estimation error. This can be understood better by examining the signal estimates obtained from the three procedures. Figure 2.7 compare the estimated signals with the true signals defined on the 1D chain graph. FL recovers the constant regions well but struggles with the smoothly increasing region, whereas the SL estimate is better in the smoothly increasing region but cannot reproduce the constant regions of the true signal. GEN, on the other hand, is able to recover the true signal in all regions. We can also make the same observations when G is the 2D grid graph (but they are harder to visualize).

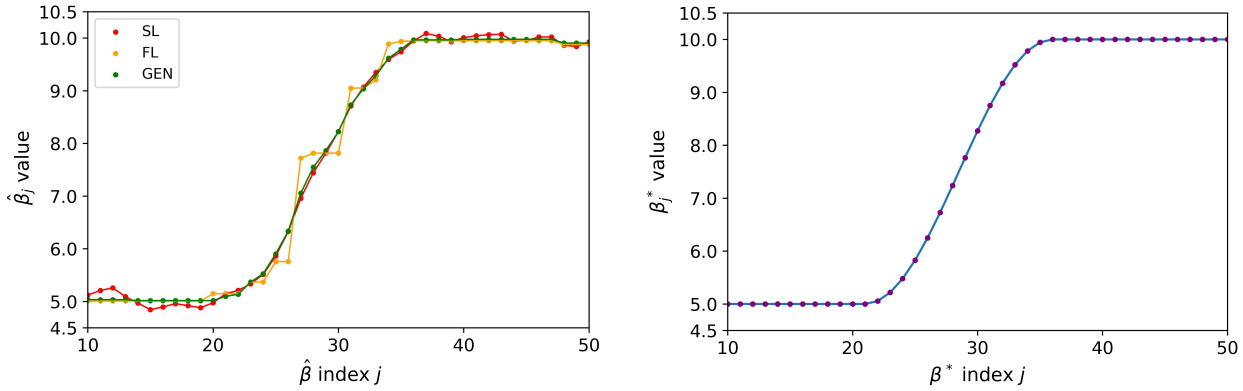


Figure 2.7: Left: estimated signals obtained from FL, SL and GEN. Right: true signal. GEN recovers the true signal well in both the constant and the smoothly increasing regions.

We also examine the performances of FL, SL and GEN as features become more correlated and thus Σ becomes more ill-conditioned. Figure 2.8 shows the estimation errors for the chain graph when Σ is the identity matrix (which is the limit of the Toeplitz covariance

matrix as $\rho \to 0$) and when Σ is the Toeplitz covariance matrix with $\rho = 0.95$. From our theoretical results, we expect that when the features are highly correlated, the performance of the standalone $\lambda_1\|\Gamma\beta\|_1$ penalty (FL) should be negatively affected by the ill-conditioned nature of Σ. However, the $\lambda_2\|\Gamma\beta\|_2^2$ penalty should improve the minimum eigenvalue term in the denominators of our error bounds, especially when $\|\Gamma\beta^*\|_\infty$ is small and λ_2 can be chosen to be larger. Such an improvement is not as noticeable when Σ is the identity, which is already well-conditioned.

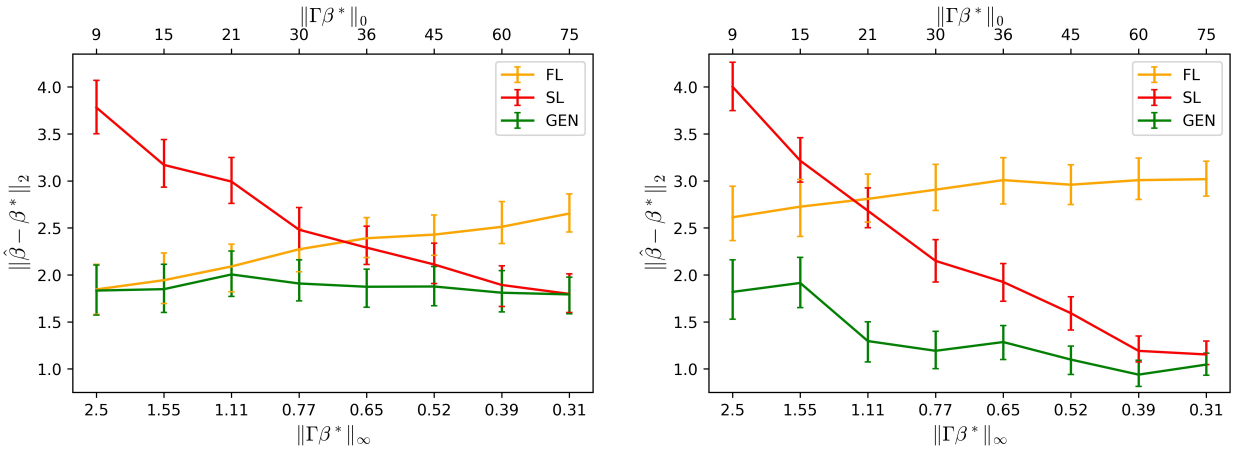


Figure 2.8: Side-by-side comparison of the estimation errors for the chain graph when Σ is the identity matrix (left) and when Σ has the Toeplitz structure with $\rho = 0.95$ (right). $\|\Gamma\beta^*\|_1$ is fixed at 15. Note the greater divergence between the estimation errors of FL and GEN when there is higher correlation.

2.4.1.4 Performance comparisons as n and p vary

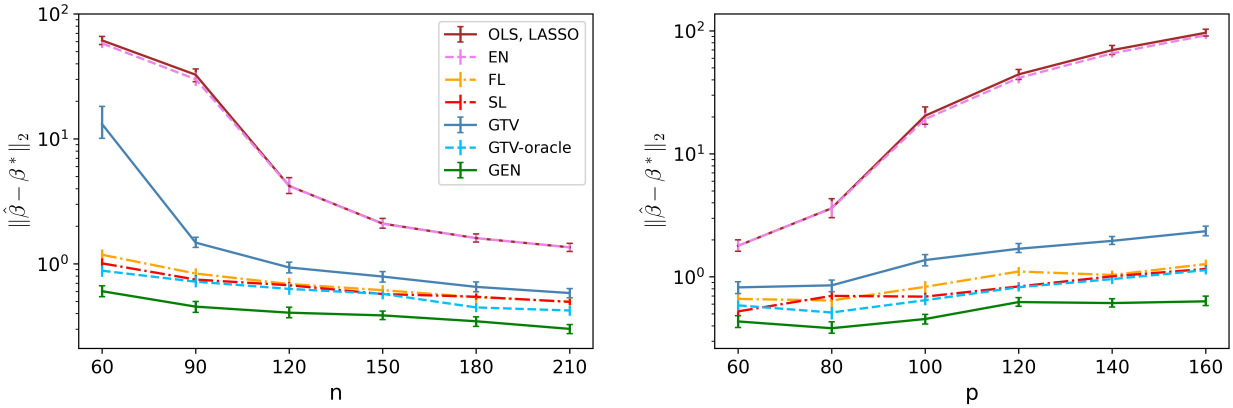
This section examines the performance of GEN relative to all other methods as n is fixed and p increases, or as p is fixed and n increases. The covariance matrix Σ is constructed as in Section 2.4.1.1, and the graphs we use are again the chain graph, the 2D grid and the barbell graph. The true signal β^* is again not sparse, but contains a mix of piecewise constant regions and smoothly varying regions on the graph G (similar to the true signals with intermediate values of $\|\Gamma\beta^*\|_\infty$ and $\|\Gamma\beta^*\|_0$ in Figure 2.4 and Figure 2.5). Note that we

do include the high-dimensional setting when n is smaller than p. As described in Section 2.4.1.2, hyperparameters for all methods are chosen based on the best prediction scores in cross-validation. We only report the estimation errors in Figure 2.9, since the prediction errors for all methods show the same trends.

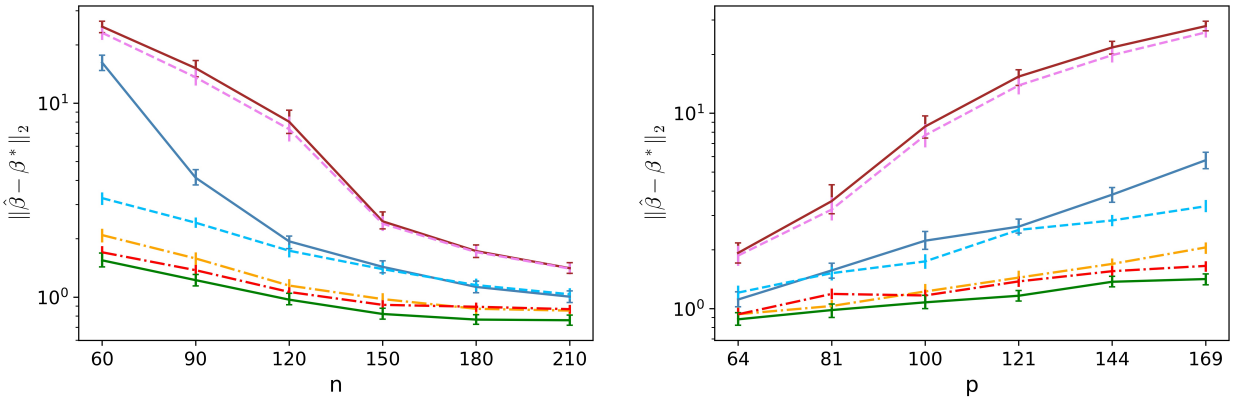
As we can see from Figure 2.9, GEN consistently has the best performance in terms of estimation errors (except for the barbell graph when GTV-oracle performs slightly better for some values of (n, p)). With regard to graph-independent methods, OLS and the Lasso clearly fail to perform well in our setting, and EN only provides limited improvements in terms of estimation errors. FL and SL, whose penalties do take into account the graph G, perform significantly better than the previous three methods but never better than GEN. The performance of GTV, whose penalty depends on the covariance estimate $\hat{\Sigma}$, is not consistent; it can be reasonable for the barbell graph but, for the other two graphs, is not very different from OLS, EN and the Lasso for certain values of (n, p). Interestingly, the performance of GTV-oracle can surpass that of GEN for the barbell graph; this can be attributed to the fact that Σ is constructed to reflect the graph structure and hence is a good estimate for the graph G itself. The divergence between the estimation errors of GTV and GTV-oracle therefore suggests that the covariance estimation error is not negligible, especially when p is small relative to n. Note that GTV requires much more time than other methods for hyperparameter selection and model training, as we have discussed in Section 2.4.1.2.

2.4.1.5 Performance comparisons when β^* is both sparse and smooth over G

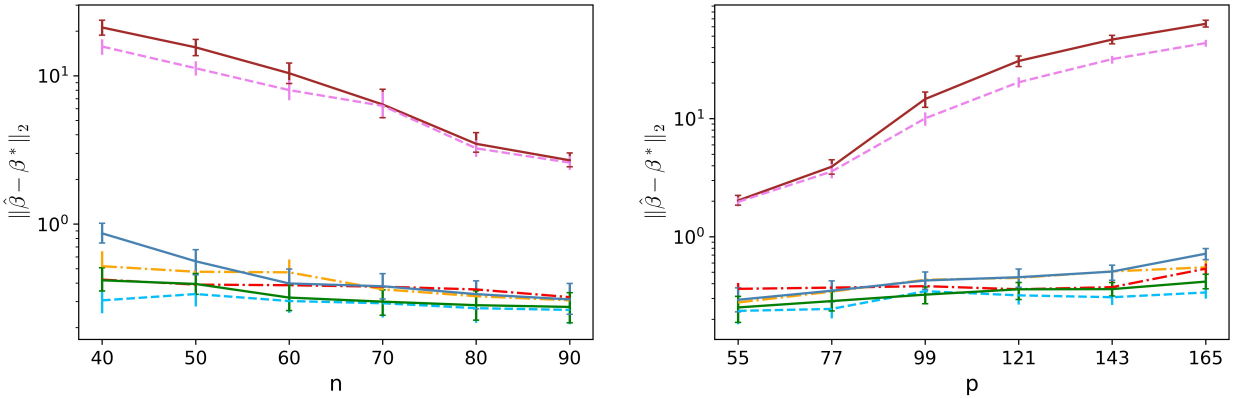
In Section 2.4.1.3 and Section 2.4.1.4, we have compared the performances of various estimators when β^* is dense. We now consider the case when G is the chain graph, and β^* is sparse and has small variations in its successive entries, as illustrated in the left plot of Figure 2.10. Such a signal structure should be more favorable to either FL or SL, and we expect at least one of them to outperform GEN in this case.



(a) 1D chain



(b) 2D grid



(c) Barbell

Figure 2.9: Estimation errors (reported on the log scale) based on 500 resamplings for all estimators as p is fixed ($p = 110$ for chain graph, $p = 121$ for 2D grid, $p = 66$ for barbell graph) but n increases (left), and as $n = 90$ is fixed but p increases (right). $\sigma = 1$ is fixed, and in each plot $\|\Gamma\beta^*\|_\infty$ is kept roughly constant. CV yields λ_L identically equal to zero for the Lasso estimator, and thus its performance coincides with that of OLS.

However, as can be seen in Table 2.3, GEN still has the best performance compared to all other estimators. FL and SL perform better than the Lasso estimator, which in turn is better than OLS as expected. Certainly, such a strong performance relative to the other estimators may depend on the smoothness and sparsity levels of the true signal. Nonetheless, this example clearly demonstrates that effectively leveraging the true signal's smoothness over G can be more important in reducing prediction and estimation errors than exploiting its sparsity structure.

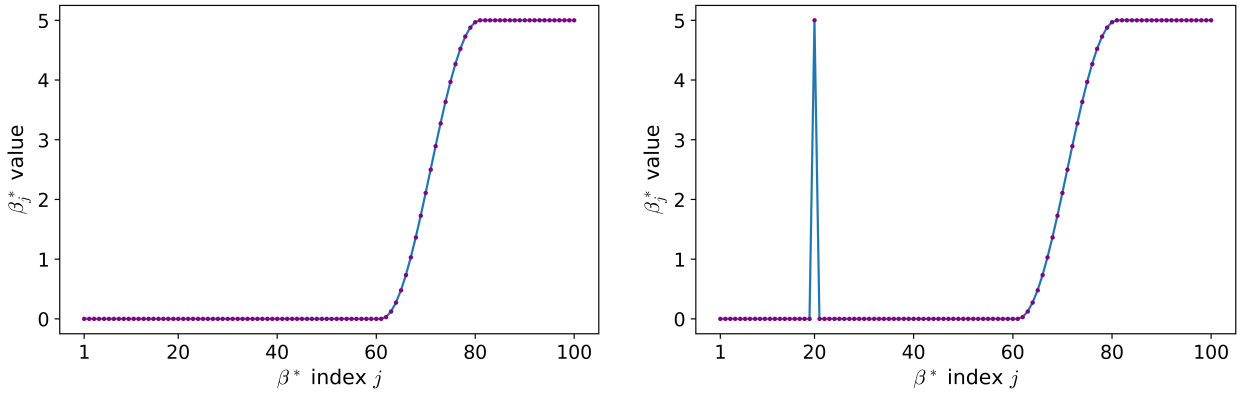


Figure 2.10: Left: Sparse and smooth signal with $p = 100$, $\|\beta^*\|_0 = 40$, $\|\Gamma\beta^*\|_\infty = 0.39$. Right: The left signal is modified to include a spike, so that $\|\Gamma\beta^*\|_\infty$ increases to 5. We use $\sigma = 1, n = 80$ and the Toeplitz covariance matrix with $\rho = 0.5$ for Σ in this section.

We also consider a slight modification to the previous example, so that we have a sharp spike in the zero region of the signal. Adding a single spike should not significantly change the radius R_q of the ℓ_q-ball to which $\Gamma\beta^*$ belongs. However, since $\|\Gamma\beta^*\|_\infty$ is now much larger, CV yields λ_2 identically equal to zero, and GEN degenerates into FL with $\lambda_L = 0$. As a result, FL performs better than GEN (and so does GTV-oracle), although the deterioration of GEN's performance is not drastic and GEN still performs better than EN, SL, GTV and the Lasso. It is therefore a question of interest for future research whether we can replace the ℓ_2 component of GEN with another penalty that is more robust to signal spikes, while retaining the benefits of having the ℓ_2 component as discussed in Section 2.4.1.3.

