DISSERTATION

STATISTICAL MODELING OF HIGH-DIMENSIONAL CATEGORICAL DATA WITH APPLICATIONS TO MUTATION FITNESS AND SPARSE TEXT TOPIC ANALYSIS

Submitted by

Bingying Dai

Department of Statistics

In partial fulfillment of the requirements

For the Degree of Doctor of Philosophy

Colorado State University

Fort Collins, Colorado

Summer 2025

Doctoral Committee:

Advisor: Yunpeng Zhao
Co-Advisor: Wen Zhou

Daniel Cooley
Tianjian Zhou
Nathaniel Blanchard

ABSTRACT

STATISTICAL MODELING OF HIGH-DIMENSIONAL CATEGORICAL DATA WITH APPLICATIONS TO MUTATION FITNESS AND SPARSE TEXT TOPIC ANALYSIS

The growing availability of large-scale categorical data has created a strong need for statistical methods capable of modeling high-dimensional discrete structures. Such data are common in fields like biological sequence analysis, natural language processing, and social network modeling, where observations often involve thousands of categorical or count-valued variables, exhibiting complex dependencies and high sparsity. Conventional statistical models, designed for continuous or low-dimensional settings, often fall short in capturing the latent structure and combinatorial complexity of such data. This dissertation introduces new statistical modeling frameworks and estimation techniques tailored for high-dimensional categorical data, supported by theoretical guarantees and validated through applications in protein sequence analysis and topic modeling.

The first part of the dissertation focuses on modeling mutational fitness in proteins, where predicting the effects of amino acid mutations is challenging due to the vast combinations of sites and amino acid types. We propose a new framework for analyzing protein sequences using the Potts model with node-wise high-dimensional multinomial regression. Our method identifies key site interactions and important amino acids, quantifying mutation effects through evolutionary energy derived from model parameters. It encourages sparsity in both site-wise and amino acid-wise dependencies through element-wise and group sparsity. We have established, for the first time to our knowledge, the ℓ_2 convergence rate for estimated parameters in the high-dimensional Potts model using sparse group Lasso, matching the existing minimax lower bound for high-dimensional linear models with a sparse group structure, up to a factor depending only on the multinomial nature of the Potts model. This theoretical guarantee enables accurate quantification of estimated energy changes. Additionally, we incorporate structural data into our model by applying penalty weights

across site pairs. Our method outperforms others in predicting mutation fitness, as demonstrated by comparisons with high-throughput mutagenesis experiments across 12 protein families.

The second part focus on topic modeling which is a fundamental technique for uncovering latent semantic structures in large text corpora. While traditional probabilistic models such as Latent Dirichlet Allocation and probabilistic Latent Semantic Indexing have been widely adopted, they often rely on assumptions that do not align well with the properties of real-world text data, particularly the pervasive presence of zero counts. These structural zeros, especially in short documents, often reflect more than random sampling variability and can indicate meaningful absence. To address these limitations, we propose a novel Zero-Inflated Poisson model that incorporates three essential components: a zero-inflation mechanism explicitly accounting for excess zeros that arise from structural rather than sampling sources; a functional link connecting the zero-inflation probability to the Poisson intensity to capture informative missingness related to topic prevalence, and document-level random effects accounting for unobserved heterogeneity across documents. An efficient alternating optimization algorithm is developed for intensity parameters estimation under a low-rank structure. We establish finite-sample error bounds for topic-word matrix recovery via a vertex hunting procedure. Empirical studies on synthetic datasets show that the model outperforms existing methods in sparse and heterogeneous settings. Application to a real-world corpus of statistical publications further confirms the model's ability to recover meaningful topics and track their evolution over time.

ACKNOWLEDGEMENTS

I extend my deepest gratitude to my co-advisor, Prof. Wen Zhou, for his invaluable insights, generous guidance, and profound perspectives that significantly shaped my approach to both research problems and broader academic understanding. I am also profoundly thankful to my advisor, Prof. Yunpeng Zhao, for his steadfast support, helpfulness, and kindness throughout my academic journey.

I sincerely appreciate the time, constructive feedback, and encouragement provided by my committee members: Prof. Daniel Cooley, Prof. Tianjian Zhou, and Prof. Nathaniel Blanchard. Their expertise was instrumental in refining this work. I am grateful for the invaluable guidance and collaborative experiences shared with scholars from external institutions: Dr. Yinan Lin, Dr. Kejue Jia, Prof. Zhao Ren, and Prof. Tianxi Li.

I would like to express my heartfelt appreciation to my parents for their unwavering support of my decision to embark on this journey. I am especially grateful to my mother for her constant support, endless care, and unconditional love. I also wish to thank my grandmother, who planted the seed of curiosity and exploration in my young heart and whom I hope would look upon me with pride. My deepest thanks also go to all my friends who have cared for me throughout the years, especially Yue Bai. Every shared moment, from the quiet rhythm of ordinary days to the laughter and the tears, has carried deep meaning and become an essential part of this journey.

DEDICATION

I would like to dedicate this thesis to my loved ones.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

Chapter 1

Introduction

High-dimensional categorical data have become increasingly prevalent in the era of data-driven science, emerging across a wide range of disciplines including computational biology, natural language processing, and the social sciences. Unlike continuous measurements, categorical observations, such as protein sequences with amino acids, or word frequencies among documents, are inherently discrete, often characterized by extreme sparsity, and structured in complex ways. These characteristics pose significant challenges to traditional statistical modeling, which typically relies on assumptions better suited to continuous or low-dimensional data. Consequently, to uncover meaningful patterns and develop reliable predictive models in such settings, new statistical methodologies that can effectively leverage domain-specific structures and address the unique properties of discrete, high-dimensional spaces are urgently needed.

This urgent need for new statistical methodologies is particularly magnified by the rise of large-scale biological and textual datasets, which present high-dimensional, heterogeneous, and often noisy categorical data. In the domain of molecular biology and biomedical engineering, one critical challenge lies in evaluating and predicting the effects of genetic mutation effects. As the catalog of observed variations continues to grow in both humans and model organisms, identifying amino acid alterations that drive phenotypic differences or contribute to complex diseases becomes increasingly important. Although high-throughput mutational scanning technologies have advanced the experimental study of mutation effects (Melamed et al., 2013; Lek et al., 2016), these approaches are often constrained by cost, scalability, and experimental feasibility. For example, analyzing protein mutations may involve hundreds of sites, making exhaustive experimentation impractical and highlighting the need for statistical models that can leverage sequence data.

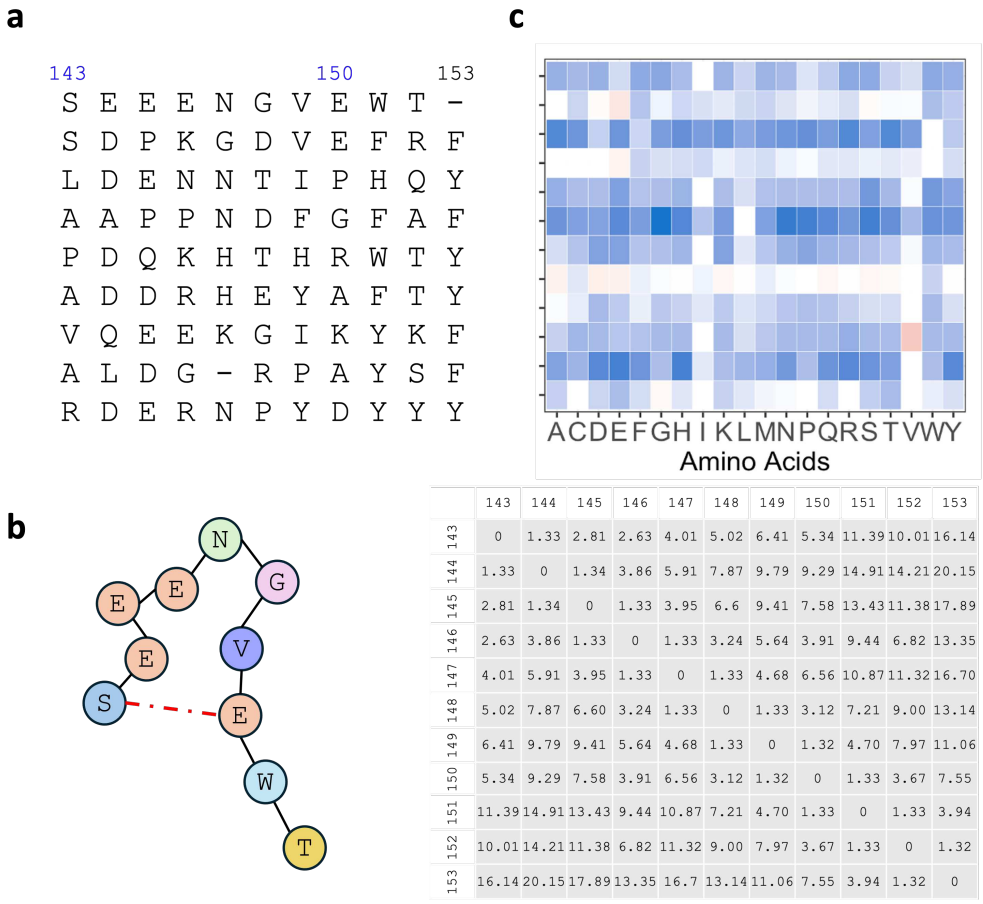
In parallel, the explosion of textual data across a variety of domains has posed significant challenges for extracting structured insights from unstructured language in natural language

processing. Topic modeling has emerged as a foundational unsupervised technique for discovering latent thematic structures in large collections of documents. While originally designed for text analysis, topic modeling methods have proven highly adaptable, finding applications in diverse areas such as bioinformatics (McMurdie and Holmes, 2014), social network analysis (Hsieh et al., 2012), and healthcare analytics (Li et al., 2021). However, traditional topic models often assume relatively long and information-rich documents, which limits their effectiveness in increasingly common settings involving short, sparse texts such as paper abstracts, or social media posts.

This dissertation investigates two distinct applications of high-dimensional categorical data analysis: the modeling and prediction of protein mutation fitness using structural information through a high-dimensional Potts model, and the topic analysis of sparse short texts using a zero-inflated Poisson model combined with a singular value decomposition (SVD) approach for a low-rank structure. Although these applications differ in scientific domain and modeling framework, they share a common statistical objective: developing structured, scalable, and theoretically grounded methodologies for learning from sparse, high-dimensional categorical data. Background for each study appears in Sections 1.1 and 1.2, respectively.