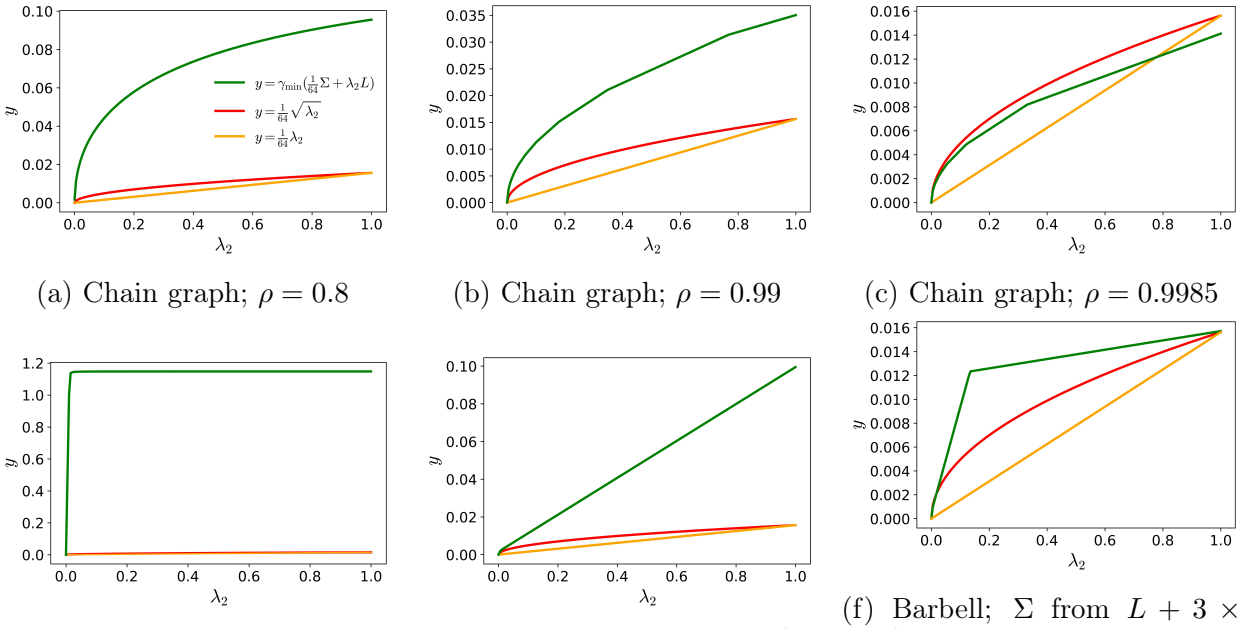
Table 2.3: Prediction and estimation errors for the true signals in Figure 2.10; 'L' and 'R' denote errors for the left and right true signals respectively. The mean and standard deviation of the errors based on 500 resamplings are shown below. Errors better than GEN's errors are shown in orange.

	OLS	**L**	**EN**	**FL**	**SL**	**GTV**	**GTV-oracle**	**GEN**
L: Est. errors	6.92 ± 1.05	1.98 ± 0.36	1.98 ± 0.36	0.56 ± 0.08	0.40 ± 0.06	0.87 ± 0.13	0.38 ± 0.06	0.27 ± 0.05
L: Pred. errors	38.12 ± 15.33	2.94 ± 1.35	2.94 ± 1.35	0.28 ± 0.12	0.33 ± 0.14	0.72 ± 0.26	0.21 ± 0.11	0.15 ± 0.08
R: Est. errors	7.25 ± 1.10	2.95 ± 0.59	2.95 ± 0.59	0.88 ± 0.17	1.75 ± 0.30	1.31 ± 0.22	0.75 ± 0.12	0.97 ± 0.18
R: Pred. errors	41.75 ± 15.95	6.19 ± 2.93	6.19 ± 2.93	0.67 ± 0.26	2.37 ± 0.93	1.44 ± 0.50	0.49 ± 0.19	0.87 ± 0.32

2.4.2 Empirical study of the quantity $\gamma_{min}\left(\frac{1}{64}\Sigma + \lambda_2 L\right)$

In Section 2.2, if Σ is ill-conditioned and we cannot assume $\gamma_{\min}(\Sigma)$ is bounded away from zero, then we assume that $\gamma_{\min}\left(\frac{1}{64}\Sigma + \lambda_2 L\right)$ may be greater than $c\lambda_2$ or $c\sqrt{\lambda_2}$. These assumptions lead to the bounds (2.24) and (2.25), which may allow for consistency in prediction and estimation respectively.

We conjecture that $\gamma_{\min}\left(\frac{1}{64}\Sigma + \lambda_2 L\right) \geq \frac{1}{64}\lambda_2$ holds for all $\lambda_2 \in [0, 1]$ under reasonable assumptions about (Σ, L); this implies $\gamma_{\min}\left(\frac{1}{64}\Sigma + \lambda_2 L\right) \geq \frac{1}{64}\min(\lambda_2, 1)$. Figure 2.11 shows the growth of the quantity $\gamma_{\min}\left(\frac{1}{64}\Sigma + \lambda_2 L\right)$ as a function of λ_2, for the various types of graphs and covariance matrices we have considered in Section 2.4. When G is the chain graph and Σ has the Toeplitz structure, we generally have $\gamma_{\min}\left(\frac{1}{64}\Sigma + \lambda_2 L\right) \geq \frac{1}{64}\sqrt{\lambda_2}$ for all $\lambda_2 \in [0, 1]$ unless $\rho > 0.99$. When G is the 2D grid or barbell graph and Σ is constructed accordingly as in Section 2.4.1.1, we can also observe the same trends. Overall, when Σ is ill-conditioned and λ_2 can be chosen to be sufficiently large, the ℓ_2 component of the GEN penalty can significantly improve our error upper bounds.



(a) Chain graph; $\rho = 0.8$

(b) Chain graph; $\rho = 0.99$

(c) Chain graph; $\rho = 0.9985$

(d) Complete graph; $\rho = 0.99$

(e) 2D grid; Σ from $L + 10^{-4} I_p$

(f) Barbell; Σ from $L + 3 \times 10^{-4} I_p$

Figure 2.11: Growth of $\gamma_{\min}(\frac{1}{64}\Sigma + \lambda_2 L)$ as a function of λ_2 for various choices of Σ and G ($p = 100$ for all plots). In (a), (b) and (c), G is the chain graph and Σ has Toeplitz structure with varying ρ. In (d), G is the complete graph and Σ has Toeplitz structure. In (e) and (f), we use the 2D grid and barbell graph, with corresponding (and highly correlated) covariance structures as in Section 2.4.1.1.

2.4.3 Real data analysis

2.4.3.1 COVID-19 trend prediction

We consider the problem of predicting the number of COVID-19 cases 14 days in advance for a given county in California, using a New York Times-curated COVID-19 dataset. This problem may be of importance for hospitals and local authorities, as they may wish to anticipate potential spikes in COVID-19 cases based on current, local data. It is reasonable to assume that the number of cases Y_{tc} on day t in county c is Poisson-distributed, as in Agosto and Giudici [2020], Bu et al. [2021] and Cori et al. [2013]. In this case, we can apply the variance-stabilizing Anscombe transform $x \mapsto 2\sqrt{x + \frac{3}{8}}$ to form $\tilde{Y}_{tc}$. Following the modeling approach in Cori et al. [2013], we may then consider the Gaussian model

$$\tilde{Y}_{tc} = \sum_{s=14}^{21} \alpha_s Y_{t-s,c} + \epsilon_{tc} \tag{2.49}$$

where $\epsilon_{tc} \sim N(0, \sigma^2)$. In order to reduce temporal correlation between observations, the days t are sampled such that consecutive time points are at least 7 days apart. We restrict our analysis (a) to the period from June 2020 to July 2021 to avoid non-stationary effects in the evolution of the pandemic due to the appearance of new virus strains, and (b) to the 25 densest counties in California where linear models are typically a better fit. As in Ngonghala et al. [2022], we use cross-validation to evaluate the accuracy of our model; 6/7 of our data is used for fitting and the remaining data is for performance evaluation (i.e. 2 months of data). Fitting an OLS model based on (2.49) usually results in a satisfactory fit with an R^2 score above 0.8 (see the Appendix).

We hypothesize that for densely populated counties, rising cases in neighboring counties may further explain a significant fraction of the remaining variance in the data due to population movements between counties. To test this hypothesis, we consider a model that incorporates the number of cases from nearby counties within a two-hop radius of the given

county c:

$$\tilde{Y}_{tc} = \sum_{k \in N_2(c)} \sum_{s=14}^{21} \alpha_{sk} Y_{t-s,k} + \epsilon_{tc} \tag{2.50}$$

Table 2.4: Median RMSE achieved by various methods for 25 counties. OLS is fitted based on model (2.49), and all other methods are based on (2.50). The best performances for each county are highlighted in bold.

County	**OLS**	**L**	**EN**	**FL**	**SL**	**GEN**
Alameda	1.08	1.18	1.14	0.96	**0.85**	0.99
Butte	3.20	1.81	1.83	**1.46**	1.88	2.10
Contra Costa	**1.21**	1.79	1.69	3.47	2.49	3.35
Fresno	8.22	7.06	9.77	**5.25**	7.71	5.95
Los Angeles	4.92	6.92	7.34	5.28	6.06	**4.56**
Marin	6.11	3.92	5.49	5.47	**3.38**	4.18
Merced	7.94	9.75	9.03	9.26	9.68	**6.08**
Napa	3.93	3.97	5.53	3.92	3.90	**2.96**
Orange	**1.84**	4.44	3.97	3.34	2.08	2.56
Placer	2.20	1.35	1.89	**1.27**	1.55	1.53
Riverside	3.46	3.32	3.62	3.21	**2.79**	3.73
Sacramento	2.23	3.11	2.53	3.61	2.31	**1.66**
San Diego	1.43	1.67	1.00	0.98	**0.88**	0.91
San Francisco	1.41	2.23	**1.05**	1.23	1.33	1.39
San Joaquin	3.64	**3.43**	**3.43**	3.93	5.24	5.41
San Mateo	**1.44**	2.34	2.45	1.68	1.56	1.75
Santa Barbara	2.84	2.02	2.02	2.02	**2.01**	3.71
Santa Clara	1.14	2.04	1.85	1.05	1.10	**1.03**
Santa Cruz	6.56	3.86	4.62	**3.55**	4.17	4.59
Solano	2.10	3.86	3.72	**2.03**	2.93	2.62
Sonoma	**1.74**	3.47	3.60	2.78	2.62	2.67
Stanislaus	9.29	6.41	9.17	**4.55**	4.76	4.88
Sutter	4.07	7.51	7.51	4.33	2.58	**1.94**
Ventura	2.02	1.22	1.20	1.23	**1.16**	1.36
Yolo	3.43	5.13	4.54	1.93	2.45	**1.79**

Fitting model (2.50) is a high-dimensional problem, where the number of parameters p can be up to 3 times the number of observations n, depending on the county. OLS therefore is not a suitable method for model (2.50). Consequently, in this experiment we fit the penalty-based methods (except GTV) based on (2.50) and compare with the performance of OLS computed based on (2.49). The graph G we consider here is such that two feature

vectors are connected if they are indexed by the same day t and by two adjacent counties, or if they are indexed by the same county k and two consecutive time points.

We perform this prediction task for each of 25 most densely populated counties in California. For each county, we report the median root mean square error (RMSE) computed on the test set in Table 2.4. As expected, incorporating the numbers of cases in neighboring counties allows us to outperform OLS based on (2.49) in 21 out of 25 counties. The graph-dependent methods FL, SL and GEN perform better than other methods in 19 out of 25 counties. Among these 19 counties, GEN has the best performance in 7 of them and is therefore a competitive candidate for this prediction task. The improvement it yields can be quite substantial; for the county Sutter in particular, GEN reduces the RMSE by 50% compared to OLS and at least 25% compared to FL and SL.

2.4.3.2 Detection of Alzheimer's disease

We test the GEN penalty's performance in detecting Alzheimer's disease, using an MRI dataset available on Kaggle. The task is to classify whether the MRI images in the dataset show signs of dementia. Since the responses are binary, we need to consider the logistic extension (2.6) of our method as well as that of all other methods. The original dataset has images labeled with moderate, mild, very mild and no dementia, but we exclude the moderate cases due to the small number of training samples. We also exclude the very mild cases since the images may be too similar to those with no dementia, thus leading to lower prediction accuracy for all methods.

Since the features are 2D MRI images, it is natural to use the 2D grid graph as our graph G, which is of size $p = 32 \times 32 = 1024$ (we compress the original images to this size for computational convenience). We use the first 800 images with no dementia and 400 images with mild dementia in the original dataset. Out of these 1200 images, $n = 480$ images are used as training data (note that $n < p$), 480 images are use for hyperparameter tuning,

and the other 240 images constitute our testing data. For computation, we use ECOS for all methods. Since GTV requires at least a 3D grid search for hyperparameter tuning, it is too slow to be considered for this experiment. All other methods take at most 5 seconds of training time.

The classification accuracies for all methods except GTV are reported in Table 2.5. As expected, GEN shows better prediction performance than all other methods in consideration.

Table 2.5: Prediction accuracies for classification of Alzhemer's disease status. Here, OLS is replaced by logistic regression (LR), and the logistic extensions of all penalty-based methods (except GTV) are used.

	LR	**L**	**EN**	**FL**	**SL**	**GEN**
Accuracy	82.08%	90.0 %	92.50%	91.25%	92.08%	92.92%

2.4.3.3 Estimation of crime patterns in Chicago

Consider the task of uncovering crime trends over time across the 77 communities of Chicago (which we denote by the set $\mathcal{C}$). Statistics on the number of crimes per community between 2004 and 2022 are available on the city's data portal. The monthly crime rates (which are defined here as the number of crimes per 100,000 inhabitants) vary over the years and across the communities, and they are also subject to significant seasonal effects. Additional details on the nature of the data and preprocessing are provided in the Appendix.

Let $Y_{my}^{(c)}$ denote the crime rate for community $c \in \mathcal{C}$, month m and year y. Since we are working with count data, it is reasonable to pre-process the data by applying the Anscombe transform to $Y_{my}^{(c)}$ to form $\tilde{Y}_{my}^{(c)}$. We then consider the following additive Gaussian model

$$\tilde{Y}_{my}^{(c)} = \sum_{i=1}^{12} \alpha_i \mathbf{1}[m = i] + \sum_{j=2004}^{2022} \beta_j \mathbf{1}[y = j] + \sum_{c \in \mathcal{C}} \gamma_k \mathbf{1}[k = c] + \epsilon_{my}^{(c)} \tag{2.51}$$

where $\epsilon_{my}^{(c)} \sim N(0, \sigma^2)$. While our design matrix here is not equal to identity as in the trend

filtering case, note that it contains "one-hot" encoding rather than i.i.d. rows from some distribution $\mathbb{P}$. The parameters (α, β) naturally exhibit *temporal smoothness*, since we expect them to vary smoothly over time. The community offset parameter γ, on the other hand, should exhibit *spatial smoothness*, as we expect neighboring communities to have similar offsets. To define our GEN penalty, we encode these prior beliefs in a regularizing graph G with 3 disconnected components: one chain graph reinforcing the temporal smoothness of the month coefficients, another chain graph for that of the years, and a third component encoding neighborhood adjacency. We compare our method's performance with all other methods except GTV.

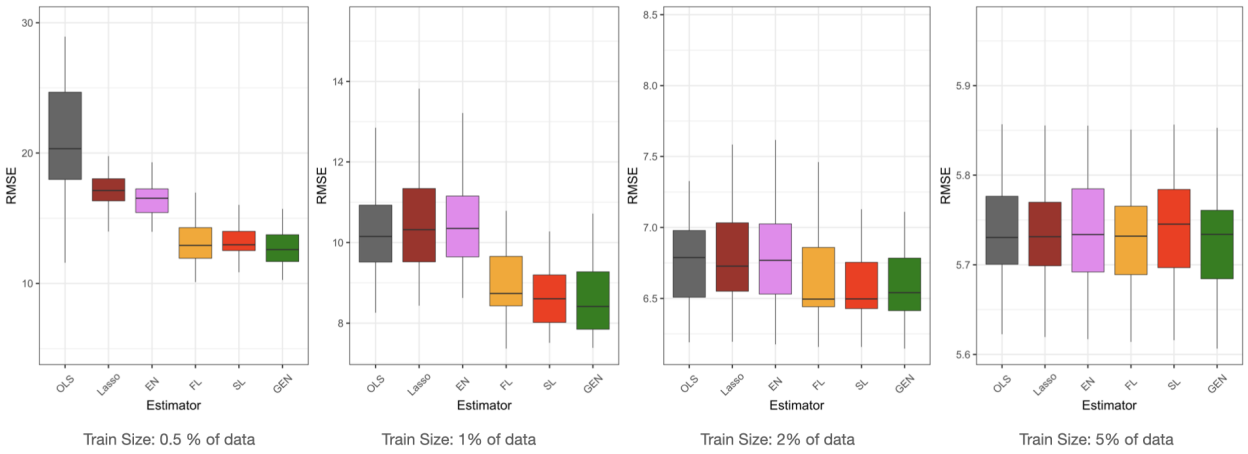


Figure 2.12: RMSE achieved by different estimators as the proportion α of data used for training varies.

Figure 2.12 compares the prediction performance (reported using RMSE computed on held-out data across 40 independent trials) for all methods. Here, performance is assessed for different data regimes: while the original dataset contains 17,094 observations, we use a fraction $\alpha \in \{0.5\%, 1\%, 2\%, 5\%\}$ of data for estimation of the $p = 108$ parameters in our model ($\alpha = 0.5\%$ and $\alpha = 1\%$ correspond to $p > n$ and $p \approx n$ respectively). As shown in Figure 2.12, GEN performs consistently better than all other methods, especially in the data-sparse regime. While in this example we are more interested in the estimation of crime patterns rather than prediction (note that the model (2.51) cannot be used to predict crime

rates beyond 2022), Figure 2.12 provides evidence for GEN's superior performance and can be of interest if we are given a dataset with many missing values that require data imputation.

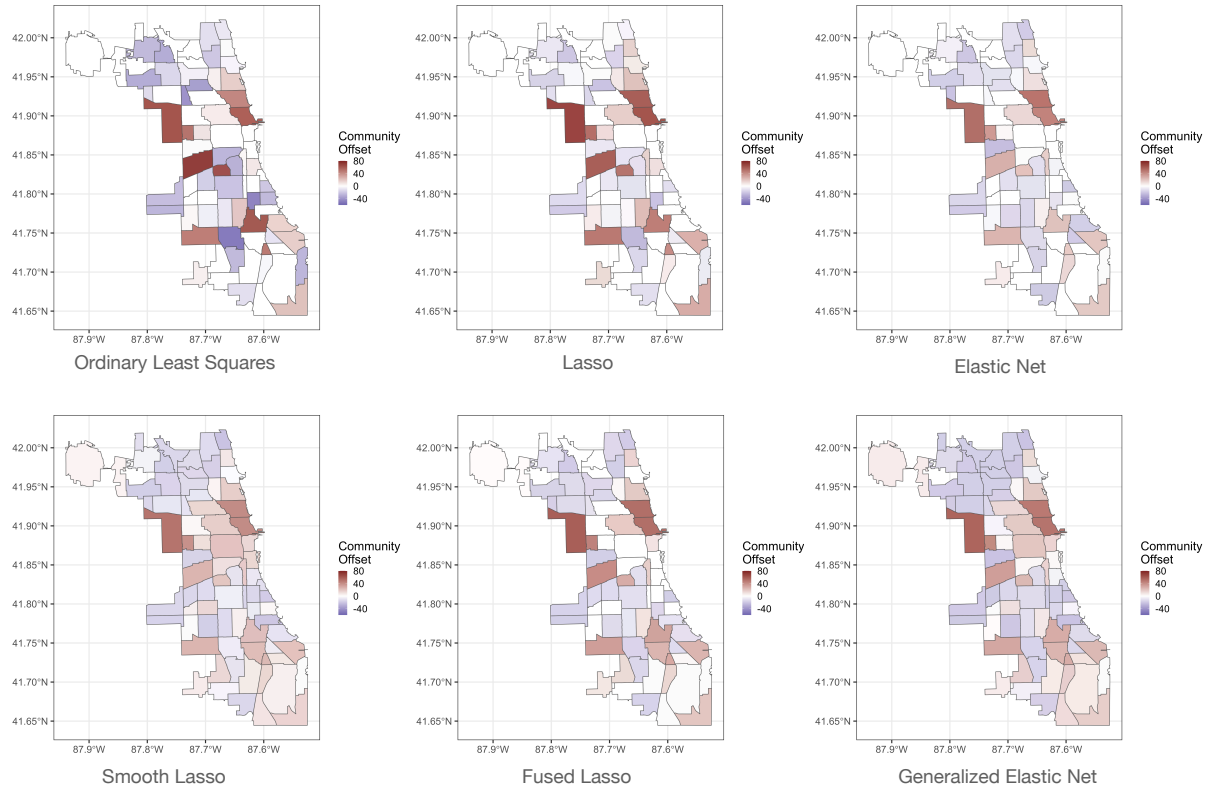
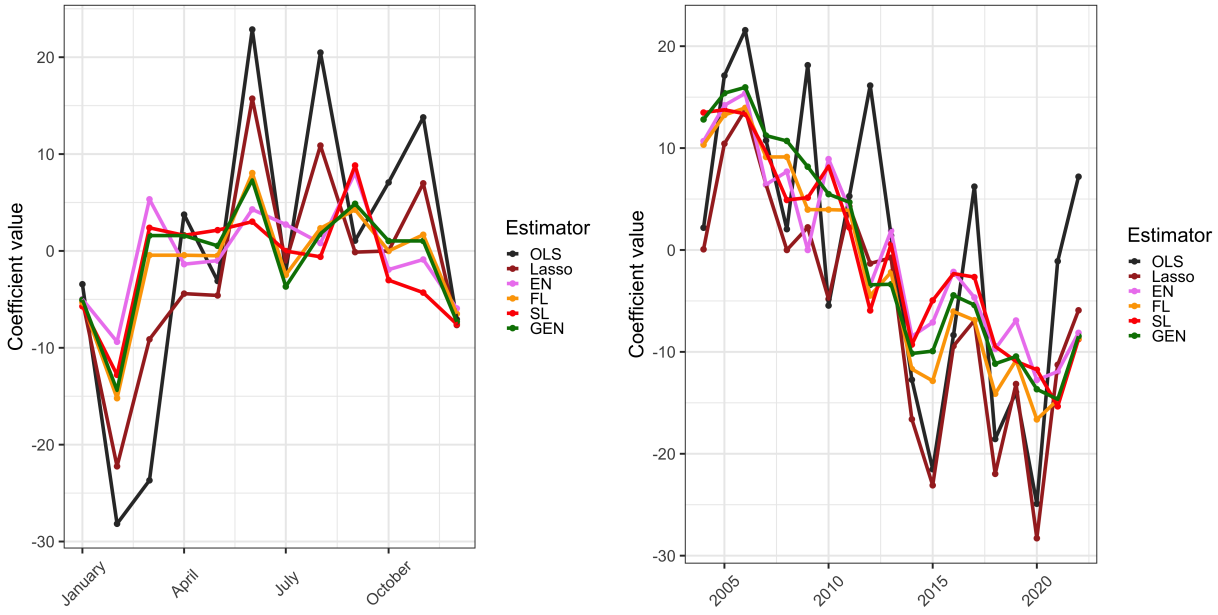


Figure 2.13: Visualization of the community offsets γ produced by different estimators. Note that GEN produces smoother estimates with greater magnitudes.

Figure 2.13 and Figure 2.14 visualize the estimates of the community offsets γ and the temporal parameters (α, β) respectively. Note that the estimate of γ obtained from GEN contains fewer zero entries and is significantly smoother when compared with other methods. We can clearly see that GEN divides the communities into clusters with similar community offsets. Only Smooth Lasso provides an estimate of γ that is close to GEN's, but interestingly the estimates obtained by GEN tend to be greater in magnitude. From Figure 2.14, we can see that GEN, FL and SL produce smooth estimates to show that crime rates tend to decrease in the colder months and that there is a general reduction in crime rates between 2002 and 2022.



(a) Estimates of α

(b) Estimates of β

Figure 2.14: Visualization of the estimates of the monthly parameters α and the yearly parameters β's produced by different methods. GEN, FL and SL produce smoother estimates relative to graph-independent methods.

2.5 Proofs and supplementary materials

2.5.1 Proofs of theoretical results

We restate our theorems in this appendix for convenience.

Theorem 47 (Theorem 33). *Fix $\delta > 0$ and choose $\lambda_1 = 32\sigma\rho(\Gamma)\sqrt{\frac{\gamma_{max}(\Sigma)\log p}{n}}$, $\lambda_2 \leq \frac{\lambda_1}{8\|\Gamma\beta^*\|_\infty}$. Given any set S satisfying both*

$$\frac{144\gamma_{max}(\Sigma)(\sqrt{n_c}+\delta)^2}{n} + \frac{36\lambda_1^2|S|k_S^{-2}}{\sigma^2} \leq \frac{1}{2}\gamma_{min}\left(\frac{1}{64}\Sigma + \lambda_2 L\right) \tag{2.52}$$

and

$$\lambda_1\|(\Gamma\beta^*)_{-S}\|_1 \leq \frac{\sigma^2}{18} \tag{2.53}$$

with probability at least $1 - c_1 \exp(-nc_2) - \frac{2}{m} - e^{-\delta^2/2}$ *we have*

$$\|\Sigma^{1/2}(\hat{\beta} - \beta^*)\|_2^2 \lesssim \frac{\sigma^2 \gamma_{max}(\Sigma)\frac{n_c+\delta^2}{n} + \lambda_1^2 |S| k_S^{-2}}{\gamma_{min}\left(\frac{1}{64}\Sigma + \lambda_2 L\right)} + \lambda_1 \|(\Gamma\beta^*)_{-S}\|_1$$

and

$$\|\hat{\beta} - \beta^*\|_2^2 \lesssim \frac{\sigma^2 \gamma_{max}(\Sigma)\frac{n_c+\delta^2}{n} + \lambda_1^2 |S| k_S^{-2}}{\gamma_{min}^2\left(\frac{1}{64}\Sigma + \lambda_2 L\right)} + \frac{\lambda_1 \|(\Gamma\beta^*)_{-S}\|_1}{\gamma_{min}\left(\frac{1}{64}\Sigma + \lambda_2 L\right)}$$

Proof. By definition,

$$\hat{\beta} = \arg\min_{\beta} \frac{1}{n}\|Y - X\beta\|_2^2 + \lambda_1 \|\Gamma\beta\|_1 + \lambda_2 \|\Gamma\beta\|_2^2$$

We can also rewrite our estimator as:

$$\hat{\beta} = \arg\min_{\beta} \frac{1}{n}\|\tilde{Y} - \tilde{X}\beta\|_2^2 + \lambda_1 \|\Gamma\beta\|_1$$

Using subdifferential calculus, we can see that $\hat{\beta}$ must satisfy

$$\frac{2\tilde{X}^T(\tilde{Y} - \tilde{X}\hat{\beta})}{n} = \lambda_1 \Gamma^T \text{sign}(\Gamma\hat{\beta})$$

where

$$[\text{sign}(x)]_i = \begin{cases} 1 & \text{if } x_i > 0, \\ \text{any value in } [-1, 1] & \text{if } x_i = 0, \\ -1 & \text{if } x_i < 0\,. \end{cases}$$

Hence, we obtain

$$\frac{2}{n}\hat{\beta}^T\tilde{X}^T(\tilde{Y} - \tilde{X}\hat{\beta}) = \lambda_1 \hat{\beta}^T\Gamma^T \text{sign}(\Gamma\hat{\beta}) = \lambda_1 \|\Gamma\hat{\beta}\|_1$$

and for any $\beta \in \mathbb{R}^p$,

$$\frac{2}{n}\beta^T \tilde{X}^T(\tilde{Y} - \tilde{X}\hat{\beta}) = \lambda_1 \beta^T \Gamma^T \text{sign}(\Gamma\hat{\beta}) \leq \lambda_1 \|\Gamma\beta\|_1$$

By subtracting the previous equality from the inequality above, for any $\beta \in \mathbb{R}^p$ we have

$$\frac{2}{n}(\beta - \hat{\beta})^T \tilde{X}^T(\tilde{Y} - \tilde{X}\hat{\beta}) \leq \lambda_1(\|\Gamma\beta\|_1 - \|\Gamma\hat{\beta}\|_1)$$

Since $\tilde{Y} = \tilde{X}\beta^* + \tilde{\epsilon}$,

$$\begin{aligned}
&\frac{2}{n}(\hat{\beta} - \beta)^T \tilde{X}^T \tilde{X}(\hat{\beta} - \beta^*) \\
&\leq \frac{2}{n}\tilde{\epsilon}^T \tilde{X}(\hat{\beta} - \beta) + \lambda_1(\|\Gamma\beta\|_1 - \|\Gamma\hat{\beta}\|_1) \\
&= \frac{2}{n}\epsilon^T X(\hat{\beta} - \beta) - 2\lambda_2(\beta^*)^T \Gamma^T \Gamma(\hat{\beta} - \beta) + \lambda_1(\|\Gamma\beta\|_1 - \|\Gamma\hat{\beta}\|_1) \\
&\leq \frac{2}{n}\epsilon^T X(\hat{\beta} - \beta) + 2\lambda_2 \|\Gamma\beta^*\|_\infty \|\Gamma(\hat{\beta} - \beta^*)\|_1 + \lambda_1(\|\Gamma\beta\|_1 - \|\Gamma\hat{\beta}\|_1) \\
&\leq \frac{2}{n}\epsilon^T X(\hat{\beta} - \beta) + \frac{\lambda_1}{4}\|\Gamma(\hat{\beta} - \beta)\|_1 + \lambda_1(\|\Gamma\beta\|_1 - \|\Gamma\hat{\beta}\|_1)
\end{aligned}$$

where the last inequality follows if we choose $\lambda_2 \leq \frac{\lambda_1}{8\|\Gamma\beta^*\|_\infty}$.

We wish to bound $\frac{2}{n}\epsilon^T X(\hat{\beta} - \beta)$. As $\Pi \in \mathbb{R}^{p \times p}$ denotes the projection matrix onto the kernel of Γ, we have $I_p = \Pi + \Gamma^\dagger \Gamma$. Hence,

$$\begin{aligned}
\frac{2}{n}\epsilon^T X(\hat{\beta} - \beta) &= \frac{2}{n}\epsilon^T X \Pi(\hat{\beta} - \beta) + \frac{2}{n}\epsilon^T X \Gamma^\dagger \Gamma(\hat{\beta} - \beta) \\
&\leq \frac{2}{n}\|\Pi X^T \epsilon\|_2 \|\hat{\beta} - \beta\|_2 + \frac{2}{n}\|(\Gamma^\dagger)^T X^T \epsilon\|_\infty \|\Gamma(\hat{\beta} - \beta)\|_1 \\
&\leq \frac{2}{n}\|\Pi X^T \epsilon\|_2 \|\hat{\beta} - \beta\|_2 + \frac{\lambda_1}{4}\|\Gamma(\hat{\beta} - \beta)\|_1
\end{aligned} \tag{2.54}$$

where the last inequality follows if we choose $\lambda_1 \geq \frac{8}{n}\|(\Gamma^\dagger)^T X^T \epsilon\|_\infty$ (with high probability).