1.1 Modeling of Protein Mutation Fitness with Potts Models

Understanding how mutations affect protein function is a fundamental problem in molecular biology, with implications for disease research, drug development, and evolutionary studies. A key source of information for such analysis comes from evolutionary data, where patterns of conservation and co-variation across protein sequences offer clues about functional constraints. Multiple sequence alignments (MSAs) align sites across evolutionarily related protein sequences within the same family and serve as primary input data for studying mutation effects. For example, Figure 1.1(a) showcases a portion of the MSA data for the Dihydrofolate reductase (DYR) protein family, a key enzyme in folate metabolism, responsible for reducing dihydrofolic



	143	144	145	146	147	148	149	150	151	152	153
143	0	1.33	2.81	2.63	4.01	5.02	6.41	5.34	11.39	10.01	16.14
144	1.33	0	1.34	3.86	5.91	7.87	9.79	9.29	14.91	14.21	20.15
145	2.81	1.34	0	1.33	3.95	6.6	9.41	7.58	13.43	11.38	17.89
146	2.63	3.86	1.33	0	1.33	3.24	5.64	3.91	9.44	6.82	13.35
147	4.01	5.91	3.95	1.33	0	1.33	4.68	6.56	10.87	11.32	16.70
148	5.02	7.87	6.60	3.24	1.33	0	1.33	3.12	7.21	9.00	13.14
149	6.41	9.79	9.41	5.64	4.68	1.33	0	1.32	4.70	7.97	11.06
150	5.34	9.29	7.58	3.91	6.56	3.12	1.32	0	1.33	3.67	7.55
151	11.39	14.91	13.43	9.44	10.87	7.21	4.70	1.33	0	1.33	3.94
152	10.01	14.21	11.38	6.82	11.32	9.00	7.97	3.67	1.33	0	1.32
153	16.14	20.15	17.89	13.35	16.7	13.14	11.06	7.55	3.94	1.32	0

Figure 1.1: (a) MSA data for the DYR family at sites 143-153. (b) Protein structure highlighting the spatial proximity of sites 143 and 150, with site-wise physical distances derived from structural data. (c) Integration of MSA and structural data facilitates mutation fitness landscape analysis.

acid to tetrahydrofolic acid (Schnell et al., 2004). Each row represents a protein sequence, with twenty distinct letters[1] encoding specific amino acids and dashed lines indicating alignment gaps.

MSA data are widely used in computational methods for assessing the functional impact of amino acid mutations. For example, SIFT (Ng and Henikoff, 2003) directly relies on conservation scores derived from MSAs to predict whether a mutation is likely to affect protein function, while PolyPhen (Adzhubei et al., 2010) incorporates these conservation features into probabilistic classifiers such as Naïve Bayes. Although effective in practice, these methods evaluate each site largely in isolation and do not explicitly model dependencies between sites.

[1] A, C, D, E, F, G, H, I, K, L, M, N, P, Q, R, S, T, V, W, and Y represent amino acids

The aforementioned computational approaches using MSA data primarily treat it as a collection of fixed amino acids, without quantifying its uncertainty or integrating it with other informative resources, such as protein structural data shown in Figure 1.1(b). In fact, MSA data for each sequence can be viewed as a multivariate categorical random vector. The Potts model (Potts, 1952), an extension of the Ising model (Ising, 1925), naturally model such data, allowing for the investigation of both single-site and pair-site parameters. In this context, single-site parameters represent site-wise effects, while pair-site parameters, known as direct coupling between sites (Morcos et al., 2011), capture both site-wise and site-amino acid-wise dependencies. This framework facilitates the calculation of evolutionary statistical energy and provides a coherent assessment of relative energy changes caused by single-site or even multiple-site mutations. Consequently, the Potts model offers a flexible statistical framework for predicting mutation effects and constructing mutation fitness landscapes, as illustrated in Figure 1.1(c).

Building on the principle of maximum entropy, which supports constructing statistical models that reproduce observed correlations without making unwarranted assumptions, direct coupling analysis (DCA) applies the Potts model to estimate direct site-site couplings from MSA data. It does so by fitting a sparse Markov random field to the sequence alignment, allowing the model to disentangle direct interactions from indirect correlations mediated through other sites within a protein family (Morcos et al., 2011; Levy et al., 2017). By estimating coupling parameters that match empirical amino acid frequencies and covariances, DCA reveals a network of interactions that are critical for understanding protein structure, function, and stability. Similarly, EVmutation (Hopf et al., 2017) builds on the Potts model to characterize mutation fitness landscapes by capturing interactions across all pairs of sites using MSA data. It extends the DCA framework by assigning a statistical energy to each sequence, allowing the effect of a mutation to be quantified as the change in energy between the wild-type and mutant sequences.

Despite their effectiveness in modeling pairwise interactions, existing Potts-based approaches such as DCA and EVmutation face limitations in fully leveraging structural information and handling the sparsity of dependencies within sequence alignments. These challenges underscore

the need for new statistical frameworks that are interpretable, theoretically grounded, and scalable to high-dimensional protein sequence data. We propose a modeling approach that builds upon and extends these ideas. Our work specifically aims to improve the prediction of mutation fitness, with a particular focus on robust integration of structural information and enhanced model interpretability.

1.2 Topic Modeling of Sparse Texts

Topic modeling is a fundamental tool in text analysis that aims to uncover latent thematic structures within large collections of documents. It provides a principled way to organize, summarize, and explore unstructured textual data by identifying coherent groups of words, known as topics, that tend to appear together. These topics reveal underlying semantic patterns and are useful in a wide range of applications including information retrieval, document classification, recommendation systems, and content summarization.

Mathematically, given a corpus of n observed documents and a fixed vocabulary of V words, we represent it as a document-word matrix $\mathbf{Z} = (Z_{ij}) \in \mathbb{R}^{n \times V}$ where Z_{ij} denotes the frequency of word j in document i. The goal of topic modeling is to recover K underlying topics and model the relationships among documents, topics, and words using a document-topic matrix $\mathbf{W} \in \mathbb{R}^{n \times K}$ and a topic-word matrix $\mathbf{A} \in \mathbb{R}^{K \times V}$. Each document is assumed to be a mixture of topics, and each topic is defined by a distribution over words.

Notably, recent work on topic modeling has increasingly focused on neural embedding-based approaches, particularly those leveraging large pretrained language models such as BERT. Methods like BERTopic (Grootendorst, 2022) apply a pretrained bidirectional encoder to generate dense semantic embeddings for documents, which are then processed using dimensionality reduction techniques (e.g., UMAP) and clustering algorithms. Topic words are extracted heuristically via TF-IDF applied to the resulting clusters. While effective at capturing high-level semantic similarity, these models lack an explicit generative process for word occurrences and provide no theoretical guarantees for topic recovery or estimation stability. Moreover, because they

rely on frozen pretrained encoders, their adaptability to domain-specific vocabulary and context is limited. These limitations have motivated a renewed interest in probabilistic models, which offer transparent generative formulations, interpretable parameter structures, and theoretical justifications for recovery guarantees.

Two prominent probabilistic models in this context are Latent Dirichlet Allocation (LDA) (Blei et al., 2003) and probabilistic Latent Semantic Indexing (pLSI) (Hofmann, 1999), both of which rely on probabilistic frameworks based on multinomial distributions to model word frequencies within individual documents. LDA is a foundational method in topic modeling. It employs a hierarchical Bayesian framework that includes a Poisson prior for document length and a Dirichlet prior for the distribution of topics within documents. This allows each document to be modeled as a mixture over topics, where each topic is represented by a multinomial distribution over words. Parameter estimation in LDA is typically performed using techniques such as Markov Chain Monte Carlo (MCMC). The model has been extended in numerous directions, including dynamic topic models (DTM) (Blei and Lafferty, 2006), which introduce a temporal structure by placing a state-space model over the Dirichlet parameters. This allows the topic proportions and word distributions to evolve over time, capturing topic shifts and trends in longitudinal corpora. Variants of Latent Dirichlet Allocation (LDA) have been adapted for supervised learning, modeling correlations between topics, and incorporating additional labels, demonstrating the model's considerable flexibility across diverse application domains.

The probabilistic Latent Semantic Indexing (pLSI) model (Hofmann, 1999) has served as a foundational approach in topic modeling by representing each document as a mixture over a set of latent topics, with each topic characterized by a probability distribution over the vocabulary. In the modern matrix factorization formulation of pLSI (Arora et al., 2012; Bing et al., 2020a,b; Wu et al., 2023; Ke and Wang, 2024; Tran et al., 2023), the parameter matrix of document-specific multinomial word distributions is denoted as $\mathbf{\Pi} \in \mathbb{R}^{n \times V}$. Each row $\mathbf{\Pi}_{i\cdot}$ of $\mathbf{\Pi}$ defines the multinomial parameters for document i, encoding the probabilities of observing each word

type in the vocabulary. The key modeling assumption is that $\mathbf{\Pi}$ admits a low-rank factorization of the form $\mathbf{\Pi} = \mathbf{WA}$ with the number of latent topics $K \ll \min(n, V)$.

To ensure that $\mathbf{W}$ and $\mathbf{A}$ admit a probabilistic interpretation, additional constraints are typically imposed. Specifically, the rows of both matrices are required to lie on the probability simplex. These constraints place the factorization within the framework of non-negative matrix factorization (NMF) with simplex structure, which has been widely studied in both theory and application. However, the non-uniqueness of matrix factorizations poses a significant challenge in recovering the true underlying topic-word matrix $\mathbf{A}$. Without additional structural assumptions, $\mathbf{W}$ and $\mathbf{A}$ are only identifiable up to a rotation within the space of rank-K factorizations, making the interpretation of recovered topics ambiguous. To address this, Arora et al. (2012) introduced the anchor words assumption, which provides a sufficient condition for the identifiability of $\mathbf{A}$. The anchor words condition assumes that each topic contains at least one word that has non-zero probability only within that topic and zero probability under all others. This assumption geometrically implies that the rows of $\mathbf{A}$ lie within a convex hull whose vertices correspond to anchor words, enabling exact recovery of the topic-word matrix under separability, which builds upon the separability condition introduced by Donoho and Stodden (2003) in the context of non-negative matrix factorization. Under the separability condition, numerous algorithms have been developed to leverage this structural property. These include geometric approaches such as the Successive Projection Algorithm (SPA), convex optimization methods, and spectral techniques that combine dimensionality reduction with vertex search procedures. Recent advances, exemplified by Bing et al. (2020b); Wu et al. (2023), investigate the sparsity of the topic–word matrix $\mathbf{A}$ and offer statistically near-optimal guarantees. The subsequent study (Tran et al., 2023), however, suggests that the underlying sparsity assumption may be overly restrictive and susceptible to overfitting. Furthermore, Bing et al. (2020b); Ke and Wang (2024) introduce scalable algorithms for anchor word recovery, which significantly enhance the practical utility of the pLSI framework for analyzing real-world text corpora.