We obtain the bound:

$$\begin{aligned}\frac{2}{n}(\hat{\beta}-\beta)^T\tilde{X}^T\tilde{X}(\hat{\beta}-\beta^*) \leq & \frac{2}{n}\|\Pi X^T\epsilon\|_2\|\hat{\beta}-\beta\|_2 + \frac{\lambda_1}{2}\|\Gamma(\hat{\beta}-\beta)\|_1 \\ & + \lambda_1\|\Gamma\beta\|_1 - \lambda_1\|\Gamma\hat{\beta}\|_1\end{aligned} \tag{2.55}$$

$$\frac{2}{n}(\hat{\beta}-\beta)^T\tilde{X}^T\tilde{X}(\hat{\beta}-\beta^*) \leq \frac{2}{n}\|\Pi X^T\epsilon\|_2\|\hat{\beta}-\beta\|_2 + \frac{\lambda_1}{2}\|\Gamma(\hat{\beta}-\beta)\|_1 + \lambda_1\|\Gamma\beta\|_1 - \lambda_1\|\Gamma\hat{\beta}\|_1$$

For any $S \subseteq [m]$:

$$\begin{aligned}&\frac{\lambda_1}{2}\|\Gamma(\hat{\beta}-\beta)\|_1 + \lambda_1\|\Gamma\beta\|_1 - \lambda_1\|\Gamma\hat{\beta}\|_1 \\ &\leq \frac{\lambda_1}{2}\|(\Gamma\hat{\beta}-\Gamma\beta)_S\|_1 + \frac{\lambda_1}{2}\|(\Gamma\hat{\beta})_{-S}\|_1 + \frac{\lambda_1}{2}\|(\Gamma\beta)_{-S}\|_1 + \lambda_1\|\Gamma\beta\|_1 - \lambda_1\|\Gamma\hat{\beta}\|_1 \\ &\leq \frac{3\lambda_1}{2}\|(\Gamma\hat{\beta}-\Gamma\beta)_S\|_1 + \frac{3\lambda_1}{2}\|(\Gamma\beta)_{-S}\|_1 - \frac{\lambda_1}{2}\|(\Gamma\hat{\beta})_{-S}\|_1 \\ &\leq \frac{3\lambda_1}{2}\|(\Gamma\hat{\beta}-\Gamma\beta)_S\|_1 + 2\lambda_1\|(\Gamma\beta)_{-S}\|_1 - \frac{\lambda_1}{2}\|(\Gamma\hat{\beta}-\Gamma\beta)_{-S}\|_1 \\ &\leq 2\lambda_1\|(\Gamma\hat{\beta}-\Gamma\beta)_S\|_1 + 2\lambda_1\|(\Gamma\beta)_{-S}\|_1 - \frac{\lambda_1}{2}\|\Gamma\hat{\beta}-\Gamma\beta\|_1\end{aligned}$$

and so we have

$$\begin{aligned}&\frac{2}{n}(\hat{\beta}-\beta)^T\tilde{X}^T\tilde{X}(\hat{\beta}-\beta^*) + \frac{\lambda_1}{2}\|\Gamma\hat{\beta}-\Gamma\beta\|_1 \\ &\leq \frac{2}{n}\|\Pi X^T\epsilon\|_2\|\hat{\beta}-\beta\|_2 + 2\lambda_1\|(\Gamma\hat{\beta}-\Gamma\beta)_S\|_1 + 2\lambda_1\|(\Gamma\beta)_{-S}\|_1 \\ &\leq 2\left(\frac{1}{n}\|\Pi X^T\epsilon\|_2 + \frac{\lambda_1\sqrt{|S|}}{k_S}\right)\|\hat{\beta}-\beta\|_2 + 2\lambda_1\|(\Gamma\beta)_{-S}\|_1 \\ &\leq 2\left(\sqrt{2\sigma^2\gamma_{\max}(\Sigma)}\frac{\sqrt{n_c}+\delta}{\sqrt{n}} + \frac{\lambda_1\sqrt{|S|}}{k_S}\right)\|\hat{\beta}-\beta\|_2 + 2\lambda_1\|(\Gamma\beta)_{-S}\|_1\end{aligned}$$

with high probability, where we used the definition of k_S and Lemma 52. If we set $\beta = \beta^*$,

we obtain

$$\frac{1}{n}\|\tilde{X}(\hat{\beta}-\beta^*)\|_2^2+\frac{\lambda_1}{4}\|\Gamma\hat{\beta}-\Gamma\beta^*\|_1\leq\left(\sqrt{2\sigma^2\gamma_{\max}(\Sigma)}\frac{\sqrt{n_c}+\delta}{\sqrt{n}}+\frac{\lambda_1\sqrt{|S|}}{k_S}\right)\|\hat{\beta}-\beta^*\|_2 + \lambda_1\|(\Gamma\beta^*)_{-S}\|_1 \tag{2.56}$$

which implies

$$\lambda_1\|\Gamma\hat{\beta}-\Gamma\beta^*\|_1\leq 4\left(\sqrt{2\sigma^2\gamma_{\max}(\Sigma)}\frac{\sqrt{n_c}+\delta}{\sqrt{n}}+\frac{\lambda_1\sqrt{|S|}}{k_S}\right)\|\hat{\beta}-\beta^*\|_2+4\lambda_1\|(\Gamma\beta^*)_{-S}\|_1$$

or that

$$\begin{aligned}
&576\rho^2(\Gamma)\frac{\gamma_{\max}(\Sigma)\log p}{n}\|\Gamma\hat{\beta}-\Gamma\beta^*\|_1^2=\frac{576}{1024}\frac{\lambda_1^2}{\sigma^2}\|\Gamma\hat{\beta}-\Gamma\beta^*\|_1^2\\
&\leq 18\frac{\lambda_1^2}{\sigma^2}\left(\sqrt{2\sigma^2\gamma_{\max}(\Sigma)}\frac{\sqrt{n_c}+\delta}{\lambda_1\sqrt{n}}+\frac{\sqrt{|S|}}{k_S}\right)^2\|\hat{\beta}-\beta^*\|_2^2+18\frac{\lambda_1^2}{\sigma^2}\|(\Gamma\beta^*)_{-S}\|_1^2\\
&\leq\left(72\gamma_{\max}(\Sigma)\frac{(\sqrt{n_c}+\delta)^2}{n}+36\frac{\lambda_1^2|S|k_S^{-2}}{\sigma^2}\right)\|\hat{\beta}-\beta^*\|_2^2+18\frac{\lambda_1^2}{\sigma^2}\|(\Gamma\beta^*)_{-S}\|_1^2\\
&\leq\left(72\gamma_{\max}(\Sigma)\frac{(\sqrt{n_c}+\delta)^2}{n}+36\frac{\lambda_1^2|S|k_S^{-2}}{\sigma^2}\right)\|\hat{\beta}-\beta^*\|_2^2+\lambda_1\|(\Gamma\beta^*)_{-S}\|_1
\end{aligned}\tag{2.57}$$

where we used the condition (2.53). Now if we apply Corollary 51 to (2.56), we have

$$\begin{aligned}
&(\hat{\beta}-\beta^*)^T\left(\frac{1}{64}\Sigma+\lambda_2 L\right)(\hat{\beta}-\beta^*)\leq\left(\sqrt{2\sigma^2\gamma_{\max}(\Sigma)}\frac{\sqrt{n_c}+\delta}{\sqrt{n}}+\frac{\lambda_1\sqrt{|S|}}{k_S}\right)\|\hat{\beta}-\beta^*\|_2\\
&+\lambda_1\|(\Gamma\beta^*)_{-S}\|_1+\frac{72\gamma_{\max}(\Sigma)n_c}{n}\|\hat{\beta}-\beta^*\|_2^2+576\rho^2(\Gamma)\frac{\gamma_{\max}(\Sigma)\log p}{n}\|\Gamma\hat{\beta}-\Gamma\beta^*\|_1^2
\end{aligned}$$

which, by (2.57) and the inequality $\frac{n_c}{n} \leq \frac{(\sqrt{n_c}+\delta)^2}{n}$, implies

$$(\hat{\beta}-\beta^*)^T\left(\frac{1}{64}\Sigma+\lambda_2 L\right)(\hat{\beta}-\beta^*) \leq \left(\sqrt{2\sigma^2\gamma_{\max}(\Sigma)}\frac{\sqrt{n_c}+\delta}{\sqrt{n}}+\frac{\lambda_1\sqrt{|S|}}{k_S}\right)\|\hat{\beta}-\beta^*\|_2$$
$$+2\lambda_1\|(\Gamma\beta^*)_{-S}\|_1+\left(\frac{144\gamma_{\max}(\Sigma)(\sqrt{n_c}+\delta)^2}{n}+\frac{36\lambda_1^2|S|k_S^{-2}}{\sigma^2}\right)\|\hat{\beta}-\beta^*\|_2^2$$

If we now apply the condition (2.52), we obtain

$$(\hat{\beta}-\beta^*)^T\left(\frac{1}{64}\Sigma+\lambda_2 L\right)(\hat{\beta}-\beta^*) \leq \left(\sqrt{2\sigma^2\gamma_{\max}(\Sigma)}\frac{\sqrt{n_c}+\delta}{\sqrt{n}}+\frac{\lambda_1\sqrt{|S|}}{k_S}\right)\|\hat{\beta}-\beta^*\|_2$$
$$+2\lambda_1\|(\Gamma\beta^*)_{-S}\|_1+\frac{1}{2}\gamma_{\min}\left(\frac{1}{64}\Sigma+\lambda_2 L\right)\|\hat{\beta}-\beta^*\|_2^2$$

which, by using $\gamma_{\min}\left(\frac{1}{64}\Sigma+\lambda_2 L\right)\|\hat{\beta}-\beta^*\|_2^2 \leq (\hat{\beta}-\beta^*)^T\left(\frac{1}{64}\Sigma+\lambda_2 L\right)(\hat{\beta}-\beta^*)$, implies both

$$\gamma_{\min}\left(\frac{1}{64}\Sigma+\lambda_2 L\right)\|\hat{\beta}-\beta^*\|_2^2 \leq 2\left(\sqrt{2\sigma^2\gamma_{\max}(\Sigma)}\frac{\sqrt{n_c}+\delta}{\sqrt{n}}+\frac{\lambda_1\sqrt{|S|}}{k_S}\right)\|\hat{\beta}-\beta^*\|_2$$
$$+4\lambda_1\|(\Gamma\beta^*)_{-S}\|_1 \tag{2.58}$$

and

$$(\hat{\beta}-\beta^*)^T\left(\frac{1}{64}\Sigma+\lambda_2 L\right)(\hat{\beta}-\beta^*) \leq 4\lambda_1\|(\Gamma\beta^*)_{-S}\|_1$$
$$+2\frac{\sqrt{2\sigma^2\gamma_{\max}(\Sigma)}\frac{\sqrt{n_c}+\delta}{\sqrt{n}}+\frac{\lambda_1\sqrt{|S|}}{k_S}}{\sqrt{\gamma_{\min}\left(\frac{1}{64}\Sigma+\lambda_2 L\right)}}\sqrt{(\hat{\beta}-\beta^*)^T\left(\frac{1}{64}\Sigma+\lambda_2 L\right)(\hat{\beta}-\beta^*)} \tag{2.59}$$

The error bounds follow from (2.58) and (2.59) if we note that $x^2-bx-c \leq 0$ implies $x^2 \leq 4\max(b^2,c) \leq 4(b^2+c)$, for $b,c>0$. □

Theorem 48 (Theorem 46)**.** *Let* Γ *be the incidence matrix of the p-vertex chain graph, and*

fix $\delta > 0$. With an appropriate choice of λ_1 and $\lambda_2 \leq \frac{\lambda_1}{8\|\Gamma\beta^*\|_\infty}$, with high probability we have

$$\|\Sigma^{1/2}(\hat{\beta} - \beta^*)\|_2^2 \lesssim \frac{\sigma^2\gamma_{max}(\Sigma)}{\gamma_{min}(\frac{1}{64}\Sigma + \lambda_2 L)}\frac{1+\delta^2}{n} + \frac{(\sigma^2\gamma_{max}(\Sigma)\|\Gamma\beta^*\|_1)^{2/3}}{\gamma_{min}^{1/3}\left(\frac{1}{64}\Sigma + \lambda_2 L\right)}\frac{(p\log p)^{1/3}}{n^{2/3}} \tag{2.60}$$

$$\|\hat{\beta} - \beta^*\|_2^2 \lesssim \frac{\sigma^2\gamma_{max}(\Sigma)}{\gamma_{min}^2(\frac{1}{64}\Sigma + \lambda_2 L)}\frac{1+\delta^2}{n} + \frac{(\sigma^2\gamma_{max}(\Sigma)\|\Gamma\beta^*\|_1)^{2/3}}{\gamma_{min}^{4/3}\left(\frac{1}{64}\Sigma + \lambda_2 L\right)}\frac{(p\log p)^{1/3}}{n^{2/3}} \tag{2.61}$$

provided that the RHS of (2.60) *is smaller than $C\sigma^2$.*

Proof. The proof is identical to that of Theorem 47 up to (2.54). However, we need to bound $\frac{2}{n}\epsilon^T X\Gamma^\dagger\Gamma(\hat{\beta} - \beta)$ differently.

Let $\Gamma = U\Xi V^T$ be the singular value decomposition of Γ, and let $\xi_1, \ldots, \xi_{p-1}$ be the nonzero singular values of Γ. Let $u_1, \ldots, u_m$ and $v_1, \ldots, v_p$ denote the columns of U and V. Denote by $V_{[k]} \in \mathbb{R}^{p\times k}$ the matrix containing the first k columns of V (k is to be specified later) and $V_{-[k]} \in \mathbb{R}^{p\times(p-k)}$ the matrix containing the other $p-k$ columns of V. Define the projection matrix $P_{[k]} := V_{[k]}V_{[k]}^T \in \mathbb{R}^{p\times p}$.

Noting that $\Gamma^\dagger\Gamma$ is a projection matrix, we have:

$$\begin{aligned}
&\frac{2}{n}\epsilon^T X\Gamma^\dagger\Gamma(\hat{\beta} - \beta) \\
&= \frac{2}{n}\epsilon^T XP_{[k]}\Gamma^\dagger\Gamma(\hat{\beta} - \beta) + \frac{2}{n}\epsilon^T X(I_p - P_{[k]})\Gamma^\dagger\Gamma(\hat{\beta} - \beta) \\
&\leq \frac{2}{n}\|P_{[k]}X^T\epsilon\|_2\|\Gamma^\dagger\Gamma(\hat{\beta} - \beta)\|_2 + \frac{2}{n}\|(\Gamma^\dagger)^T(I_p - P_{[k]})X^T\epsilon\|_\infty\|\Gamma(\hat{\beta} - \beta)\|_1 \\
&\leq \frac{2}{n}\|P_{[k]}X^T\epsilon\|_2\|\hat{\beta} - \beta\|_2 + \frac{\lambda_1}{2}\|\Gamma(\hat{\beta} - \beta)\|_1
\end{aligned} \tag{2.62}$$

if we choose $\lambda_1 \geq \frac{8}{n}\|(\Gamma^\dagger)^T(I_p - P_{[k]})X^T\epsilon\|_\infty$ with high probability.

In order to choose k, we need to bound $\frac{8}{n}\|(\Gamma^\dagger)^T(I_p - P_{[k]})X^T\epsilon\|_\infty$. Let $s_1', \ldots, s_m'$ be the columns of $(I_p - P_{[k]})\Gamma^\dagger$. Let e_j, $j \in [m]$, denote the j^{th} canonical basis element. As in

the proof of Theorem 6 of Wang et al. [2015], we have:

$$\begin{aligned}\|s_j'\|_2^2 &= \|(I_p - V_{[k]}V_{[k]}^T)V\Xi^\dagger U^T e_j\|_2^2 \\ &= \|\begin{bmatrix}0 & V_{-[k]}\end{bmatrix}\Xi^\dagger U^T e_j\|_2^2 = \left\|\sum_{i=k+1}^{p-1}\xi_i^{-1}\langle u_i, e_j\rangle v_i\right\|_2^2 \\ &= \sum_{i=k+1}^{p-1}\xi_i^{-2}\langle u_i, e_j\rangle^2 \leq \frac{2}{p}\sum_{i=k+1}^{p-1}\xi_i^{-2}\end{aligned}$$

where we made use of the fact that the left singular vectors $\{u_i\}_{i=1}^m$ of Γ, when Γ is the incidence matrix of the chain graph with p vertices, satisfy $\forall i \in [m] : \|u_i\|_\infty \leq \sqrt{\frac{2}{p}}$.

For the chain graph, the nonzero singular values ξ_i are such that

$$\xi_i^2 = 4\sin^2\left(\frac{\pi i}{2p}\right) = 2 - 2\cos\left(\frac{\pi i}{p}\right), \text{ for } i = 1, \ldots, p-1$$

Hence, as in Wang et al. [2015],

$$\begin{aligned}\max_{j\in[m]}\|s_j'\|_2^2 \leq \frac{2}{p}\sum_{i=k+1}^{p-1}\xi_i^{-2} &= \frac{1}{2p}\sum_{i=k+1}^{p-1}\sin^{-2}\left(\frac{\pi i}{2p}\right) \\ &\leq \frac{1}{2p}\int_k^p \sin^{-2}\left(\frac{\pi x}{2p}\right)dx = \frac{\cos\left(\frac{\pi k}{2p}\right)}{\pi\sin\left(\frac{\pi k}{2p}\right)} \leq \frac{4p}{\pi^2 k}\end{aligned}$$

where we used $\sin(x) \geq x/2$ and $\cos(x) \leq 1$ for $x \in [0, \pi/2]$.

Similar to Lemma 53, we can then select $\lambda_1 = \frac{64}{\pi}\sigma\sqrt{p/k}\sqrt{\frac{\gamma_{\max}(\Sigma)\log p}{n}}$. We also have $\frac{2}{n}\|P_{[k]}X^T\epsilon\|_2 \leq 4\sqrt{\frac{2\sigma^2\gamma_{\max}(\Sigma)k}{n}}$ with probability at least $1 - e^{-n/8} - e^{-k^2/2}$, as in Lemma 52. The rest of the proof is again identical to that of Theorem 33, and we obtain for any S that

$$\|\Sigma^{1/2}(\hat{\beta} - \beta^*)\|_2^2 \lesssim \frac{\sigma^2\gamma_{\max}(\Sigma)\left[\frac{1+\delta^2}{n} + \frac{k}{n}\right] + \lambda_1^2|S|k_S^{-2}}{\gamma_{\min}\left(\frac{1}{64}\Sigma + \lambda_2 L\right)} + \lambda_1\|(\Gamma\beta^*)_{-S}\|_1$$

with high probability. By setting $S = \emptyset$ and choosing k such that

$$\frac{\sigma^2 \gamma_{\max}(\Sigma) k}{\gamma_{\min}\left(\frac{1}{64}\Sigma + \lambda_2 L\right) n} \asymp \lambda_1 \|\Gamma\beta^*\|_1 \asymp \sigma \sqrt{\frac{p\gamma_{\max}(\Sigma)\log p}{kn}} \|\Gamma\beta^*\|_1$$

we obtain

$$k \asymp \left(\frac{p \log p \|\Gamma\beta^*\|_1^2 \gamma_{\min}^2\left(\frac{1}{64}\Sigma + \lambda_2 L\right) n}{\sigma^2 \gamma_{\max}(\Sigma)} \right)^{1/3}$$

and with this choice of k

$$\frac{\sigma^2 \gamma_{\max}(\Sigma) k}{\gamma_{\min}\left(\frac{1}{64}\Sigma + \lambda_2 L\right) n} \asymp \frac{(\sigma^2 \gamma_{\max}(\Sigma) \|\Gamma\beta^*\|_1)^{2/3}}{\gamma_{\min}^{1/3}\left(\frac{1}{64}\Sigma + \lambda_2 L\right)} \frac{(p \log p)^{1/3}}{n^{2/3}}$$

□

We will often use the following lemma to compare probabilities involving two Gaussian vectors.

Lemma 49 (Anderson's Gaussian comparison inequality Anderson [1955])**.** *Let X and Y be two zero-mean Gaussian vectors with covariance Σ_X and Σ_Y respectively. If $\Sigma_Y - \Sigma_X$ is positive semi-definite, then for any convex set C satisfying $C = -C$,*

$$\mathbb{P}(X \in C) \geq \mathbb{P}(Y \in C)$$

Lemma 50 (Lemma 34)**.** *If $X \in \mathbb{R}^{n \times p}$ has i.i.d. $N(0, \Sigma)$ rows and $m \geq 2$, $n \geq 10$, then the event*

$$\left\{ \forall v \in \mathbb{R}^p : \frac{\|Xv\|_2}{\sqrt{n}} \geq \frac{1}{4}\|\Sigma^{1/2} v\|_2 - 3\sqrt{\frac{\gamma_{max}(\Sigma) n_c}{n}} \|v\|_2 - 6\sqrt{2}\rho(\Gamma)\sqrt{\frac{\gamma_{max}(\Sigma)\log p}{n}} \|\Gamma v\|_1 \right\}$$

holds with probability at least $1 - c_1 \exp(-nc_2)$, for some universal constants $c_1, c_2 > 0$.

Proof. We follow the proof outline of Raskutti et al. [2010]. First note that we can restrict

our attention to $v \in \mathbb{R}^p$ satisfying $\|\Sigma^{1/2}v\|_2 = 1$, as the inequality that defines the event above is invariant to scaling of v. Define

$$V(r,s) := \{v \in \mathbb{R}^p : \|\Sigma^{1/2}v\|_2 = 1, \|\Gamma v\|_1 \leq r, \|v\|_2 \leq s\}$$

and

$$M(r,s,X) := \sup_{v \in V(r,s)} \left(1 - \frac{\|Xv\|_2}{\sqrt{n}}\right)$$

Bounding the expectation $\mathbb{E}(M(r,s,X))$: By an application of Gordon's inequality (see Section 4.2 of Raskutti et al. [2010] for the details),

$$\begin{aligned}\mathbb{E}\left(\sup_{v \in V(r,s)} (-\|Xv\|_2)\right) = \mathbb{E}\left(\sup_{v \in V(r,s)} \inf_{u \in S^{n-1}} u^T X v\right) &\leq \mathbb{E}\left(\sup_{v \in V(r,s)} \inf_{u \in S^{n-1}} g^T u + h^T \Sigma^{1/2} v\right) \\ &= -\mathbb{E}\|g\|_2 + \mathbb{E}\left(\sup_{v \in V(r,s)} h^T \Sigma^{1/2} v\right)\end{aligned}$$

where $g \sim N(0, I_n)$ independent of $h \sim N(0, I_p)$. We know that $\mathbb{E}\|g\|_2 \geq \frac{3}{4}\sqrt{n}$ when $n \geq 10$, so we just need to upper bound $\mathbb{E}\left(\sup_{v \in V(r,s)} h^T \Sigma^{1/2} v\right)$.

Since $\Pi + \Gamma^\dagger \Gamma = I_p$,

$$h^T \Sigma^{1/2} v = h^T \Sigma^{1/2} (\Pi + \Gamma^\dagger \Gamma) v \leq \|\Pi \Sigma^{1/2} h\|_2 \|v\|_2 + \|(\Gamma^\dagger)^T \Sigma^{1/2} h\|_\infty \|\Gamma v\|_1$$

and by definition of $V(r,s)$ we have $\|v\|_2 \leq s$ and $\|\Gamma v\|_1 \leq r$ for all $v \in V(r,s)$, so we obtain

$$\mathbb{E}\left(\sup_{v \in V(r,s)} h^T \Sigma^{1/2} v\right) \leq s\mathbb{E}\|\Pi \Sigma^{1/2} h\|_2 + r\mathbb{E}\|(\Gamma^\dagger)^T \Sigma^{1/2} h\|_\infty$$

Note that the spectral decomposition of $\Pi = U \Lambda U^T$, where U is an orthogonal matrix, is such that Λ is a diagonal matrix with n_c ones and $p - n_c$ zeros on the diagonal. Since

$\gamma_{\max}(\Sigma)\Pi - \Pi\Sigma\Pi$ is positive semi-definite, by Lemma 49 we know that $\|\sqrt{\gamma_{\max}(\Sigma)}\Pi h\|_2$ stochastically dominates $\|\Pi\Sigma^{1/2}h\|_2$, and hence

$$\begin{aligned}
\mathbb{E}\|\Pi\Sigma^{1/2}h\|_2 &\leq \sqrt{\gamma_{\max}(\Sigma)}\mathbb{E}\|\Pi h\|_2 \\
&= \sqrt{\gamma_{\max}(\Sigma)}\mathbb{E}\|U\Lambda U^T h\|_2 \\
&= \sqrt{\gamma_{\max}(\Sigma)}\mathbb{E}\|\Lambda h\|_2 \\
&= \sqrt{\gamma_{\max}(\Sigma)}\mathbb{E}\sqrt{h_1^2 + \cdots + h_{n_c}^2} \\
&\leq \sqrt{\gamma_{\max}(\Sigma) n_c}
\end{aligned}$$

where we have used Jensen's inequality and the rotational invariance of the standard Gaussian distribution in the above derivations.

By Exercise 2.12 b) of Wainwright [2019], we also have for $m \geq 2$:

$$\begin{aligned}
\mathbb{E}\|(\Gamma^\dagger)^T\Sigma^{1/2}h\|_\infty &= \mathbb{E}\max_{j\in[m]}|\langle s_j, \Sigma^{1/2}h\rangle| \\
&\leq 2\sqrt{\gamma_{\max}(\Sigma)}\rho(\Gamma)\sqrt{\log m} \leq 2\sqrt{2}\sqrt{\gamma_{\max}(\Sigma)}\rho(\Gamma)\sqrt{\log p}
\end{aligned}$$

since $\{\langle s_j, \Sigma^{1/2}h\rangle : j = 1, \cdots, m\}$ is a collection of m zero-mean Gaussian variables with variance at most $\gamma_{\max}(\Sigma)\max_{j\in[m]}\|s_j\|_2^2 = \gamma_{\max}(\Sigma)\rho(\Gamma)^2$ (and in the last inequality we used $m \leq p^2$).

We can therefore conclude

$$\mathbb{E}\left(-\inf_{v\in V(r,s)}\|Xv\|_2\right) \leq -\frac{3}{4}\sqrt{n} + s\sqrt{\gamma_{\max}(\Sigma)n_c} + 2\sqrt{2}r\sqrt{\gamma_{\max}(\Sigma)}\rho(\Gamma)\sqrt{\log p}$$

Dividing by $\sqrt{n}$ and adding 1 on both sides, we obtain

$$\mathbb{E}\left(M(r,s,X)\right) \leq \frac{1}{4} + s\sqrt{\frac{\gamma_{\max}(\Sigma)n_c}{n}} + 2\sqrt{2}r\sqrt{\gamma_{\max}(\Sigma)}\rho(\Gamma)\sqrt{\frac{\log p}{n}}$$

Concentration around the mean for $M(r,s,X)$: As $M(r,s,X)$ is a Lipschitz function of a Gaussian vector (see Section 4.3 of Raskutti et al. [2010] for details), for all $t > 0$ we have:

$$\mathbb{P}(|M(r,s,X) - \mathbb{E}M(r,s,X)| \geq t/2) \leq 2\exp(-nt^2/8)$$

Substituting $t = t(r,s) := \frac{1}{4} + s\sqrt{\frac{\gamma_{\max}(\Sigma)n_c}{n}} + 2\sqrt{2}r\sqrt{\gamma_{\max}(\Sigma)}\rho(\Gamma)\sqrt{\frac{\log p}{n}}$, we obtain

$$\mathbb{P}\left(M(r,s,X) \geq \frac{3t(r,s)}{2}\right) \leq 2\exp(-nt(r,s)^2/8)$$

Peeling: This part is adapted from Section 4.4 of Raskutti et al. [2010]. We have shown that

$$\mathbb{P}\left(\sup_{\substack{\|v\|_2\leq s, \|\Gamma v\|_1 \leq r \\ \|\Sigma^{1/2}v\|_2 = 1}} \left(1 - \frac{\|Xv\|_2}{\sqrt{n}}\right) \geq \frac{3}{8} + \frac{3}{2}\sqrt{\frac{\gamma_{\max}(\Sigma)n_c}{n}}s + 3\sqrt{2}\rho(\Gamma)\sqrt{\frac{\gamma_{\max}(\Sigma)\log p}{n}}r\right)$$
$$\leq 2\exp\left(-\frac{n}{18}\left(\frac{3}{8} + \frac{3}{2}s\sqrt{\frac{\gamma_{\max}(\Sigma)n_c}{n}} + 3\sqrt{2}\rho(\Gamma)r\sqrt{\frac{\gamma_{\max}(\Sigma)\log p}{n}}\right)^2\right)$$

Let $g_1(s) := \frac{3}{16} + \frac{3}{2}\sqrt{\frac{\gamma_{\max}(\Sigma)n_c}{n}}s$ and $g_2(r) := \frac{3}{16} + 3\sqrt{2}\rho(\Gamma)\sqrt{\frac{\gamma_{\max}(\Sigma)\log p}{n}}r$. We can rewrite the above as

$$\mathbb{P}\left(\sup_{\substack{\|v\|_2\leq s, \|\Gamma v\|_1 \leq r \\ \|\Sigma^{1/2}v\|_2 = 1}} \left(1 - \frac{\|Xv\|_2}{\sqrt{n}}\right) \geq g_1(s) + g_2(r)\right) \leq 2\exp\left(-\frac{n}{18}[g_1(s) + g_2(r)]^2\right)$$

Note that $g_1 \geq \mu$ and $g_2 \geq \mu$ where $\mu := \frac{3}{16}$. For $i = 1, 2, \cdots$, and $j = 1, 2, \cdots$, we

define the sets

$$A_{ij} := \{v \in \mathbb{R}^p : \|\Sigma^{1/2}v\|_2 = 1, 2^{i-1}\mu \leq g_1(\|v\|_2) < 2^i\mu, 2^{j-1}\mu \leq g_2(\|\Gamma v\|_1) < 2^j\mu\}$$

Also, we define the events

$$\mathcal{E}_{ij} := \left\{\exists v \in A_{ij} : 1 - \frac{\|Xv\|_2}{\sqrt{n}} \geq 2[g_1(\|v\|_2) + g_2(\|\Gamma v\|_1)]\right\}$$

as well as the event

$$\mathcal{E} := \left\{\exists v \in \mathbb{R}^p : \|\Sigma^{1/2}v\|_2 = 1 \text{ and } 1 - \frac{\|Xv\|_2}{\sqrt{n}} \geq 2\left[g_1(\|v\|_2) + g_2(\|\Gamma v\|_1)\right]\right\}$$

Note that $\mathcal{E} = \cup_{i=1}^{\infty} \cup_{j=1}^{\infty} \mathcal{E}_{ij}$. Our goal is to prove that $\mathbb{P}(\mathcal{E}) \leq c_1 \exp(-nc_2)$, from which the lemma follows.

If we have $v \in A_{ij}$ such that $1 - \frac{\|Xv\|_2}{\sqrt{n}} \geq 2[g_1(\|v\|_2) + g_2(\|\Gamma v\|_1)]$ holds, then by definition of A_{ij},

$$1 - \frac{\|Xv\|_2}{\sqrt{n}} \geq 2(2^{i-1}\mu + 2^{j-1}\mu) = 2^i\mu + 2^j\mu = g_1(g_1^{-1}(2^i\mu)) + g_2(g_2^{-1}(2^j\mu))$$

Again by definition of A_{ij}, $g_1(\|v\|_2) \leq 2^i\mu$ and $g_2(\|\Gamma v\|_1) \leq 2^j\mu$, and so

$$\|v\|_2 \leq g_1^{-1}(2^i\mu) \qquad \text{and} \qquad \|\Gamma v\|_1 \leq g_2^{-1}(2^j\mu)$$

Therefore, we must have

$$\begin{aligned}
\mathbb{P}(\mathcal{E}_{ij}) &\leq 2\exp\left(-\frac{n}{18}[g_1(g_1^{-1}(2^i\mu)) + g_2(g_2^{-1}(2^j\mu))]^2\right) \\
&= 2\exp\left(-\frac{n}{18}(2^i + 2^j)^2\mu^2\right) \\
&\leq 2\exp\left(-\frac{n}{18}2^{2i}\mu^2\right)\exp\left(-\frac{n}{18}2^{2j}\mu^2\right)
\end{aligned}$$

Hence,

$$\begin{aligned}
\mathbb{P}(\mathcal{E}) &\leq 2\sum_{i=1}^{\infty}\sum_{j=1}^{\infty}\exp\left(-\frac{n}{18}2^{2i}\mu^2\right)\exp\left(-\frac{n}{18}2^{2j}\mu^2\right) \\
&= 2\left(\sum_{i=1}^{\infty}\exp\left(-\frac{n}{18}2^{2i}\mu^2\right)\right)^2 \\
&\leq 2\left(\sum_{i=1}^{\infty}\exp\left(-\frac{ni}{18}\mu^2\right)\right)^2 \\
&= 2\left(\frac{\exp\left(-\frac{n}{18}\mu^2\right)}{1-\exp\left(-\frac{n}{18}\mu^2\right)}\right)^2 \leq c_1\exp(-nc_2)
\end{aligned}$$

□

Corollary 51. *Under the settings of Lemma 50,*

$$\frac{\|\tilde{X}v\|_2^2}{n} \geq v^T\left(\frac{1}{64}\Sigma + \lambda_2 L\right)v - \frac{72\gamma_{max}(\Sigma)n_c}{n}\|v\|_2^2 - 576\rho(\Gamma)^2\frac{\gamma_{max}(\Sigma)\log p}{n}\|\Gamma v\|_1^2$$

holds for all $v \in \mathbb{R}^p$ with probability at least $1 - c_1\exp(-nc_2)$

Proof. We argue in a manner similar to the proof of Theorem 7.16 in Wainwright [2019]. For any real numbers a, b, c such that $c \geq \max(a-b, 0)$, we claim that $c^2 \geq (1-\delta)^2a^2 - \frac{b^2}{\delta^2}$ for any $\delta \in (0,1)$. This is because if $b \geq a\delta$, then $(1-\delta)^2a^2 - \frac{b^2}{\delta^2} \leq a^2[(1-\delta)^2 - 1] \leq 0 \leq c$, and if $b < a\delta$, then since $c \geq a - b$, we have $c \geq a - a\delta = (1-\delta)a$.