While pLSI and LDA are effective for many text corpora, their underlying multinomial assumption presents limitations in practical applications. In particular, the conventional treatment of document length as a fixed parameter within the multinomial framework constrains the model's capacity to capture the relationship between document size, topic complexity and vocabulary richness. Empirical evidence, as demonstrated in Figure 1.2(b), reveals a positive correlation between document length and vocabulary coverage, suggesting a systematic relationship that merits explicit modeling. However, this relationship is overlooked under the multinomial assumption, which treats document length as exogenous. Although one could model document length as a function of vocabulary size, the relationship is not trivial to specify.

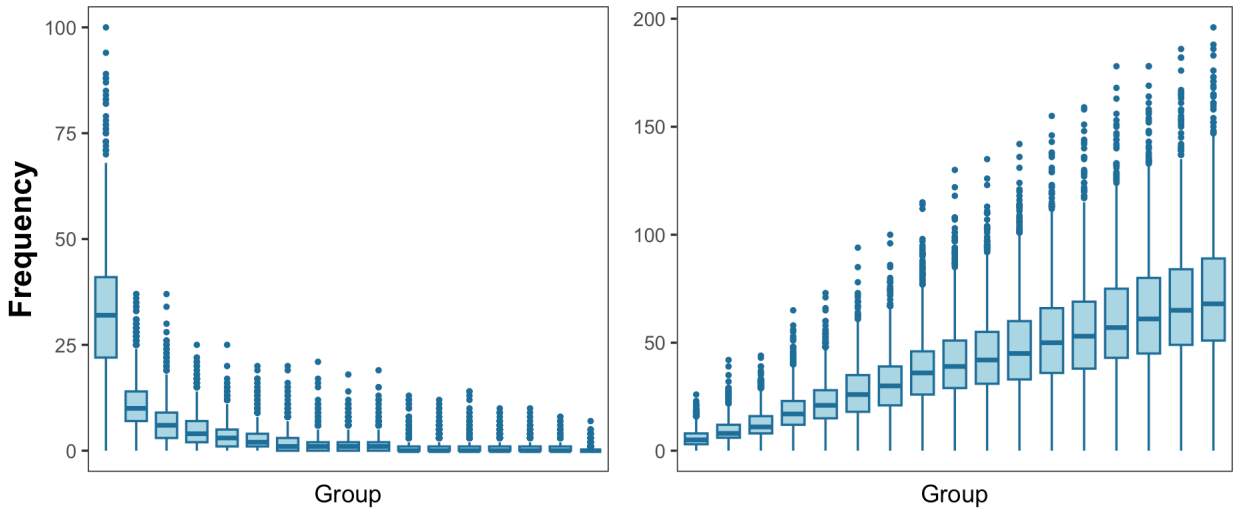


Figure 1.2: Analysis of a corpus of scientific abstracts, where each document is represented as a bag-of-words vector. Left: (a) All words in the vocabulary are sorted by their TF-IDF scores in descending order and divided into consecutive 100-word groups. For each group, the total frequency of the included words is computed per document, and the distribution across documents is visualized using a boxplot. Right: (b) All words in the vocabulary are randomly shuffled and divided into consecutive 100-word groups. Let G_k denote the union of the first k groups in this shuffled list. For each document and each k, the total frequency of the words in G_k that appear in the document is computed. The distribution of these cumulative frequencies across documents is visualized using a boxplot.

Furthermore, Figure 1.2(a) illustrates a highly skewed word frequency distribution, where a small set of words accounts for a large portion of total usage, while most words are relatively rare. In contrast, the distribution observed in synthetic corpora generated by LDA with Dirichlet priors

is notably less heavy-tailed and less skewed. This discrepancy suggests that LDA fails to capture the degree of imbalance in word frequencies characteristic of natural language.

In addition to length-vocabulary dependency, sparsity is another key challenge. Many real-world corpora, particularly short texts or domain-specific documents, contain words that appear rarely or not at all across most documents. To explore this issue, we examine ten commonly used datasets: the Multi-Attribute Dataset on Statisticians (MADStat) introduced by Ke et al. (2024); the Associated Press (AP) dataset from Ke and Wang (2024); the COVID-19 Open Research Dataset (CORD-19) from Wang et al. (2020); research articles from Kaggle; the DBLP Discovery Dataset (D3) from Wahle et al. (2022); and the Enron Emails, NIPS Papers, KOS Blogs, NYTimes, and PubMed Abstracts provided by Newman (2008). Table 1.1 reports the sparsity of each dataset, computed as the proportion of zero entries in the corpus matrix relative to the total number of entries which is nV. All datasets exhibit sparsity levels exceeding 95%, indicating that sparsity is a substantial factor that cannot be ignored in downstream analysis. We replicate this level of sparsity in our numerical experiments.

Table 1.1: Sparsity levels of ten commonly used text datasets.

Data	n	V	Sparsity	Data	n	V	Sparsity
MADStat	14000	2102	98.16%	AP	10473	2246	98.72%
CORD-19	10694	8129	98.33 %	Enron Emails	39861	28102	99.67%
NIPS Paper	1500	12419	95.99%	KOS Blog	3430	6906	98.51%
NYTimes	300000	102660	99.77%	PubMed Abstracts	8200000	141043	99.95%
Research Abstracts	29961	10572	98.93%	D3	6300000	16000	99.19%

While some studies have explored adaptations of pLSI to sparse settings (Bing et al., 2020b; Wu et al., 2023; Tran et al., 2023), the multinomial framework inherently lacks mechanisms to handle structured sparsity or distinguish between *structural zeros* (words that cannot appear in certain contexts) and *sampling zeros* (words that could appear but happen not to). Although zero-inflated extensions of LDA (Deek and Li, 2021) have been proposed, they typically rely on MCMC for estimation, which incurs substantial computational overhead with limited theoretical guarantees.

These observations reveal a fundamental mismatch between the traditional multinomial assumption and the statistical realities of real-world text corpora, especially given their extreme sparsity and structured zeros. To overcome these limitations, we introduce a novel probabilistic framework centered on a zero-inflated Poisson (ZIP) model. This model is designed to explicitly account for excess zeros and accommodate low-rank structure to effectively determine the topic-word matrix. Our approach ensures reliable estimation, preserves interpretability, and offers theoretical guarantees for consistency.

1.3 Outline

The subsequent chapters of this dissertation are organized as follows, each addressing a distinct modeling challenge involving high-dimensional categorical data. Chapter 2 introduces a high-dimensional Potts model for predicting mutation fitness from multiple sequence alignments. It presents the model formulation, describes the estimation procedure, establishes theoretical properties, and includes empirical evaluations based on real biological data. Chapter 3 develops a zero-inflated Poisson (ZIP) topic modeling framework with a low-rank structure designed to address the challenges posed by short and sparse text corpora. It details the statistical formulation, outlines an efficient estimation algorithm, and demonstrates the method's advantages through simulation studies and applications to real-world datasets. Finally, Chapter 4 concludes the dissertation by summarizing the main contributions and outlining directions for future research in the statistical modeling of high-dimensional categorical data.

Chapter 2

Modeling and prediction of mutation fitness on protein functionality with structural information using high-dimensional Potts model

2.1 Introduction

Understanding the effects of amino acid mutations on protein function is a central problem in molecular biology, with broad implications for disease mechanism studies, protein engineering, and evolutionary analysis. As discussed in Chapter 1, statistical models that aim to predict the functional impact of mutations from MSA data must capture not only single-site conservation patterns but also site-amino acid-wise dependencies. The Potts model has emerged as a powerful framework for this task, as it enables the joint modeling of site-specific effects and pairwise site-amino acid-wise interactions across sites.

2.1.1 Related works on estimating high-dimensional Potts model

Bayesian approaches using Metropolis-Hastings (MH) algorithms are commonly employed for parameter estimation in the Potts model (Møller et al., 2006; Li et al., 2017). Due to the computational intractability of the partition function, calculating the MH ratio requires introducing auxiliary variables. This approach leverages the conditional distributions based on the data and additional parameters to eliminate the intractable term. However, several challenges arise, including selecting priors for Potts model parameters, managing burn-in processes, and determining the conditional distributions of auxiliary variables. Sampling these auxiliary variables could be fairly challenging, particularly for high-dimensional cases, requiring either unbiased sampling or accepting Markov chains whose stationary distribution may deviate from the desired

posterior (Park and Haran, 2018). These complexities underscore the need for careful consideration when drawing inferences on the results.

An alternative strategy for parameter estimation in the Potts model is the pseudolikelihood (Hopf et al., 2017), which relies on the product of conditional probabilities of one variable given the others (Balakrishnan et al., 2011). While these methods avoid involving prior knowledge of parameters or auxiliary variables, they are computationally expensive for simultaneous estimation of all parameters. To address this, the Potts-Ising model (Razaee and Amini, 2020) reduces the dimensionality by constraining pair-site parameters to match the Ising model form. However, this model is tailored for categorical data with a distinctive category, where dependencies depend only on the presence or absence of that category.

As a special case of the Potts model, the Ising model has received significant attention in the literature. Node-wise ℓ_1-regularized logistic regression (Ravikumar et al., 2010) is among the most popular approaches for parameter estimation in the Ising model, with practical algorithms and well-studied theoretical guarantees. It is natural to extend this approach and apply node-wise multinomial regression for parameter estimation in the Potts model, which allows the incorporation of specific penalties. For example, recent work in Tian et al. (2024) investigates the estimation and prediction errors for ℓ_1-penalized multinomial regression, while other studies explore multinomial regression with structured penalties (Tutz and Gertheiss, 2016; Nibbering and Hastie, 2022; Levy and Abramovich, 2023), assuming element-wise or group-wise sparsity. Although Tutz and Gertheiss (2016) mentions the idea of combining penalties for recapitulating data-specific structures, simultaneous element-wise and group-wise sparsity remains largely underexplored in multinomial regression.