Letting $c = \frac{\|Xv\|_2}{\sqrt{n}}$, $a = \frac{1}{4}\|\Sigma^{1/2}v\|_2$, $b = 3\sqrt{\frac{\gamma_{\max}(\Sigma)n_c}{n}}\|v\|_2 + 6\sqrt{2}\rho(\Gamma)\sqrt{\frac{\gamma_{\max}(\Sigma)\log p}{n}}\|\Gamma v\|_1$ and $\delta = \frac{1}{2}$, we obtain for all $v \in \mathbb{R}^p$ with probability at least $1 - c_1\exp(-nc_2)$:

$$\begin{aligned}\frac{\|Xv\|_2^2}{n} &\geq \frac{1}{64}\|\Sigma^{1/2}v\|_2^2 - 36\left(\sqrt{\frac{\gamma_{\max}(\Sigma)n_c}{n}}\|v\|_2 + 2\sqrt{2}\rho(\Gamma)\sqrt{\frac{\gamma_{\max}(\Sigma)\log p}{n}}\|\Gamma v\|_1\right)^2 \\ &\geq \frac{1}{64}\|\Sigma^{1/2}v\|_2^2 - 72\frac{\gamma_{\max}(\Sigma)n_c}{n}\|v\|_2^2 - 576\rho(\Gamma)^2\frac{\gamma_{\max}(\Sigma)\log p}{n}\|\Gamma v\|_1^2\end{aligned}$$

By adding $\lambda_2 v^T L v$ to both sides, we obtain what we need to prove. □

Lemma 52 (High-probability bound on $\|\Pi X^T\epsilon\|_2$)**.** *For any $\delta > 0$,*

$$\|\Pi X^T\epsilon\|_2 \leq \sqrt{2\sigma^2 n\gamma_{max}(\Sigma)}(\sqrt{n_c} + \delta)$$

with probability at least $1 - e^{-n/8} - e^{-\delta^2/2}$.

Proof. We make use of the fact that X and ϵ are independent. Note that $X\Pi$ has i.i.d. $N(0, \Pi\Sigma\Pi)$ rows, which we denote by $\tilde{x}_1, \cdots, \tilde{x}_n$. Then

$$\|\Pi X^T\epsilon\|_2 = \left\|\sum_{i=1}^n \epsilon_i\tilde{x}_i\right\|_2 = \|\epsilon\|_2\left\|\frac{1}{\|\epsilon\|_2}\sum_{i=1}^n \epsilon_i\tilde{x}_i\right\|_2$$

which has the same distribution as $\|\epsilon\|_2\|\tilde{x}\|_2$, where $\tilde{x} \sim N(0, \Pi\Sigma\Pi)$ is independent of ϵ. Since $\gamma_{\max}(\Sigma)\Pi - \Pi\Sigma\Pi$ is positive semi-definite, by Lemma 49, $\|\tilde{x}\|_2$ is stochastically dominated by $\sqrt{\gamma_{\max}(\Sigma)}\|\Pi h\|_2$ (where $h \sim N(0, I_p)$), which in turn has the same distribution as $\sqrt{\gamma_{\max}(\Sigma)}\|h'\|_2$ where $h' \sim N(0, I_{n_c})$.

By an application of a concentration inequality for Lipschitz functions of Gaussian vec-

tors, we have for any $\delta > 0$ (see Example 2.28 of Wainwright [2019]):

$$\mathbb{P}(\|h'\|_2 \geq \sqrt{n_c} + \delta) \leq e^{-\delta^2/2}$$

and we also have $\|\epsilon\|_2 \leq \sigma\sqrt{2n}$ with probability at least $1 - e^{-n/8}$ (see Example 2.11 of Wainwright [2019]). Combining all the pieces yields the result. □

Lemma 53 (Choice of λ_1)**.** *With probability at least* $1 - \frac{2}{m} - e^{-n/8}$*, we have*

$$\|(\Gamma^\dagger)^T X^T \epsilon\|_\infty \leq 4\sigma\rho(\Gamma)\sqrt{\gamma_{max}(\Sigma) n \log p}$$

and hence λ_1 *should be chosen such that* $\lambda_1 \geq 32\sigma\rho(\Gamma)\sqrt{\frac{\gamma_{max}(\Sigma)\log p}{n}}$

Proof. Recall that the columns of $\Gamma^\dagger \in \mathbb{R}^{p\times m}$ are denoted as $s_1, \cdots, s_m$, and let the rows of X be $x_1, \cdots, x_n$, which by assumption are i.i.d. $N(0, \Sigma)$ vectors.

For any $t > 0$:

$$\begin{aligned}
&\mathbb{P}(\|(\Gamma^\dagger)^T X^T \epsilon\|_\infty \geq t) \\
&= \mathbb{P}\left(\max_{j\in[m]} \left|\left\langle s_j, \sum_{i=1}^n \epsilon_i x_i \right\rangle\right| \geq t\right) \\
&= \mathbb{P}\left(\|\epsilon\|_2 \max_{j\in[m]} \left|\left\langle s_j, \frac{1}{\|\epsilon\|_2}\sum_{i=1}^n \epsilon_i x_i \right\rangle\right| \geq t\right) \\
&\leq \mathbb{P}\left(\sqrt{2n}\sigma \max_{j\in[m]} \left|\left\langle s_j, \frac{1}{\|\epsilon\|_2}\sum_{i=1}^n \epsilon_i x_i \right\rangle\right| \geq t\right) + \mathbb{P}(\|\epsilon\|_2 > \sigma\sqrt{2n})
\end{aligned}$$

Using the same trick as in Lemma 52, $x := \frac{1}{\|\epsilon\|_2}\sum_{i=1}^n \epsilon_i x_i \sim N(0, \Sigma)$ independent of ϵ. Also, we note again that $\mathbb{P}(\|\epsilon\|_2 > \sigma\sqrt{2n}) \leq e^{-n/8}$. Hence, $\mathbb{P}(\|(\Gamma^\dagger)^T X^T \epsilon\|_\infty \geq t)$ is

bounded above by

$$\mathbb{P}\left(\sqrt{2n}\sigma \max_{j\in[m]} |\langle s_j, x\rangle| \geq t\right) + e^{-n/8} \leq \mathbb{P}\left(\sqrt{2n\gamma_{\max}(\Sigma)}\sigma \max_{j\in[m]} |\langle s_j, g\rangle| \geq t\right) + e^{-n/8}$$

where $g \sim N(0, I_p)$ and we used Lemma 49 in the last inequality. Since $\{\langle s_j, g\rangle : j \in [m]\}$ are Normal variables with variance at most $\rho(\Gamma)^2$, by applying the union bound, the expression above can be bounded above by

$$2\exp\left(-\frac{t^2}{4\gamma_{\max}(\Sigma)n\sigma^2\rho(\Gamma)^2} + \log m\right) + e^{-n/8}$$

If t is chosen such that $t^2 = 8\log(m)\gamma_{\max}(\Sigma)n\sigma^2\rho(\Gamma)^2$, we can conclude that

$$\|(\Gamma^\dagger)^T X^T \epsilon\|_\infty \leq 2\sqrt{2}\sigma\rho(\Gamma)\sqrt{\gamma_{\max}(\Sigma)n\log m} \leq 4\sigma\rho(\Gamma)\sqrt{\gamma_{\max}(\Sigma)n\log p}$$

with probability at least $1 - \frac{2}{m} - e^{-n/8}$, where we used $m \leq p^2$. □

Lemma 54 (Lemma 3 of Hütter and Rigollet [2016])**.** *If Γ is the incidence matrix of a graph $G = (V, E)$ with maximum degree d and $\emptyset \neq S \subseteq E$, then*

$$k_S^{-2} \leq 4\min(d, |S|)$$

Lemma 55 (Lower bound in Lemma 38)**.** *If Γ is the incidence matrix of the 2D grid, then $\rho(\Gamma) \gtrsim 1$.*

Proof. Let $N := \sqrt{p}$. In the proof of Proposition 4 of Hütter and Rigollet [2016], it was shown that $\Gamma^\dagger$ has $2N(N-1)$ columns $((s^{(1)}_{i,j})_{\substack{i\in[N-1]\\ j\in[N]}}, (s^{(2)}_{i,j})_{\substack{j\in[N-1]\\ i\in[N]}})$, each of which has column norm such that

$$\|s^{(\diamond)}_{i,j}\|_2^2 = \sum_{k=0}^{N-1}\sum_{l=1}^{N-1} \frac{1}{(\lambda_k + \lambda_l)^2}\langle v_l, d_i\rangle^2\langle v_k, e_j\rangle^2$$

where $\diamond \in \{1,2\}$, $\lambda_k = 2 - 2\cos\frac{k\pi}{N}$ for $0 \leq k \leq N-1$ (these are the eigenvalues of the Laplacian of the one-dimensional chain graph with N vertices), d_i is the i^{th} column of D_1^T where D_1 is the incidence matrix of the chain graph with N vertices, $e_1, \ldots, e_n$ are the canonical basis vectors of $\mathbb{R}^N$, and $v_k \in \mathbb{R}^N$ $(0 \leq k \leq N-1)$ are the orthonormal eigenvectors of the Laplacian of the one-dimensional chain graph with N vertices:

$$(v_0)_j = \frac{1}{\sqrt{N}}$$

$$(v_k)_j = \sqrt{\frac{2}{N}}\cos\left(\frac{(j+1/2)k\pi}{N}\right) \text{ for } 0 \leq j \leq N-1, 1 \leq k \leq N-1$$

Since $\rho(\Gamma)$ is defined as the maximum column norm of $\Gamma^\dagger$, we can bound it below by the column norm of $s_{i,j}^{(\diamond)}$ where $i = j = \diamond = 1$. We have:

$$\|s_{1,1}^{(1)}\|_2^2 = \sum_{k=0}^{N-1}\sum_{l=1}^{N-1} \frac{1}{\left(4 - 2\cos\frac{k\pi}{N} - 2\cos\frac{l\pi}{N}\right)^2}\langle v_l, d_1\rangle^2\langle v_k, e_1\rangle^2$$

Using the inequality $2 - 2\cos(x) \leq x^2$, we have $4 - 2\cos\frac{k\pi}{N} - 2\cos\frac{l\pi}{N} \leq \frac{\pi^2}{N^2}(k^2+l^2)$. Furthermore, note that

$$\begin{aligned}\langle v_l, d_1\rangle^2 &= \frac{2}{N}\left(\cos\frac{(5/2)l\pi}{N} - \cos\frac{(3/2)l\pi}{N}\right)^2 \\ &= \frac{2l^2\pi^2}{N^3}\sin^2(x')\end{aligned}$$

for some $x' \in \left[\frac{(3/2)l\pi}{N}, \frac{(5/2)l\pi}{N}\right]$, by the mean value theorem. Given the inequality $\sin(x) \geq x/2$ for $x \in [0, \pi/2]$, we can conclude that $\sin^2(x') \geq (x')^2/4 \geq \frac{9}{16}\frac{l^2\pi^2}{N^2}$ if we assume $l \leq \frac{N}{5}$, and so

$$\langle v_l, d_1\rangle^2 \geq \frac{9\pi^4}{8}\frac{l^4}{N^5}$$

if $l \leq N/5$. Moreover, $\langle v_k, e_1\rangle^2 = \frac{1}{N}$ if $k = 0$, and if we assume $k \leq \frac{2}{3\pi}N$, then $1 - \frac{9}{8}\frac{k^2\pi^2}{N^2} \geq \frac{1}{2}$ and an application of $1 - \cos(x) \leq \frac{x^2}{2}$ gives

$$\begin{aligned}\langle v_k, e_1\rangle^2 = \frac{2}{N}\left[\cos\left(\frac{3k\pi}{2N}\right)\right]^2 &= \frac{2}{N}\left[1 - \left(1 - \cos\frac{3k\pi}{2N}\right)\right]^2 \\ &\geq \frac{2}{N}\left[1 - \frac{9k^2\pi^2}{8N^2}\right]^2 \geq \frac{1}{2N}\end{aligned}$$

Hence, if $k \leq \frac{2}{3\pi}N$, we have $\langle v_k, e_1\rangle^2 \geq \frac{1}{2N}$. Let $c = \min\left(\frac{2}{3\pi}, \frac{1}{5}\right) = \frac{1}{5}$. Now,

$$\begin{aligned}\|s_{1,1}^{(1)}\|_2^2 &\geq \sum_{k=0}^{\lfloor cN\rfloor}\sum_{l=1}^{\lfloor cN\rfloor} \frac{1}{\left(4 - 2\cos\frac{k\pi}{N} - 2\cos\frac{l\pi}{N}\right)^2}\langle v_l, d_1\rangle^2\langle v_k, e_1\rangle^2 \\ &\geq \sum_{k=0}^{\lfloor cN\rfloor}\sum_{l=1}^{\lfloor cN\rfloor}\left(\frac{N^4}{\pi^4(k^2+l^2)^2}\right)\left(\frac{9\pi^4}{8}\frac{l^4}{N^5}\right)\left(\frac{1}{2N}\right) \\ &= \frac{9}{16N^2}\sum_{l=1}^{\lfloor cN\rfloor}\sum_{k=0}^{\lfloor cN\rfloor}\frac{l^4}{(k^2+l^2)^2}\end{aligned}$$

Since $\frac{1}{(k^2+l^2)^2}$ is a decreasing function of k,

$$\begin{aligned}\|s_{1,1}^{(1)}\|_2^2 &\geq \frac{9}{16N^2}\sum_{l=1}^{\lfloor cN\rfloor} l^4\int_0^{cN}\frac{1}{(x^2+l^2)^2}dx \\ &= \frac{9}{16N^2}\sum_{l=1}^{\lfloor cN\rfloor} l^4\frac{(l^2+c^2N^2)\arctan(cN/l) + cNl}{2l^3(l^2+c^2N^2)} \\ &= \frac{9}{32N^2}\sum_{l=1}^{\lfloor cN\rfloor} l\frac{(l^2+c^2N^2)\arctan(cN/l) + cNl}{l^2+c^2N^2} \\ &\geq \frac{9}{32N^2}\sum_{l=1}^{\lfloor cN\rfloor} l\arctan(c) \\ &= \frac{9\arctan(c)}{32N^2}\frac{\lfloor cN\rfloor(\lfloor cN\rfloor+1)}{2} \gtrsim 1\end{aligned}$$

□

Lemma 56 (Lower bound in Lemma 40)**.** *If* Γ *is the incidence matrix of the* r*-dimensional grid for* $r \geq 3$*, then* $\rho(\Gamma) \geq c(r)$*, where* $c(r)$ *is a constant depending only on* r*.*

Proof. Note that the $\gtrsim$ sign is used in this proof to omit constant multipliers that may depend on r. Similarly to the previous lemma, it is sufficient to lower bound

$$\|s_{\mathbf{1}}^{(1)}\|_2^2 = \sum_{l=1}^{N-1} \sum_{k_1=0}^{N-1} \cdots \sum_{k_{r-1}=0}^{N-1} \frac{\langle v_l, d_1\rangle^2 \prod_{j=1}^{r-1} \langle v_{k_j}, e_1\rangle^2}{(\lambda_l + \sum_{j=1}^{r-1} \lambda_{k_j})^2}$$

where d_1, e_1 as well as $\lambda_0, \dots \lambda_{N-1}$ and $v_0, \dots, v_{N-1}$ are defined in relation to the chain graph with N vertices as in the previous lemma.