2.1.2 Our contributions

We propose a high-dimensional Potts model-based statistical framework to analyze MSA data for studying protein mutation effects. We develop parameter estimation for the Potts model using node-wise multinomial regression, which is computationally efficient. Biologically, site-

wise dependencies are not expected to be universal to maintain evolutionary flexibility (Jones et al., 2012; Jernigan et al., 2021), considered as *site-wise sparsity*, and strong dependencies between specific amino acids are not densely observed in general, considered as *element-wise sparsity*. Thus, we enforce group-wise and element-wise sparsity using the sparse group Lasso penalty (Simon et al., 2013), resulting in a node-wise multinomial regression with a double sparse structure at both group and element levels. Additionally, since site-wise dependencies tend to be stronger at spatially closer sites (Morcos et al., 2011; Marks et al., 2012), we integrate protein structural data with MSA data by deriving group weights in the sparse group Lasso penalty based on spatial distance between sites. Our framework integrates the Potts model, sparse group Lasso, and combined MSA and structural data, providing an efficient statistical approach to predict protein mutation fitness. We apply our method to 12 protein families to estimate both single-site and pair-site parameters, aiding in the estimation of energy changes due to mutations. As demonstrated in Section 2.6, our method achieves much higher correlation with experimentally measured mutation fitness compared to the benchmark EVM (Hopf et al., 2017).

Beyond methodological development, we rigorously analyze the theoretical properties of the proposed method. Estimators with sparse group penalties pose challenges due to their non-decomposability (Cai et al., 2022). Existing approaches primarily rely on analysis of the Karush–Kuhn–Tucker condition and require additional assumptions, such as incoherence-type condition (Cai et al., 2022) or strong sparsity conditions (Zhang and Li, 2023), to establish estimation error bounds. We take a different strategy that avoids these assumptions by deriving a tight upper bound for the stochastic term $\sum_{k=1}^{K} \epsilon_k^{\mathrm{T}} \mathbf{X} u_k$, induced by the proposed loss function Equation (2.8). Here, $\mathbf{X} \in \mathbb{R}^{n \times D}$ is the design matrix, $\epsilon_k \in \mathbb{R}^n$ are errors, and u_k is a D-dimensional vector. Traditional treatments for $\sum_{k=1}^{K} \epsilon_k^{\mathrm{T}} \mathbf{X} u_k$, such as $\epsilon_k^{\mathrm{T}} \mathbf{X} u_k \leq \|\mathbf{X}^{\mathrm{T}} \epsilon_k\|_\infty \|u_k\|_1$, may result in sub-optimal convergence rates for estimation (Lounici et al., 2011), especially for non-decomposable penalties like sparse group penalties. By modifying the new oracle inequalities for high-dimensional linear models (Bellec et al., 2018), we derive a finer upper bound for $\sum_{k=1}^{K} \epsilon_k^{\mathrm{T}} \mathbf{X} u_k$, as detailed after Theorem 2.3.1. This refined bound allows us to establish ℓ_q error

bounds ($q = 1, 2$) for the high-dimensional Potts model using sparse group Lasso, matching the minimax lower bound for high-dimensional linear models with sparse group structures, up to a factor dependent on the multinomial nature of the Potts model. This factor highlights fundamental distinctions between multinomial regression and linear or logistic regression, which remain underexplored in existing literature. Furthermore, these error bounds provide consistency guarantees for the estimated energy changes, forming the theoretical foundation for Potts model-based mutation analysis in the literature.

The remainder of this chapter is organized as follows. Section 2.2 provides background on the Potts model and defines evolutionary statistical energy. We then introduce our estimation procedure for the high-dimensional Potts model, employing sparse group Lasso in node-wise multinomial regression. In Section 2.3, we establish the ℓ_1 and ℓ_2 convergence rates of the proposed estimators and demonstrate guarantees of the estimated energy changes. Section 2.4 discusses the integration of structural information into the Potts model for mutation analysis. Sections 2.6 and 2.5 evaluate our method through comprehensive real data mutation analyses and simulation studies, both demonstrating the superiority of the proposed method.

2.2 Methodology

2.2.1 Potts model and evolutionary energy for protein mutations

Consider an MSA data with n protein sequences, each containing d sites. The amino acids at each site are encoded by $K+1$ possible states, starting from 0, where K generally equals to 20 and represents the 20 amino acid types, with 0 denoting the alignment gap. Denote $[r] = \{1, 2, ..., r\}$ and $[r]_0 = \{0, 1, 2, ..., r\}$ for any positive integer r. For each $j \in [d]$, the MSA data at site j can be represented as $\mathbf{z}_j := (z_{j0}, ..., z_{jK})^{\mathrm{T}} \in \{0, 1\}^{K+1}$, where the binary value z_{jk} indicates whether amino acid k is present at site j. It follows that $\sum_{k\in[K]_0} z_{jk} = 1$. An individual protein sequence can then be expressed as $\mathbf{z} := (\mathbf{z}_1^{\mathrm{T}}, \ldots, \mathbf{z}_d^{\mathrm{T}})^{\mathrm{T}} \in \{0, 1\}^{d(K+1)}$. With these notations, the Potts model

for a sequence $\mathbf{z}$ with d sites and $K+1$ states is given by

$$P(\mathbf{z}) = \frac{1}{C} \exp\left\{ \sum_{j=1}^{d} \sum_{k=0}^{K} \theta_{jk} z_{jk} + \sum_{1 \le j < r \le d} \sum_{k=0}^{K} \sum_{l=0}^{K} \gamma_{jr,kl} z_{jk} z_{rl} \right\}, \tag{2.1}$$

where C is the partition function, the parameter θ_{jk} represents the single-site effect on energy at site j for amino acid k, while $\gamma_{jr,kl}$ quantifies the direct coupling between site j with amino acid k and site r with amino acid l.

Following the definition in Levy et al. (2017), the evolutionary statistical energy (short for *energy* hereafter) of a protein sequence $\mathbf{z}$ is defined as

$$E(\mathbf{z}) = \sum_{j=1}^{d} \sum_{k=0}^{K} \theta_{jk} z_{jk} + \sum_{1 \le j < r \le d} \sum_{k=0}^{K} \sum_{l=0}^{K} \gamma_{jr,kl} z_{jk} z_{rl}. \tag{2.2}$$

The overall energy of the sequence is thus the sum of single-site effects and direct couplings between sites, subject to the constraint that only one amino acid type can be assigned to each site. We can represent the energy change of a mutation at single or multiple sites with these parameters. Specifically, consider a mutation at site j that changes the amino acid from the wild-type a_j to another amino acid k, while leaving all other sites unchanged. The energy change of this single-site mutation can be calculated as

$$\Delta E_{j,k} = \theta_{jk} - \theta_{ja_j} + \sum_{r \ne j} (\gamma_{jr,ka_r} - \gamma_{jr,a_j a_r}). \tag{2.3}$$

This can be naturally extended of a multiple-site mutation. Let $\mathcal{J}$ denote the set of sites where mutations occur, and $\mathbf{k}_{\mathcal{J}} = (k_j : j \in \mathcal{J})$ denote the amino acids at the mutated sites transitioning from the wild-types. Assuming the cardinality $|\mathcal{J}| > 0$, the energy change of such a multiple-site mutation is

$$\Delta E_{\mathcal{J},\mathbf{k}_{\mathcal{J}}} = \sum_{j \in \mathcal{J}} (\theta_{jk_j} - \theta_{ja_j}) + \sum_{j \in \mathcal{J}} \sum_{r \notin \mathcal{J}} (\gamma_{jr,k_j a_r} - \gamma_{jr,a_j a_r}) + \sum_{j,j' \in \mathcal{J}, j \ne j'} (\gamma_{jj',k_j k_{j'}} - \gamma_{jj',a_j a_{j'}}). \tag{2.4}$$

The energy change corresponds to the log-likelihood ratio between the mutant and wild type, with higher values indicating favorable mutations and lower values for unfavorable ones.

Previous studies (Hopf et al., 2017) have demonstrated a notable agreement between energy changes and experimental results on mutation fitness. Thus, to analyze mutation fitness using MSA data modeled by the Potts model, it is sufficient to estimate the model parameters. For parameter identifiability, we assign $k = 0$ to the wild-type amino acid and set the corresponding parameters θ_{j0}, $\gamma_{jr,0l}$, and $\gamma_{jr,k0}$ to zero. This identifiability adjustment does not affect the definition of $\Delta E_{j,k}$ in Equation (2.3).

2.2.2 Node-wise sparse multinomial regression

As mentioned in Section 2.1.1, directly estimating all parameters in the Potts model Equation (2.1) can be computationally challenging. Instead, it is more practical to estimate the parameters at each site individually. With a slight abuse of notation, we exclude the wild-type amino acids from $\mathbf{z}$. For each $j \in [d]$, let $\mathbf{z}_j = (z_{j1}, ..., z_{jK})^{\mathrm{T}} \in \{0,1\}^K$ represent the amino acid type at site j, and let $\mathbf{z}_{-j} = (\mathbf{z}_1^{\mathrm{T}}, \ldots, \mathbf{z}_{j-1}^{\mathrm{T}}, \mathbf{z}_{j+1}^{\mathrm{T}}, \ldots, \mathbf{z}_d^{\mathrm{T}})^{\mathrm{T}} \in \{0,1\}^{(d-1)K}$ represent the states at all other sites. Given this setup, when focusing on a single site j while conditioning on the other sites, the conditional probability in the Potts model follows an exponential family distribution that

$$P(\mathbf{z}_j|\mathbf{z}_{-j}) = \frac{\prod_{k=1}^{K} \exp(\theta_{jk} + \sum_{r\neq j}\sum_{l=1}^{K} \gamma_{jr,kl} z_{rl})^{z_{jk}}}{1 + \sum_{k=1}^{K} \exp(\theta_{jk} + \sum_{r\neq j}\sum_{l=1}^{K} \gamma_{jr,kl} z_{rl})}, \tag{2.5}$$

Thus, a node-wise multinomial regression can be applied by considering $\mathbf{z}_j$ for each $j \in [d]$ as the response, representing different amino acid states, and $\mathbf{z}_{-j}$, the amino acids at all other sites, as covariates. A similar approach is adopted by Cai et al. (2019) to estimate parameters in an Ising model using node-wise logistic regression.

For each site $j \in [d]$, we first derive the loss function for our node-wise multinomial regression. Since the MSA data at site j is treated as the response, let $\mathbf{Y}_j = (\mathbf{Y}_j^{(1)}, \mathbf{Y}_j^{(2)}, \ldots, \mathbf{Y}_j^{(n)})^{\mathrm{T}}$ represent the response matrix of n observed sequences, where $\mathbf{Y}_j^{(i)} = (y_{j1}^{(i)}, \ldots, y_{jK}^{(i)})^{\mathrm{T}} \in \{0,1\}^K$ is the ith sequence of $\mathbf{z}_j$, and $y_{jk}^{(i)} = 1$ indicates that the amino acid at site j in the ith sequence is in

state k. Meanwhile, since the MSA data at all other sites are considered as covariates, let $\mathbf{X}_{-j} \in \{0,1\}^{n\times(d-1)K}$ represent the design matrix, where the ith row is denoted as $\mathbf{X}_{-j,i} = \mathbf{x}_{-j}^{(i)}$. Here, $\mathbf{x}_{-j}^{(i)}$ represents the ith observation of $\mathbf{z}_{-j}$, specifically $\mathbf{x}_{-j}^{(i)} = (\mathbf{x}_1^{(i)\mathrm{T}}, \ldots, \mathbf{x}_{j-1}^{(i)\mathrm{T}}, \mathbf{x}_{j+1}^{(i)\mathrm{T}}, \ldots, \mathbf{x}_d^{(i)\mathrm{T}})^{\mathrm{T}} \in \{0,1\}^{(d-1)K}$, where $\mathbf{x}_r^{(i)} = (\mathbf{x}_{r1}^{(i)}, \ldots, \mathbf{x}_{rK}^{(i)})^{\mathrm{T}} \in \{0,1\}^K$ for $r \neq j$. Here, $\mathbf{x}_{rk}^{(i)} = 1$ indicates the amino acid at site r in the ith sequence is in state k. Given $\mathbf{X}_{-j}$ and $\mathbf{Y}_j$, the negative log-likelihood function is

$$\ell(\boldsymbol{\theta}_j, \boldsymbol{\gamma}_j; \mathbf{Y}_j, \mathbf{X}_{-j}) = \sum_{i=1}^{n} \left[\log\left(1 + \sum_{l=1}^{K} \exp\{\theta_{jl} + \boldsymbol{\gamma}_{j\bullet,k\bullet}^{\mathrm{T}} \mathbf{x}_{-j}^{(i)}\}\right) - \sum_{k=1}^{K} y_{jk}^{(i)} (\theta_{jk} + \boldsymbol{\gamma}_{j\bullet,k\bullet}^{\mathrm{T}} \mathbf{x}_{-j}^{(i)}) \right], \tag{2.6}$$

where $\boldsymbol{\theta}_j = (\theta_{j1}, ..., \theta_{jK})^{\mathrm{T}} \in \mathbb{R}^K$ consists of the single-site effects for the K amino acids at site j, and $\boldsymbol{\gamma}_j = (\boldsymbol{\gamma}_{j\bullet,1\bullet}^{\mathrm{T}}, ..., \boldsymbol{\gamma}_{j\bullet,K\bullet}^{\mathrm{T}}) \in \mathbb{R}^{(d-1)K^2}$ consists of all possible dependencies between the K amino acids at site j and the K amino acids at each of the other $d-1$ sites. For each $k \in [K]$, $\boldsymbol{\gamma}_{j\bullet,k\bullet} = (\boldsymbol{\gamma}_{j1,k\bullet}^{\mathrm{T}}, \ldots, \boldsymbol{\gamma}_{j(j-1),k\bullet}^{\mathrm{T}}, \boldsymbol{\gamma}_{j(j+1),k\bullet}^{\mathrm{T}}, \ldots, \boldsymbol{\gamma}_{jd,k\bullet}^{\mathrm{T}})^{\mathrm{T}} \in \mathbb{R}^{(d-1)K}$ represents the dependencies between site j with amino acid k and the K amino acids at each of the other $d-1$ sites. In addition, $\boldsymbol{\gamma}_{jr,k\bullet} = (\gamma_{jr,k1}, \ldots, \gamma_{jr,kK})^{\mathrm{T}} \in \mathbb{R}^K$ for $r \neq j$ collects the dependencies between site j with amino acid k and the K amino acids at site r.

Given a fixed site j and for each site pair (j,r) with $r \neq j$, the direct coupling

$$\boldsymbol{\gamma}_{j(r)} = (\boldsymbol{\gamma}_{jr,1\bullet}^{\mathrm{T}}, \ldots, \boldsymbol{\gamma}_{jr,K\bullet}^{\mathrm{T}})^{\mathrm{T}} \in \mathbb{R}^{K^2} \tag{2.7}$$

between the K amino acids at site j and the K amino acids at site r can be viewed as a subgroup within $\boldsymbol{\gamma}_j$. Particularly, sites j and r are independent conditional on other sites if and only if $\boldsymbol{\gamma}_{j(r)} = \mathbf{0}$. As discussed in Section 2.1, dependencies are not expected for all site pairs, indicating *site-wise sparsity*. Also, strong dependencies between specific amino acids, primarily driven by biochemical properties like polarity and charge, are neither expected to be universal, suggesting *element-wise sparsity* within amino acid groups. To account for these, we enforce both group-wise and element-wise sparsity in our procedure. This leads to the penalty $h(\boldsymbol{\gamma}_j) = \lambda_g \sum_{r\neq j} \|\boldsymbol{\gamma}_{j(r)}\|_2 + \lambda \sum_{r\neq j} \|\boldsymbol{\gamma}_{j(r)}\|_1$, where $\lambda_g, \lambda > 0$ are the tuning parameters for the

group Lasso and Lasso penalties, respectively, and $\|\cdot\|_q$ denotes the ℓ_q-norm of a vector for a positive integer q. In this context, the group Lasso penalty (Yuan and Lin, 2006) selects sites with strong dependencies with the site j, while the Lasso penalty (Tibshirani, 1996) identifies significant dependencies at the amino acid level. Combining these penalties results in the sparse group Lasso penalty (Simon et al., 2013), which enables a double sparse structure at both the group and element levels. Thus, for each site $j \in [d]$, we propose the following risk function with sparse group Lasso as the objective:

$$L(\boldsymbol{\theta}_j, \boldsymbol{\gamma}_j) = \ell(\boldsymbol{\theta}_j, \boldsymbol{\gamma}_j; \mathbf{Y}_j, \mathbf{X}_{-j}) + h(\boldsymbol{\gamma}_j), \tag{2.8}$$

where $\ell(\boldsymbol{\gamma}_j, \boldsymbol{\theta}_j; \mathbf{Y}_j, \mathbf{X}_{-j})$ is specified in Equation (2.6). With this objective function, our estimator for $(\boldsymbol{\theta}_j, \boldsymbol{\gamma}_j)$ of site j is

$$(\widehat{\boldsymbol{\theta}}_j, \widehat{\boldsymbol{\gamma}}_j) = \underset{\boldsymbol{\theta}_j, \boldsymbol{\gamma}_j}{\operatorname{argmin}} L(\boldsymbol{\theta}_j, \boldsymbol{\gamma}_j). \tag{2.9}$$

Remark 2.2.1. *To further incorporate additional structural information, group-specific weights w_{jr} can be introduced into the group Lasso penalty, resulting in the following modified penalty function*

$$h(\boldsymbol{\gamma}_j) = \lambda_g \sum_{r \neq j} w_{jr} \|\boldsymbol{\gamma}_{j(r)}\|_2 + \lambda \sum_{r \neq j} \|\boldsymbol{\gamma}_{j(r)}\|_1. \tag{2.10}$$

These weights can be selected to reflect the expectation that sites in closer proximity are likely to exhibit stronger dependencies, allowing the model to account for underlying spatial structures more effectively. A further discussion is detailed in Section 2.4.

2.2.3 Parameter estimations

For each $j \in [d]$, although the optimization in Equation (2.9) does not have a closed-form solution, it can be solved with a coordinate gradient descent algorithm (Vincent and Hansen, 2014). The algorithm consists of a sequence of nested loops: the outer one relies on a quadratic approximation of Equation (2.6) at the current iteration, while the middle and inner loops focus on solving the within-group subproblems as described in Simon et al. (2013). Notably, the estimates

$(\widehat{\boldsymbol{\theta}}_j, \widehat{\boldsymbol{\gamma}}_j)$ obtained from the node-wise regressions do not inherently guarantee the symmetry of $\gamma_{jr,kl}$ and $\gamma_{rj,lk}$. To enforce symmetry, we post-process the estimates by averaging the two values.

Let the gradient of $\ell(\boldsymbol{\theta}_j, \boldsymbol{\gamma}_j; \mathbf{Y}_j, \mathbf{X}_{-j})$ at $\boldsymbol{\gamma}_j$ be denoted as $\nabla\ell(\boldsymbol{\gamma}_j)$ and its Hessian matrix as $\nabla^2\ell(\boldsymbol{\gamma}_j)$. Here, $\nabla\ell(\boldsymbol{\gamma}_j) \in \mathbb{R}^{(d-1)K^2}$ and $\nabla^2\ell(\boldsymbol{\gamma}_j) \in \mathbb{R}^{(d-1)K^2\times(d-1)K^2}$. We use $\nabla\ell(\boldsymbol{\gamma}_j)_{(r)}$ and $[\nabla^2\ell(\boldsymbol{\gamma}_j)\nabla\ell(\boldsymbol{\gamma}_j)]_{(r)} \in \mathbb{R}^{K^2}$, following the same group order defined in Equation (2.7). Define the coordinate-wise soft thresholding operator $S(\mathbf{a}, b)$ for a vector $\mathbf{a} = (a_1, ..., a_n)^{\mathrm{T}}$ and a constant b as $S(\mathbf{a}, b) = (\text{sign}(a_1)\max\{|a_1| - b, 0\}, ..., \text{sign}(a_n)\max\{|a_n| - b, 0\})^{\mathrm{T}}$. The estimation procedure for fixed λ and λ_g is outlined in Algorithm 1, while (λ, λ_g) can be tuned via 5-fold cross-validation, as described in Section 2.5.1. In Section 2.2.2, we take the MSA data at each site as the response and the data from the remaining sites as covariates. This allows us to obtain $(\widehat{\boldsymbol{\theta}}_j, \widehat{\boldsymbol{\gamma}}_j)$ for all $j \in [d]$ less costly by leveraging parallel computation.