By applying the inequality $2 - 2\cos(x) \leq x^2$, we have

$$\left(\lambda_l + \sum_{j=1}^{r-1} \lambda_{k_j}\right)^2 \leq \left(2 - 2\cos\frac{l\pi}{N} + \sum_{j=1}^{r-1}\left(2 - 2\cos\frac{k_j\pi}{N}\right)\right)^2 \leq \frac{\pi^4}{N^4}\left(l^2 + \sum_{j=1}^{r-1} k_j^2\right)^2$$

Also, $k \leq \frac{2}{3\pi}N$ implies $\langle v_k, e_1\rangle^2 \geq \frac{1}{2N}$, and $l \leq N/5$ implies $\langle v_l, d_1\rangle^2 \geq \frac{9\pi^4}{8}\frac{l^4}{N^5}$. Hence, if we define $c = \min(\frac{2}{3\pi}, \frac{1}{5}) = \frac{1}{5}$ as in the previous lemma,

$$\begin{aligned}
\|s_{\mathbf{1}}^{(1)}\|_2^2 &\gtrsim \frac{1}{N^r} \sum_{l=1}^{\lfloor cN\rfloor} \sum_{k_1=0}^{\lfloor cN\rfloor} \cdots \sum_{k_{r-1}=0}^{\lfloor cN\rfloor} \frac{l^4}{(l^2 + \sum_{j=1}^{r-1} k_j^2)^2} \\
&\geq \frac{1}{N^r} \sum_{l=1}^{\lfloor cN\rfloor} \int_{0 \leq x_j \leq cN, j=1,\dots,r-1} \frac{l^4}{(l^2 + \|x\|_2^2)^2} dx \\
&\geq \frac{1}{N^r} \sum_{l=1}^{\lfloor cN\rfloor} \int_{\|x\|_2 \leq cN} \frac{l^4}{(l^2 + \|x\|_2^2)^2} dx \\
&= \frac{1}{N^r} \sum_{l=1}^{\lfloor cN\rfloor} \int_0^{cN} \int_{S_{r-2}} \frac{l^4 R^{r-2}}{(l^2 + R^2)^2} d\sigma_{r-2}(u) dR
\end{aligned}$$

where we changed to polar coordinates in the last equality; here, S_{r-2} is the unit sphere in

$\mathbb{R}^{r-1}$, and σ_{r-2} is a measure on S_{r-2} such that, if $A \subseteq S_{r-2}$ is a Borel set and $\tilde{A}$ is the set of all points ru with $0 < r < 1$ and $u \in A$, then $\sigma_{r-2}(A) = (r-1)m_{r-1}(\tilde{A})$, where m_{r-1} is the Lebesgue measure on $\mathbb{R}^{r-1}$ (see Exercise 6, Chapter 8 of Rudin [1974]). We continue:

$$\begin{aligned}
\|s_{\mathbf{1}}^{(1)}\|_2^2 &\gtrsim \frac{1}{N^r} \sum_{l=1}^{\lfloor cN \rfloor} \int_0^{cN} \frac{l^4 R^{r-2}}{(l^2+R^2)^2} dR \\
&\geq \frac{1}{N^r} \sum_{l=1}^{\lfloor cN \rfloor} \int_{cN/2}^{cN} \frac{l^4 R^{r-2}}{(l^2+R^2)^2} dR \\
&\gtrsim \frac{1}{N^3} \sum_{l=1}^{\lfloor cN \rfloor} \int_{cN/2}^{cN} \frac{l^4 R}{(l^2+R^2)^2} dR
\end{aligned}$$

where we used the fact that $r \geq 3$. Note that $\int_a^b \frac{R}{(l^2+R^2)^2} dR = \frac{b^2-a^2}{2(b^2+l^2)(a^2+l^2)}$ and hence

$$\|s_{\mathbf{1}}^{(1)}\|_2^2 \gtrsim \frac{1}{N^3} \sum_{l=1}^{\lfloor cN \rfloor} \frac{l^4 N^2}{(l^2+N^2)^2} \geq \frac{1}{N} \int_0^{N/10} \frac{l^2}{(l^2+N^2)^2} dl$$

where we used the fact that $\frac{l^2}{(l^2+N^2)^2}$ is increasing in l. Since

$$\int_0^{N/10} \frac{l^2}{(l^2+N^2)^2} dl = \frac{151 - 1515 \arctan(1/10)}{1010} N$$

and $151 - 1515 \arctan(1/10) > 0$, the proof is complete. □

2.5.2 The interior point method on the dual objective

For the special case where the design matrix is the identity and $\lambda_2 = 0$, Kim et al. [2009] applies the interior point method on the dual objective. Similarly, we can apply interior point method to solve our more general dual objective

$$\hat{\beta} := \arg\min_{\beta \in \mathbb{R}^p} \frac{1}{2} \|\tilde{Y} - \tilde{X}\beta\|_2^2 + \lambda_1 \|\Gamma\beta\|_1 \tag{2.63}$$

We will specify the update directions, update step size, and measure of suboptimality. Let $\mathbf{1}_m$ denote the vector in $\mathbb{R}^m$ with all entries equal to 1. The dual problem only has inequality constraints $f_1(u) = u - \lambda_1 \mathbf{1}_m \leq 0$ and $f_2(u) = -u - \lambda_1 \mathbf{1}_m \leq 0$. Let μ_1, μ_2 be the dual variables corresponding to f_1, f_2. We apply the standard Newton's updates on the perturbed KKT conditions (by parameter t) of this dual problem. That is, the directions of the updates $(\Delta u, \Delta\mu_1, \Delta\mu_2)$ are solutions of the following linear system:

The residuals are:

$$r_t(u, \mu_1, \mu_2) = \begin{bmatrix} \check{\Gamma}\check{\Gamma}^T u - \check{\Gamma}\check{Y} + \mu_1 - \mu_2 \\ -\text{diag}(\mu_1) f_1(u) - \frac{1}{t}\mathbf{1}_m \\ -\text{diag}(\mu_2) f_2(u) - \frac{1}{t}\mathbf{1}_m \end{bmatrix} \tag{2.64}$$

By Newton's method, we need to solve:

$$\nabla r_t(u, \mu_1, \mu_2) \begin{bmatrix} \Delta u \\ \Delta\mu_1 \\ \Delta\mu_2 \end{bmatrix} = -r_t(u, \mu_1, \mu_2) \tag{2.65}$$

That simplifies to 3 linear equations below, where divisions between vectors are element-wise:

$$\left[\check{\Gamma}\check{\Gamma}^T - \text{diag}(\mu_1/f_1(u)) - \text{diag}(\mu_2/f_2(u))\right] \Delta u = -\left[\check{\Gamma}\check{\Gamma}^T u - \check{\Gamma}\check{Y} - \frac{\mathbf{1}_m}{t f_1(u)} + \frac{\mathbf{1}_m}{t f_2(u)}\right] \tag{2.66}$$

$$\Delta\mu_1 = -\left[\text{diag}(\mu_1/f_1(u))\Delta u + \mu_1 + \frac{\mathbf{1}_m}{t f_1(u)}\right] \tag{2.67}$$

$$\Delta\mu_2 = -\left[-\text{diag}(\mu_2/f_2(u))\Delta u + \mu_2 + \frac{\mathbf{1}_m}{t f_2(u)}\right] \tag{2.68}$$

The step size for each update are computed in a standard way as Section 11.7.3 of Boyd et al. [2004]. We apply standard backtracking line search to find the step size s for the updates. Choose parameters $\alpha, \gamma \in (0, 1)$ for backtracking. Denote the updates as (u^+, μ_1^+, μ_2^+). For example, $u^+ = u + s\Delta u$. To ensure the updates to be feasible, we first make sure that $\mu_1^+, \mu_2^+ \geq 0$. That is, we set $s_{\max} = \min\{1, \min\{-\mu_{1i}/\Delta\mu_{1i} | \Delta\mu_{1i} < 0\}, \min\{-\mu_{2i}/\Delta\mu_{2i} | \Delta\mu_{2i} < 0\}\}$. Next, continuously set $s = \gamma s$ until $f_1(u^+), f_2(u^+) < 0$. Finally, set $s = \gamma s$ until $\|r_t(u^+, \mu_1^+, \mu_2^+)\|_2 \leq (1 - \alpha s)\|r_t(u, \mu_1, \mu_2)\|_2$.

As a standard measure of suboptimality, the surrogate duality gap (see Section 11.7.2 of Boyd et al. [2004] for details) at the k^{th} iteration is:

$$\eta^{(k)} = -f_1(u^{(k)})^T \mu_1^{(k)} - f_2(u^{(k)})^T \mu_2^{(k)} \tag{2.69}$$

And the residual at the k^{th} iteration is:

$$r^{(k)} = r_t(u^{(k)}, \mu_1^{(k)}, \mu_2^{(k)}) \tag{2.70}$$

Our interior point algorithm is presented below.

Algorithm 2: Interior point method on the dual objective

Input: $\lambda_1, \lambda_2, \Gamma, Y, X$, tolerance ϵ

Output: $\hat{\beta}$ as defined in (2.63)

Initialize $u^{(0)} = 0$, $\mu_1^{(0)}, \mu_2^{(0)} > 0$, $\tau > 1$

while $r^{(k)} > \epsilon$ *or* $\eta^{(k)} > \epsilon$ **do**

  Set $t = 2\tau m/\eta^{(k)}$

  Compute update direction $(\Delta u, \Delta\mu_1, \Delta\mu_2)$ as in (2.66), (2.67), (2.68)

  Determine step size s by using α, γ backtracking line search

  Update:

$$u^{(k+1)} = u^{(k)} + s\Delta u$$
$$\mu_1^{(k+1)} = \mu_1^{(k)} + s\Delta\mu_1$$
$$\mu_2^{(k+1)} = \mu_2^{(k)} + s\Delta\mu_2$$

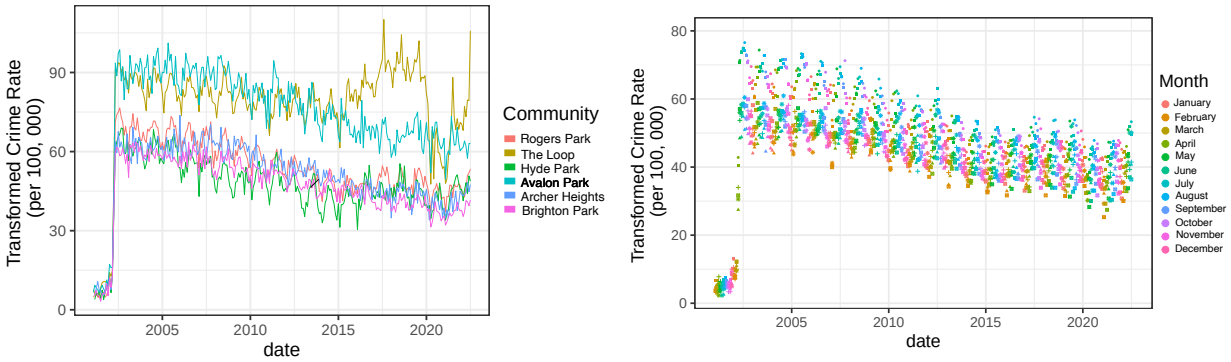
Compute $\hat{\beta} \leftarrow \tilde{X}^{\dagger}(\check{Y} - \check{\Gamma}^T u)$

Return $\hat{\beta}$

2.5.3 Additional details on data processing

Chicago Crime Data. As per the main text, statistics on the number of crimes per community between 2001 and 2022 are available on the city's data portal. For the purpose of our analysis, we consider the data between 2004 and 2022, since by preliminary inspection of the data, the first years of collection seem to have more missing data (see Figure 2.15a). We define the monthly crime rates as the number of crimes per 100,000 inhabitants. The latter are computed from the raw crime data by aggregating crime counts over neighborhoods and dividing by neighborhood population estimates found at the following link. These crime rates are usually modeled by Poisson distributions (see Osgood [2000]), which we transform here into a normal distribution through the use of an Anscombe transform. Examples of

the resulting estimates are displayed on Figure 2.15b. We note that the crime rates vary substantially over the years and across the communities, and are also subject to significant seasonal effects.



(a) Temporal evolution of the Anscombe transformed crime rate (per 100,000) as a function of time across 6 specific neighborhoods.

(b) Crime rate across neighborhoods, coloured by month of the year to highlight seasonal effects.

Figure 2.15: Anscombe transform of the number of crimes per month per 100,000 inhabitants in a few neighborhoods of Chicago. Note the seasonal effect in the crime rate and the consistent drop across neighborhoods during the coldest months of the year.

COVID Data. We consider the problem of predicting the number of COVID-19 cases 14 days in advance for a given county in California. As decribed in the main text, this could be an interesting use case for local public health decisions, such as for instance, trying to plan 2 weeks in advance appropriate resources at a local clinic. To this end, we used the New York Times-curated COVID database. The NYT COVID data provides a description of the total number of cases across all US counties, from January 2020 to October 2022 (time of writing). For the purpose of our analysis, we focus more specifically on analyzing new cases in the 25 densest California counties using data from June 1^{st}, 2020 to July 1^{st}, 2021. This time window was selected to provide more consistency in the epidemics dynamics: by June 2020, all counties in California had non zero daily incidence data. On the other hand, restricting the analysis to before July 2021 allows selecting a more cohesive window of time where the epidemic propagation was not dominated by (other unobserved) covariates, such

as the advent of new contagious strains of the virus (Delta in Summer 2021, and subsequently Omicron in Winter 2022). We pre-process the data and make it amenable to data analysis through the following steps:

1. Conversion of cumulative case counts to incidence data

2. Correction of aberrations and smoothing: we fix data aberrations (e.g. negative incidences, due to small errors in the reporting) by imposing the lower bound on the number of new cases to be 0. We further transform the incidence data using a seven-day rolling average so as to get rid of known spurious phenomena (e.g. the "weekend effect", by which the number of new cases is lower over the weekend but typically followed by a spike on the following Monday).

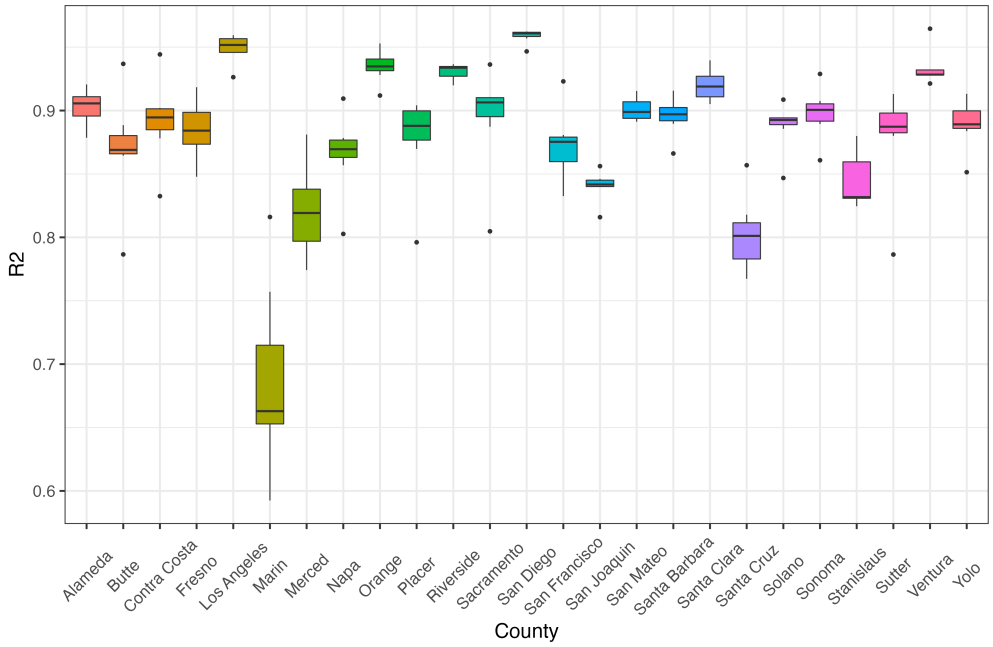


Figure 2.16: R^2 for the simple autoregressive model of Equation 2.49 on the seven different folds (see main text). Note that most models have R^2 of over 0.8, thus indicating the validity of the model.

3. Anscombe transform: We apply a variance stabilizing transform to transform incidence data (here modeled as a Poisson process, as per Agosto and Giudici [2020], Bu et al. [2021], Cori et al. [2013], Toharudin et al. [2020]): $\tilde{x} \leftarrow 2\sqrt{x + \frac{3}{8}}$.

REFERENCES

Alzheimer's disease mri preprocessed dataset. `https://www.kaggle.com/datasets/sachinkumar413/alzheimer-mri-dataset`. Accessed: September 15, 2022.

Arianna Agosto and Paolo Giudici. A Poisson autoregressive model to understand COVID-19 contagion dynamics. *Risks*, 8(3):77, 2020.

Anima Anandkumar, Dean P Foster, Daniel J Hsu, Sham M Kakade, and Yi-Kai Liu. A spectral algorithm for latent dirichlet allocation. *Advances in Neural Information Processing Systems*, 25, 2012.