Algorithm 1: Parameter estimations for the Potts model Equation (2.1)

Input: MSA data, initialization of $\widehat{\boldsymbol{\theta}}_j$, $\widehat{\boldsymbol{\gamma}}_j$ for $j \in [d]$, and penalty parameters $\{\lambda, \lambda_g\}$
For each $j \in [d]$, extract $\{\mathbf{X}_{-j}^{(i)}, \mathbf{Y}_j^{(i)}\}_{i=1}^n$, **repeat**
  For each $r \in [d]$, $r \neq j$, **repeat**
    Update $\widehat{\boldsymbol{\theta}}_j^{\text{new}}$ by mean centering as suggested by Friedman et al. (2010);
    Define the block gradient

$$\nabla g_{jr}(\boldsymbol{\gamma}_j) = \nabla\ell(\widehat{\boldsymbol{\gamma}}_j)_{(r)} + [\nabla^2\ell(\widehat{\boldsymbol{\gamma}}_j)(\boldsymbol{\gamma}_j - \widehat{\boldsymbol{\gamma}}_j)]_{(r)} \in \mathbb{R}^{K^2}.$$

    if $\|S(\nabla g_{jr}(\mathbf{0}), \lambda)\|_2 \leq \lambda_g$, **then** $\widehat{\boldsymbol{\gamma}}_{jr}^{\text{new}} = \mathbf{0}$.
    else Minimize the objective function over each component in $\widehat{\boldsymbol{\gamma}}_{jr}^{\text{new}}$ (Simon et al., 2013).
  until *convergence*;
  $\widehat{\boldsymbol{\gamma}}_j = \widehat{\boldsymbol{\gamma}}_j + t(\widehat{\boldsymbol{\gamma}}_j - \widehat{\boldsymbol{\gamma}}_j^{\text{new}})$ and $\widehat{\boldsymbol{\theta}}_j = \widehat{\boldsymbol{\theta}}_j + t(\widehat{\boldsymbol{\theta}}_j - \widehat{\boldsymbol{\theta}}_j^{\text{new}})$ with step size t from line search.
until *convergence*;
for $j \in [d], r > j$ *and* $k, l \in [K]$ **do**
  Post-process estimates for symmetry: $\tilde{\gamma}_{jr,kl} = \tilde{\gamma}_{rj,lk} = \frac{1}{2}(\widehat{\gamma}_{jr,kl} + \widehat{\gamma}_{rj,lk})$.
end
Output: $\widehat{\boldsymbol{\theta}}_j$, $\tilde{\boldsymbol{\gamma}}_j$

2.3 Theoretical Guarantee

In this section, we derive non-asymptotic ℓ_1 and ℓ_2 error bounds for the proposed sparse group Lasso estimator Equation (2.9). Building on these results, we also establish theoretical guarantees for the plug-in estimators of energy changes.

2.3.1 Preliminaries and assumptions

As described in Section 2.2.2, for each site $j \in [d]$, the dependency parameter vector $\boldsymbol{\gamma}_j \in \mathbb{R}^D$, where $D = (d-1)K^2$, has a group structure such that elements in $\boldsymbol{\gamma}_j$ can be divided into $d-1$ groups $\{\boldsymbol{\gamma}_{j(r)} : r \neq j\}$, where $\boldsymbol{\gamma}_{j(r)}$ corresponds to the group (site) r as defined in Equation (2.7). Moreover, let γ_{jl} for $l \in [D]$ represent the lth element of $\boldsymbol{\gamma}_j$. Given two positive integers s and s_g satisfying $s_g \leq d-1$ and $s_g \leq s \leq s_g K^2$, we say $\boldsymbol{\gamma}_j$ is (s, s_g)-sparse if $\|\boldsymbol{\gamma}_j\|_{0,2} := \sum_{r \neq j} \mathbb{1}\{\boldsymbol{\gamma}_{j(r)} \neq 0\} \leq s_g$ and $\|\boldsymbol{\gamma}_j\|_0 := \sum_{r \neq j} \sum_{k,l \in [K]} \mathbb{1}\{\gamma_{jr,kl} \neq 0\} \leq s$, where $\|\cdot\|_0$ represents the vector ℓ_0-norm. Let $\boldsymbol{\theta}_j^*$ and $\boldsymbol{\gamma}_j^*$ represent the true parameter vectors in the Potts model Equation (2.1) for the site j, and define $\boldsymbol{\gamma}_j^{\circ *} = (\boldsymbol{\theta}_j^{*\mathrm{T}}, \boldsymbol{\gamma}_j^{*\mathrm{T}})^{\mathrm{T}}$. We assume for each $j \in [d]$ that the true coefficient vector $\boldsymbol{\gamma}_j^*$ is (s, s_g)-sparse. That is, for the Potts model, we have

$$\max_{j \in [d]} \|\boldsymbol{\gamma}_j\|_{0,2} \leq s_g \quad \text{and} \quad \max_{j \in [d]} \|\boldsymbol{\gamma}_j\|_0 \leq s.$$

Since entries in $\boldsymbol{\theta}_j^*$ are often assumed to be non-zero for all $j \in [d]$ to model the MSA data, then $\boldsymbol{\gamma}_j^{\circ *}$ is (s°, s_g°)-sparse with $s^\circ = s + K$ and $s_g^\circ = s_g + 1$ by treating $\boldsymbol{\theta}_j^*$ as an additional group.

Recall that, for each $j \in [d]$ and $i \in [n]$, the ith row in the design matrix $\mathbf{X}_{-j}$ is $\mathbf{x}_{-j}^{(i)} \in \{0,1\}^{(d-1)K}$. To investigate the theoretical properties of $\widehat{\boldsymbol{\gamma}}_j^\circ = (\widehat{\boldsymbol{\theta}}_j^{\mathrm{T}}, \widehat{\boldsymbol{\gamma}}_j^{\mathrm{T}})^{\mathrm{T}}$ from Equation (2.9), we propose the following assumptions.

Assumption 1. *For each $j \in [d]$, suppose the rows $\{\mathbf{x}_{-j}^{(i)} : 1 \leq i \leq n\}$ in the design matrix are independent and identically distributed random vectors with $\boldsymbol{\Sigma} = \mathbb{E}[(1\ (\mathbf{x}_{-j}^{(1)})^{\mathrm{T}})^{\mathrm{T}}(1\ (\mathbf{x}_{-j}^{(1)})^{\mathrm{T}})]$, where $\boldsymbol{\Sigma}$ satisfies $\lambda_{\min}(\boldsymbol{\Sigma}) \geq c_\lambda$ for some $c_\lambda > 0$.*

Assumption 2. *With probability larger than* $1 - M_n$ *where* $M_n \to 0$ *as* $n \to \infty$*, there exists an absolute constant* $C_0 > 0$*, such that* $\max_{i\in[n],k\in[K],j\in[d]} |\zeta_{jk}^{(i)}| \leq C_0$ *with* $\zeta_{jk}^{(i)} = \theta_{jk}^* + \boldsymbol{\gamma}_{j\bullet,k\bullet}^{*\mathrm{T}} \mathbf{x}_{-j}^{(i)}$*, which implies that* $\min_{i\in[n],k\in[K],j\in[d]} \exp(\zeta_{jk}^{(i)})\{1 + \sum_{l=1}^K \exp(\zeta_{jl}^{(i)})\}^{-1} \geq c_*$ *and* $\min_{i\in[n],j\in[d]}\{1 + \sum_{l=1}^K \exp(\zeta_{jl}^{(i)})\}^{-1} \geq c_*$*, for some* $c_* > 0$*.*

Conditions similar to Assumption 1 regarding the eigenvalues of $\boldsymbol{\Sigma}$ are commonly adopted in high-dimensional statistics (Cai et al., 2022; Zhang and Li, 2023; Tian and Feng, 2023). Unlike many studies on high-dimensional regressions, the rows in our design matrix are not sub-Gaussian vectors but instead have bounded entries. This distinction imposes a stronger sparsity requirement compared to sub-Gaussian cases; see the discussions following Theorem 2.3.1 for more details. Conditions similar to Assumption 2 are also frequently used for the high-dimensional multinomial regression (Tian et al., 2024; Abramovich et al., 2021) and high-dimensional logistic regression (Guo et al., 2021; Ma et al., 2022). The factors c_λ and c_* play important roles in establishing the convergence rate of an estimator in multinomial regression; see more discussions after Theorem 2.3.1. Notably, the Potts model reduces to the Ising model when $K = 1$, and similar assumptions on $c_\lambda, c_* = O(1)$ are often imposed in literature (Ravikumar et al., 2010; Cai et al., 2019).

2.3.2 Guarantees on estimated parameters

For a constant $\eta \in (0, 1/2)$, define $\delta(\lambda) = \exp[-\{(\eta\lambda\sqrt{n})/40\}^2]$ so that

$$\lambda = 40\eta^{-1}\sqrt{n^{-1}\log(1/\delta(\lambda))}$$

. Set

$$\lambda_\sharp = 40\sigma(\eta n^{1/2})^{-1}\{2(s_0^\circ)^{-1}\log(4ed/s_g^\circ) + \log(2eK^2/s_0^\circ)\}, \tag{2.11}$$

where $\sigma^2 = \max_{i,k} \mathrm{Var}(y_k^{(i)}|\mathbf{x}^{(i)})$ is the conditional variance of the responses, and $s_0^\circ = s^\circ/s_g^\circ$ represents the average sparsity per group in the true groups. For non-negative sequences $\{a_n\}$ and $\{b_n\}$, $a_n = o(b_n)$ or $a_n \ll b_n$ means $a_n/b_n \to 0$ as $n \to \infty$, and $a_n = O(b_n)$ or $a_n \lesssim b_n$ means $a_n \leq cb_n$ for some absolute constant $c > 0$ and sufficiently large n. Recall that $\boldsymbol{\gamma}_j^*$ are (s, s_g)-sparse

for each j and $\boldsymbol{\gamma}_j^{*\circ}$ are (s°, s_g°)-sparse with $s^\circ = s + K$ and $s_g^\circ = s + 1$. We have the following guarantees on the proposed estimators.

Theorem 2.3.1. *Suppose Assumptions 1 and 2 hold, and* $s^\circ \ll c_*^2\sqrt{n/\log(dK)}$. *Let* $\delta_0 \in (0,1)$ *satisfies* $\log(1/\delta_0)\{s^\circ \log(1/\delta(\lambda_\sharp))\}^{-1} = O(1)$. *For each site* $j \in [d]$, *there exists an absolute constant* $C > 0$, *such that with probability at least* $1 - C((d-1)K+1)^{-1} - M_n - \delta_0$, *the estimator* $(\widehat{\boldsymbol{\theta}}_j, \widehat{\boldsymbol{\gamma}}_j)$ *from Equation* (2.9) *with* $\lambda = 2\lambda_\sharp$ *and* $\lambda_g = \sqrt{s/s_g}\lambda$ *satisfies for* $q = 1, 2$

$$(\|\widehat{\boldsymbol{\theta}}_j - \boldsymbol{\theta}_j^*\|_q^q + \|\widehat{\boldsymbol{\gamma}}_j - \boldsymbol{\gamma}_j^*\|_q^q)^{1/q} \leq C R_K B_n, \tag{2.12}$$

where $R_K = \sigma/(c_\lambda c_*^2)$ *and* $B_n = (s^\circ)^{1/q} n^{-1/2}\{2(s_0^\circ)^{-1}\log(4ed/s_g^\circ) + \log(2eK^2/s_0^\circ)\}^{1/2}$.

The condition $s^\circ \ll c_*^2\sqrt{n/\log(dK)}$ is necessary for two main reasons. First, the factor c_*^2 arises from the multinomial nature of the Potts model, particularly in the quadratic approximation analysis for the likelihood function. Second, while sub-Gaussian designs for high-dimensional regression typically require $s^\circ \ll c_*^2 n/\log(dK)$, our design matrix, with rows $\{\mathbf{x}_{-j}^{(i)}, 1 \leq i \leq n\}$, has bounded elements and falls outside the sub-Gaussian framework. In such cases, stricter sparsity requirement for s° is often required, as noted in Theorem 2.4 of van de Geer et al. (2014).

In Theorem 2.3.1, λ and λ_g depend on (s, s_g), while the convergence rate is determined by (s°, s_g°). This arises because our estimation procedure penalizes only the dependency vector $\boldsymbol{\gamma}_j$. As the true dependency vector $\boldsymbol{\gamma}_j^*$ is (s, s_g)-sparse, it is intuitive that the tuning parameters are related to (s, s_g). Since the single-site effects vector $\boldsymbol{\theta}_j$ is not penalized and $\boldsymbol{\gamma}_j^{\circ *}$ is (s°, s_g°)-sparse, leading to a convergence rate dependent on (s°, s_g°).

The convergence rate in Equation (2.12) is B_n when R_K is an absolute constant, which holds if K is fixed, including the important Ising model with $K = 1$. In this case, the convergence rate in Theorem 2.3.1 matches the minimax optimal rate for linear models. By taking

$$\delta_0 = \exp[-C_2\{2s_g^\circ \log(4ed/s_g^\circ) + s^\circ \log(2eK^2/s_0^\circ)\}],$$

we have $\delta_0 = o(1)$ as long as d diverges.

The condition $\log(1/\delta_0)\{s^\circ \log(1/\delta(\lambda_\sharp))\}^{-1} = O(1)$ is then satisfied with this δ_0. As a result, with probability approaching one as $n \to \infty$, we have

$$(\|\widehat{\boldsymbol{\theta}}_j - \boldsymbol{\theta}_j^*\|_2^2 + \|\widehat{\boldsymbol{\gamma}}_j - \boldsymbol{\gamma}_j^*\|_2^2)^{1/2} \lesssim \frac{1}{\sqrt{n}} \left\{ 2s_g^\circ \log\left(\frac{4ed}{s_g^\circ}\right) + s^\circ \log\left(\frac{2es_g^\circ K^2}{s^\circ}\right) \right\}^{1/2}. \tag{2.13}$$

Compared to the ℓ_2 error bounds in Cai et al. (2022) and Zhang and Li (2023) for linear models with sparse group structures, our error bound in Equation (2.13) sharpens their results by reducing logarithmic factors. Specifically, the second term in Equation (2.13) from previous works is at least of order $s^\circ \log(eK^2)$, which is strictly larger than our bound. Moreover, our error bound tightly matches the minimax lower bound, including the logarithmic factors, derived in Cai et al. (2022) for linear models. Our analysis leverages sharp oracle inequalities recently developed in Bellec et al. (2018) and Li et al. (2023) for high-dimensional linear models with ℓ_1- and sparse group Lasso penalties. Unlike standard Lasso-type analysis (Wainwright, 2019), the new oracle inequalities are based on a refined analysis of the concentration of the sum of the leading entries in the non-increasing rearrangement of the magnitudes of the products of covariates and error terms. However, adapting this analysis to the Potts model requires substantial modifications; see, for example, Lemma A.2.1 in Chapter A.2 for details.

In Equation (2.12), R_K may depend on K and is determined by three factors: the smallest eigenvalue of $\boldsymbol{\Sigma}$ for the covariates, governed by c_λ; the minimum success probability level c_*; and the conditional variance σ^2 of the response. The product $c_\lambda c_*^2$ provides a lower bound for the smallest eigenvalue of the Hessian matrix of the log-likelihood function, which is typically assumed to be $O(1)$ in linear or logistic regression. In logistic regression, $\sigma^2 = O(1)$ under a common boundedness condition similar to Assumption 2. In multinomial regression, even with Assumption 2, the variance level σ^2 is of order $O(1/K)$, differing fundamentally from logistic regression. While $c_\lambda = O(1)$ is a natural assumption, the smallest eigenvalue of the Hessian matrix for the multinomial log-likelihood is tightly bounded below by $c_*^2 = O(1/K^2)$, as shown in Lemma A.2.5 in Chapter A.2. This suggests that the multinomial regression likelihood

exhibits singularities as K diverges. This characterization of c_*^2 is a substantial distinction between multinomial and logistic regression that remains underexplored in the literature. Notably, in Tian et al. (2024), the authors studied ℓ_1-penalized multinomial regression under similar assumptions as ours, with $c_\lambda = O(1)$. However, their ℓ_2 error bound involves $K^{5/2}$ factor, significantly larger than our result with $K^{3/2}$. Finally, the Potts model shares all the intrinsic features of multinomial regression but admits $c_\lambda = O(1/K)$ due to its unique covariate structure. This can be intuitively understood by considering a case where all sites in $\mathbf{z}_{-j}$ are independent for some site j. Here, $c_\lambda = \min_{0\le k\le K} p_{jk} = O(c_*) = O(1/K)$, where $p_{jk} = P(z_{jk} = 1)$.

Theorem 2.3.1 shows that the proposed estimator $(\widehat{\boldsymbol{\theta}}_j, \widehat{\boldsymbol{\gamma}}_j)$ from Equation (2.9) achieves an improved error bound when the parameter vector is simultaneously element-wise and group-wise sparse. Recall that $\boldsymbol{\gamma}_j^{\circ *} = (\boldsymbol{\theta}_j^*, \boldsymbol{\gamma}_j^*)$ has length $D^\circ = K + (d-1)K^2$, with d groups and is (s°, s_g°)-sparse. Using only an ℓ_1 penalty $\lambda\|\boldsymbol{\gamma}\|_1$ in Equation (2.9) yields an ℓ_2 estimation error of order $\sqrt{s^\circ n^{-1}\log(D^\circ/s^\circ)}$ (Bellec et al., 2018), larger than the order in Equation (2.13) when $s_g^\circ/d = o(1)$ and $s_g^\circ/s^\circ = o(1)$. Similarly, using only a group penalty $\lambda_g \sum_{r\neq j}\|\gamma_{j(r)}\|_2$ results in an ℓ_2 error bound of order $\sqrt{s_g^\circ n^{-1}(\log d + K^2)}$ (Lounici et al., 2011), which exceeds the order in Equation (2.13) when $\log d/K^2 = o(1)$ and $s^\circ/s_g^\circ = o(K^2/\log K^2)$. Thus, the proposed estimator achieves a tighter ℓ_2 error bound compared to estimators using only ℓ_1- or group penalties, when the true underlying parameter vector is both element-wise and group-wise sparse.

2.3.3 Consistency on estimated energy for mutation fitness

From Equation (2.12), we also have the ℓ_1 error bound of the proposed estimator:

$$\|\widehat{\boldsymbol{\theta}}_j - \boldsymbol{\theta}_j^*\|_1 + \|\widehat{\boldsymbol{\gamma}}_j - \boldsymbol{\gamma}_j^*\|_1 \lesssim R_K\sqrt{\frac{s^\circ}{n}}\left\{2s_g^\circ\log\left(\frac{4ed}{s_g^\circ}\right) + s^\circ\log\left(\frac{2es_g^\circ K^2}{s^\circ}\right)\right\}^{1/2},$$

which immediately implies the consistency for $\widehat{\Delta E}_{j,k}$, the plug-in estimated relative energy change using $(\widehat{\boldsymbol{\theta}}_j, \widehat{\boldsymbol{\gamma}}_j)$ from Equation (2.9) for $\Delta E_{j,k}$ defined in Equation (2.3), as summarized below.

Corollary 2.3.2. *Under the assumptions of Theorem 2.3.1, for each site $j \in [d]$ and amino acid type $k \in [K]$, with the same probability therein, we have $|\widehat{\Delta E}_{j,k} - \Delta E_{j,k}| \lesssim R_K B_n$, where R_K and B_n are defined in Theorem 2.3.1.*

Building on the results for each site in the Potts model, we can straightforwardly establish global results for the entire sequence. To enforce symmetry in the estimator $(\widehat{\boldsymbol{\theta}}_j, \widehat{\boldsymbol{\gamma}}_j)$ from Equation (2.9) for site j, we average the estimates in Section 2.2.3, yielding the symmetrized estimator $\tilde{\boldsymbol{\gamma}}_j$. Specifically, $\tilde{\gamma}_{jr,kl} = \tilde{\gamma}_{rj,lk} = (\widehat{\gamma}_{jr,kl} + \widehat{\gamma}_{rj,lk})/2$ for each $r \neq j$. Below, we summarize the global results for $\{(\widehat{\boldsymbol{\theta}}_j, \tilde{\boldsymbol{\gamma}}_j) : j \in [d]\}$, based on which the theoretical guarantees for the plug-in estimator of the energy changes for multiple-site mutations can be readily established.