Theodore W Anderson. The integral of a symmetric unimodal function over a symmetric convex set and some probability inequalities. *Proceedings of the American Mathematical Society*, 6(2):170–176, 1955.

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

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

Manuele Bicego, Pietro Lovato, Alessandro Perina, Marianna Fasoli, Massimo Delledonne, Mario Pezzotti, Annalisa Polverari, and Vittorio Murino. Investigating topic models' capabilities in expression microarray data classification. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 9(6):1831–1836, 2012.

Peter J Bickel and Elizaveta Levina. Covariance regularization by thresholding. *Annals of Statistics*, 36(6):2577–2604, 2008.

Peter J Bickel, Ya'acov Ritov, and Alexandre B Tsybakov. Simultaneous analysis of lasso and dantzig selector. *Annals of Statistics*, 37(4):1705–1732, 2009.

Lidong Bing, Wai Lam, and Tak-Lam Wong. Using query log and social tagging to refine queries based on latent topics. In *Proceedings of the 20th ACM International Conference on Information and Knowledge Management*, pages 583–592, 2011.

Xin Bing, Florentina Bunea, and Marten Wegkamp. A fast algorithm with minimax optimal guarantees for topic models with an unknown number of topics. *Bernoulli*, 2020a.

Xin Bing, Florentina Bunea, and Marten Wegkamp. Optimal estimation of sparse topic models. *The Journal of Machine Learning Research*, 21(1):7189–7233, 2020b.

David M Blei and John D Lafferty. Dynamic topic models. In *Proceedings of the 23rd International Conference on Machine Learning*, pages 113–120, 2006.

David M Blei and John D Lafferty. A correlated topic model of science. 2007.

David M Blei, Andrew Y Ng, and Michael I Jordan. Latent dirichlet allocation. *Journal of Machine Learning Research*, 3(Jan):993–1022, 2003.

Stephen Boyd, Stephen P Boyd, and Lieven Vandenberghe. *Convex optimization*. Cambridge University Press, 2004.

Fan Bu, Allison E Aiello, Alexander Volfovsky, and Jason Xu. Likelihood-based inference for partially observed stochastic epidemics with individual heterogeneity. *arXiv preprint arXiv:2112.07892*, 2021.

Peter Bühlmann and Sara van de Geer. *Statistics for high-dimensional data: methods, theory and applications*. Springer Science & Business Media, 2011.

T Tony Cai and Harrison H Zhou. Optimal rates of convergence for sparse covariance matrix estimation. *Annals of Statistics*, pages 2389–2420, 2012.

Benjamin J Callahan, Daniel B DiGiulio, Daniela S Aliaga Goltsman, Christine L Sun, Elizabeth K Costello, Pratheepa Jeganathan, Joseph R Biggio, Ronald J Wong, Maurice L Druzin, Gary M Shaw, et al. Replication and refinement of a vaginal microbial signature of preterm birth in two racially distinct cohorts of us women. *Proceedings of the National Academy of Sciences*, 114(37):9966–9971, 2017.

Zhenghao Chen, Ilya Soifer, Hugo Hilton, Leeat Keren, and Vladimir Jojic. Modeling multiplexed images with spatial-lda reveals novel tissue microenvironments. *Journal of Computational Biology*, 27(8):1204–1218, 2020.

Fan RK Chung and Fan Chung Graham. *Spectral graph theory*, volume 92. American Mathematical Society, 1997.

Anne Cori, Neil M Ferguson, Christophe Fraser, and Simon Cauchemez. A new framework and software to estimate time-varying reproduction numbers during epidemics. *American Journal of Epidemiology*, 178(9):1505–1512, 2013.

Álvaro Corral, Gemma Boleda, and Ramon Ferrer-i Cancho. Zipf's law for word frequencies: Word forms versus lemmas in long texts. *PloS one*, 10(7):e0129031, 2015.

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

Kenneth R Davidson and Stanislaw J Szarek. Local operator theory, random matrices and banach spaces. *Handbook of the geometry of Banach spaces*, 1(317-366):131, 2001.

Alexander Domahidi, Eric Chu, and Stephen Boyd. ECOS: An SOCP solver for embedded systems. In *2013 European Control Conference (ECC)*, pages 3071–3076. IEEE, 2013.

David Donoho and Victoria Stodden. When does non-negative matrix factorization give a correct decomposition into parts? *Advances in Neural Information Processing Systems*, 16, 2003.

Michael Elad, Peyman Milanfar, and Ron Rubinstein. Analysis versus synthesis in signal priors. *Inverse problems*, 23(3):947, 2007.

Ingo Feinerer, Kurt Hornik, and Maintainer Ingo Feinerer. Package 'tm'. *Corpus*, 10(1), 2015.

Julia Fukuyama, Kris Sankaran, and Laura Symul. Multiscale analysis of count data through topic alignment. *arXiv preprint arXiv:2109.05541*, 2021.

Rong Ge and James Zou. Intersecting faces: Non-negative matrix factorization with new guarantees. In *International Conference on Machine Learning*, pages 2295–2303. PMLR, 2015.

Nicolas Gillis and Stephen A Vavasis. Fast and robust recursive algorithmsfor separable nonnegative matrix factorization. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 36(4):698–714, 2013.

Yury Goltsev, Nikolay Samusik, Julia Kennedy-Darling, Salil Bhate, Matthew Hale, Gustavo Vazquez, Sarah Black, and Garry P Nolan. Deep profiling of mouse splenic architecture with codex multiplexed imaging. *Cell*, 174(4):968–981, 2018.

Anne Greenbaum, Ren-cang Li, and Michael L Overton. First-order perturbation theory for eigenvalues and eigenvectors. *SIAM review*, 62(2):463–482, 2020.

Thomas L Griffiths and Mark Steyvers. Finding scientific topics. *Proceedings of the National Academy of Sciences*, 101(suppl_1):5228–5235, 2004.

Adityanand Guntuboyina, Donovan Lieu, Sabyasachi Chatterjee, and Bodhisattva Sen. Adaptive risk bounds in univariate total variation denoising and trend filtering. *Annals of Statistics*, 48(1):205–229, 2020.

Mohamed Hebiri and Sara van de Geer. The Smooth-Lasso and other $\ell_1 + \ell_2$-penalized methods. *Electronic Journal of Statistics*, 5:1184–1226, 2011.

Thomas Hofmann. Probabilistic latent semantic indexing. In *Proceedings of the 22nd annual international ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 50–57, 1999.

Liangjie Hong and Brian D Davison. Empirical study of topic modeling in twitter. In *Proceedings of the first workshop on social media analytics*, pages 80–88, 2010.

Roger A Horn and Charles R Johnson. *Matrix analysis*. Cambridge University Press, 2012.

Jan-Christian Hütter and Philippe Rigollet. Optimal rates for total variation denoising. In *Conference on Learning Theory*, pages 1115–1146. PMLR, 2016.

Hamid Javadi and Andrea Montanari. Nonnegative matrix factorization via archetypal analysis. *Journal of the American Statistical Association*, 115(530):896–907, 2020.

P Ji, J Jin, ZT Ke, and W Li. Meta-analysis on citations for statisticians. *manuscript.[1, 10]*, 2021.

Jiashun Jin. Fast community detection by score. *The Annals of Statistics*, 2015.

Jiashun Jin, Zheng Tracy Ke, and Shengming Luo. Estimating network memberships by simplex vertex hunting. *arXiv preprint arXiv:1708.07852*, 12, 2017.

Zheng Tracy Ke and Minzhe Wang. Using svd for topic modeling. *Journal of the American Statistical Association*, pages 1–16, 2022.

Seung-Jean Kim, Kwangmoo Koh, Stephen Boyd, and Dimitry Gorinevsky. ℓ_1 trend filtering. *SIAM review*, 51(2):339–360, 2009.

Olga Klopp, Maxim Panov, Suzanne Sigalla, and Alexandre Tsybakov. Assigning topics to documents by successive projections. *arXiv preprint arXiv:2107.03684*, 2021.

E.D. Kolaczyk. Statistical analysis of network data: Methods and models. *Springer Series In Statistics*, page 386, 2009.

Vladimir Koltchinskii. *Oracle inequalities in empirical risk minimization and sparse recovery problems: École D'Été de Probabilités de Saint-Flour XXXVIII-2008*, volume 2033. Springer Science & Business Media, 2011.

Stephanie R Land and Jerome H Friedman. Variable fusion: A new adaptive signal regression method, 1997.

Caiyan Li and Hongzhe Li. Network-constrained regularization and variable selection for analysis of genomic data. *Bioinformatics*, 24(9):1175–1182, 2008.

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

Tianxi Li, Elizaveta Levina, and Ji Zhu. Prediction models for network-linked data. *Annals of Applied Statistics*, 13(1):132–164, 2019.

Yuan Li, Benjamin Mark, Garvesh Raskutti, and Rebecca Willett. Graph-based regularization for regression problems with highly-correlated designs. In *2018 IEEE Global Conference on Signal and Information Processing (GlobalSIP)*, pages 740–742. IEEE, 2018.

Yan Liu, Alexandru Niculescu-Mizil, and Wojciech Gryc. Topic-link lda: joint models of topic and author community. In *Proceedings of the 26th annual International Conference on Machine Learning*, pages 665–672, 2009.

Zongming Ma. Sparse principal component analysis and iterative thresholding. *The Annals of Statistics*, 2013.

Sahand Negahban, Bin Yu, Martin J Wainwright, and Pradeep Ravikumar. A unified framework for high-dimensional analysis of m-estimators with decomposable regularizers. *Advances in neural information processing systems*, 22, 2009.

David Newman, Chaitanya Chemudugunta, Padhraic Smyth, and Mark Steyvers. Analyzing entities and topics in news articles using statistical topic models. In *Intelligence and Security Informatics: IEEE International Conference on Intelligence and Security Informatics, ISI 2006, San Diego, CA, USA, May 23-24, 2006. Proceedings 4*, pages 93–104. Springer, 2006.

David Newman, Youn Noh, Edmund Talley, Sarvnaz Karimi, and Timothy Baldwin. Evaluating topic models for digital libraries. In *Proceedings of the 10th annual Joint Conference on Digital Libraries*, pages 215–224, 2010.

Calistus N Ngonghala, Hemaho B Taboe, Salman Safdar, and Abba B Gumel. Unraveling the dynamics of the Omicron and Delta variants of the 2019 coronavirus in the presence of vaccination, mask usage, and antiviral treatment. *Applied Mathematical Modelling*, 2022.

XuanLong Nguyen. Posterior contraction of the population polytope in finite admixture models. 2015.

Valeria Nikolaenko, Stratis Ioannidis, Udi Weinsberg, Marc Joye, Nina Taft, and Dan Boneh. Privacy-preserving matrix factorization. In *Proceedings of the 2013 ACM SIGSAC Conference on Computer & Communications Security*, pages 801–812, 2013.

Edouard Ollier and Vivian Viallon. Regression modelling on stratified data with the lasso. *Biometrika*, 104(1):83–96, 2017.

Francesco Ortelli and Sara van de Geer. Synthesis and analysis in total variation regularization. *arXiv preprint arXiv:1901.06418*, 2019.

Francesco Ortelli and Sara van de Geer. Prediction bounds for higher order total variation regularized least squares. *Annals of Statistics*, 49(5):2755–2773, 2021.

D Wayne Osgood. Poisson-based regression analysis of aggregate crime rates. *Journal of Quantitative Criminology*, 16(1):21–43, 2000.

Steven T Piantadosi. Zipf's word frequency law in natural language: A critical review and future directions. *Psychonomic Bulletin & Review*, 21:1112–1130, 2014.

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

Garvesh Raskutti, Martin J Wainwright, and Bin Yu. Restricted eigenvalue properties for correlated Gaussian designs. *Journal of Machine Learning Research*, 11:2241–2259, 2010.

Garvesh Raskutti, Martin J Wainwright, and Bin Yu. Minimax rates of estimation for high-dimensional linear regression over ℓ_q-balls. *IEEE transactions on information theory*, 57 (10):6976–6994, 2011.

Gabriel K Reder, Adamo Young, Jaan Altosaar, Jakub Rajniak, Noémie Elhadad, Michael Fischbach, and Susan Holmes. Supervised topic modeling for predicting molecular substructure from mass spectrometry. *F1000Research*, 10:Chem–Inf, 2021.

Walter Rudin. Real and Complex Analysis. *McGraw-Hill Inc.,*, 1974.

Walter Rudin et al. *Principles of mathematical analysis*, volume 3. McGraw Hill, New York, 1976.

Kris Sankaran and Susan P Holmes. Latent variable modeling for the microbiome. *Biostatistics*, 20(4):599–614, 2019.

Ville Satopaa, Jeannie Albrecht, David Irwin, and Barath Raghavan. Finding a" kneedle" in a haystack: Detecting knee points in system behavior. In *2011 31st International Conference on Distributed Computing Systems Workshops*, pages 166–171. IEEE, 2011.

Mark R Segal, Kam D Dahlquist, and Bruce R Conklin. Regression approaches for microarray data analysis. *Journal of Computational Biology*, 10(6):961–980, 2003.

Yiyuan She. *Sparse regression with exact clustering*. Stanford University, 2008.

Jian Tang, Zhaoshi Meng, Xuanlong Nguyen, Qiaozhu Mei, and Ming Zhang. Understanding the limiting factors of topic modeling via posterior contraction analysis. In *International Conference on Machine Learning*, pages 190–198. PMLR, 2014.

Yee Teh, Michael Jordan, Matthew Beal, and David Blei. Sharing clusters among related groups: Hierarchical dirichlet processes. *Advances in Neural Information Processing Systems*, 17, 2004.

Robert Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 58(1):267–288, 1996.

Robert Tibshirani, Michael Saunders, Saharon Rosset, Ji Zhu, and Keith Knight. Sparsity and smoothness via the fused lasso. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 67(1):91–108, 2005.

Ryan J Tibshirani and Jonathan Taylor. The solution path of the generalized lasso. *Annals of Statistics*, 39(3):1335–1371, 2011.

Toni Toharudin, Rezzy Eko Caraka, Rung Ching Chen, Nyityasmono Tri Nugroho, M Sueb, I Jaya, and R Pontoh. Bayesian Poisson model for COVID-19 in West Java, Indonesia. *Sylwan*, 164(6):279–290, 2020.

Paul Tseng. Convergence of a block coordinate descent method for nondifferentiable minimization. *Journal of Optimization Theory and Applications*, 109(3):475–494, 2001.

Sara A van de Geer. *Estimation and testing under sparsity*. Springer, 2016.

Sara A Van de Geer and Sara van de Geer. *Empirical Processes in M-estimation*, volume 6. Cambridge university press, 2000.

Martin J Wainwright. *High-dimensional statistics: A non-asymptotic viewpoint*, volume 48. Cambridge University Press, 2019.

Lingxiao Wang, Boxin Zhao, and Mladen Kolar. Differentially private matrix completion through low-rank matrix factorization. In *International Conference on Artificial Intelligence and Statistics*, pages 5731–5748. PMLR, 2023.

Yu-Xiang Wang, James Sharpnack, Alex Smola, and Ryan Tibshirani. Trend filtering on graphs. In *Artificial Intelligence and Statistics*, pages 1042–1050. PMLR, 2015.

Ruijia Wu, Linjun Zhang, and T Tony Cai. Sparse topic modeling: Computational efficiency, near-optimal algorithms, and statistical inference. *Journal of the American Statistical Association*, pages 1–13, 2022.

Bo Xin, Yoshinobu Kawahara, Yizhou Wang, and Wen Gao. Efficient generalized fused lasso and its application to the diagnosis of Alzheimer's disease. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 28, 2014.

Shinichi Yachida, Sayaka Mizutani, Hirotsugu Shiroma, Satoshi Shiba, Takeshi Nakajima, Taku Sakamoto, Hikaru Watanabe, Keigo Masuda, Yuichiro Nishimoto, Masaru Kubo, et al. Metagenomic and metabolomic analyses reveal distinct stage-specific phenotypes of the gut microbiota in colorectal cancer. *Nature Medicine*, 25(6):968–976, 2019.

Xiaolin Zheng, Weifeng Ding, Zhen Lin, and Chaochao Chen. Topic tensor factorization for recommender system. *Information Sciences*, 372:276–293, 2016.

George Kingsley Zipf. *The Psycho-biology of Language: An Introduction to Dynamic Philology*. Houghton-Mifflin, 1936.

Hui Zou and Trevor Hastie. Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 67(2):301–320, 2005.