Corollary 2.3.3. *Under the assumptions of Theorem 2.3.1, with the same probability and R_K, B_n specified therein, we have*

$$\sum_{j=1}^{d} (\|\widehat{\boldsymbol{\theta}}_j - \boldsymbol{\theta}_j^*\|_2^2 + \|\tilde{\boldsymbol{\gamma}}_j - \boldsymbol{\gamma}_j^*\|_2^2) \lesssim d R_K^2 B_n^2 \ \text{ and } \ \sum_{j=1}^{d} (\|\widehat{\boldsymbol{\theta}}_j - \boldsymbol{\theta}_j^*\|_1 + \|\tilde{\boldsymbol{\gamma}}_j - \boldsymbol{\gamma}_j^*\|_1) \lesssim d R_K B_n.$$

2.4 Integration of Structural Information via Group Weights

The sparse group Lasso penalty ensures the desired site-wise and element-wise sparsity in our method but may not fully leverage the dependencies between spatially proximal sites (Marks et al., 2012). To address this, we first retrieve one representative 3-dimensional protein structure (Alphafold prediction) for each family from the UniProt database (Consortium, 2020). For each site pair (j, r), spatial proximity is calculated as the Euclidean distance (in $\mathring{A}$) between the 3-dimensional coordinates of their respective alpha-carbon atoms, denoted as D_{jr}.

To integrate this structural information into our method, we then introduce group weights determined by physical distances within the protein's structure. That is, for each site $j \in [d]$, we define the group weight w_{jr} in Equation (2.10) as

$$w_{jr} = (\sqrt{d_r/n} + \sqrt{2\log(d-1)/n}) \mathcal{K}(D_{jr}), \tag{2.14}$$

where d_r denotes the group size of $\boldsymbol{\gamma}_{j(r)}$, specifically K^2 as defined in Equation (2.7), and $\mathcal{K}(D_{jr})$ is a function that incorporates structural information into the weights.

To determine the form of $\mathcal{K}(\cdot)$ from data, we draw inspiration from the adaptive group Lasso (Huang et al., 2008). We consider four protein families: DYR, Trypsin-2 (TRY2), Tyrosine-protein kinase Fyn (FYN), and Yes-associated protein (YAP), and estimate $\widehat{\boldsymbol{\gamma}}_{j(r)}^{(g)}$ using multinomial regression with group Lasso for all site pairs (j, r). Next, we examine the relationship between $\|\widehat{\boldsymbol{\gamma}}_{j(r)}^{(g)}\|_2$ and D_{jr}. As shown in Figure 2.1, there is a consistent trend across the protein families under considerations, with greater magnitude of dependency parameters observed at spatially closer sites. This aligns with the biological intuition that direct couplings within sequences are expected to be stronger at spatially closer sites. It suggests that $\mathcal{K}(\cdot)$ should assign larger group weights to more distant sites to encourage greater sparsity in their dependency parameters. Based on this observation, we consider

$$\mathcal{K}(D_{jr}) = 1 - \exp(-D_{jr}^2/\mathrm{MS}_j), \tag{2.15}$$

where $\mathrm{MS}_j = (d-1)^{-1}\sum_{r\neq j}[D_{jr} - (d-1)^{-1}\sum_{r'\neq j} D_{jr'}]^2$ is chosen to normalize the distances between site j and other sites (Li and Luan, 2003; Wang et al., 2009). In Section 2.5, we also explore alternative forms of $\mathcal{K}(\cdot)$ to demonstrate the robustness of the choice in Equation (2.15).

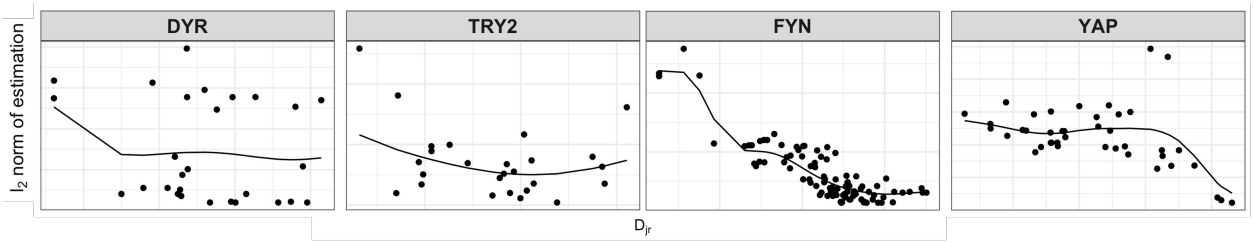


Figure 2.1: The distances D_{jr} between sites j and r versus $\|\widehat{\boldsymbol{\gamma}}_{j(r)}^{(g)}\|_2$ with the fitted trend.

Remark 2.4.1. *When group weights are considered, the theoretical results in Section 2.3 can be easily extended to cases where the weights are fixed or derived from an independent dataset, such*

as the protein's structural data, by adjusting $\lambda_\sharp$ *of Equation* (2.11) *to* $\lambda_\sharp / w_{\min}$*, where* $w_{\min} = \min_{r \neq j, w_{jr} \neq 0} w_{jr}$*. As* $w_{\min}$ *is independent of the sample size of the training dataset, this adjustment has minimal impact on the convergence rate.*

2.5 Evidences from Numerical Experiments

In this section, we evaluate our proposed method numerically and compare it with following: node-wise Lasso with $\lambda_g = 0$ in Equation (2.10), node-wise sparse group Lasso with $w_{jr} = 1$ for all j, r in Equation (2.10), and node-wise ridge regression as implemented in EVM (Hopf et al., 2017).

2.5.1 Settings and implementations

For all numerical experiments, we generate n independent d-dimensional sequences. Data generation for the Potts model is non-trivial due to the computational challenges of the large state space of $\mathbf{z}$, even for modest K and d (Izenman, 2021). To address this, we use a Gibbs sampler based on the conditional probability in Equation (2.5) to sequentially sample each site. Further details on the data generation are provided in Section A.3.1.

The independent entries of model parameters $\boldsymbol{\theta}_{d \times K}$ are generated from $\mathrm{Unif}(0, 2)$. To introduce structural information among sites, we generate a distance matrix $\mathbf{D}$ with independent symmetric entries $D_{jr} \sim 40\mathrm{Beta}(2, 2)$ for $j < r$, where D_{jr} provides the distance between sites j and r.

We also generate a binary adjacency matrix $\mathbf{A}$ with independent entries $A_{jr} \sim \mathrm{Ber}(p_{jr})$, where $A_{jr} = 0$ indicates $\boldsymbol{\gamma}_{j(r)} = 0$. Here, $p_{jr} \in (0, 1)$ controls group-wise sparsity s_g. For element-wise sparsity, we set $\gamma_{jr,kl} = 0$ for $6 \leq k, l \leq K$ and each j, r. We consider two settings of coefficients: (M1) where the magnitude of coefficients, i.e., the signal strength between sites, is related to their distance, and (M2) where the connection probability between sites is distance-dependent.

(M1) Set $p_{jr} = \log d/(2d)$, making the site-wise connections sparse (Bollobás and Riordan, 2011). If $A_{jr} = 1$, we set $\gamma_{jr,kl} = \exp(-D_{jr}^2/\mathrm{MS}_j) u_{jr,kl}$ as nonzero entries for all $1 \leq k, l \leq 5$,

where $u_{jr,kl}$ is independently sampled from $\mathrm{Unif}([-2,-0.5]\cup[0.5,2])$, and MS_j is defined in Equation (2.15).

(M2) The nonzero entries $\gamma_{jr,kl}$ are independently sampled from $\mathrm{Unif}([-2,-0.5]\cup[0.5,2])$, with $p_{jr} = \tau\exp(-D_{jr}^2/\mathrm{MS}_j)[\sum_{r'}\exp(-D_{jr'}^2/\mathrm{MS}_j)]^{-1}$, where MS_j is defined as in (M1). To control the group sparsity s_g, $\tau = 3$ is used for $d = 25$ and $\tau = 1.5$ for $d = 50$.

Tuning parameters for all methods are selected via 5-fold cross-validation. For our method and node-wise sparse group Lasso, which involve two tuning parameters λ_g and λ, we search over the grid $\{(i2^j,(1-i)2^j) : i \in I; j \in J\}$ to reduce computational cost with $I = \{0, 0.1, 0.2, ..., 1\}$ and $J = \{-5,-4,-3,-2,-1,-0.5,0,0.5,1,2\}$. If the simulated data are highly imbalanced such that frequency of $z_{jk_0} = 1$ falls below 10 for some $j \in [d]$ and $k_0 \in [K]$, the corresponding observations are excluded, and we set $\widehat{\gamma}_{i\bullet,k_0\bullet} = 0$ by default to ensure valid cross-validation.

We also explore two ways for constructing group weights w_{jr} of the form Equation (2.14) using different $\mathcal{K}(D_{jr})$ to examine how the choice of $\mathcal{K}(D_{jr})$ affects coefficient estimation. In (N1), $\mathcal{K}(D_{jr})$ matches the form used in the data generating process, while in (N2), it differs but retains a similar trend.

(N1) Set $\mathcal{K}(D_{jr}) = 1 - \exp(-D_{jr}^2/\mathrm{MS}_j)$, consistent with how D_{jr} is used for generating coefficients in (M1) and (M2).

(N2) Set $\mathcal{K}(D_{jr}) = \exp(D_{jr})[1+\exp(D_{jr})]^{-1}$, commonly used to generate adjacency matrices in the Erdős-Rényi model in network analysis.

2.5.2 Results

To assess the estimation accuracy of a method, we use the mean squared estimation error (MSE), i.e., $\sum_{j=1}^d \|\boldsymbol{\gamma}_j - \tilde{\boldsymbol{\gamma}}_j\|_2^2$. For selection accuracy, we consider the true positive rate (TPR) and the false discovery rate (FDR) for identifying nonzero $\gamma_{jr,kl}$ and groups $\boldsymbol{\gamma}_{j(r)}$. To clarify, TPR is defined as TP/(TP+FN), where TP and FN are the numbers of correctly and incorrectly estimated nonzero parameters, and FDR is FP/(FP+TP), where FP is the number of incorrectly estimated

nonzero parameters. The TPR and FDR for identifying nonzero groups are defined similarly and denoted as TPR_{g} and FDR_{g}. We evaluate our model and competitors under (M1) and (M2) with $n = 1000, 2000, 4000$, $d = 25, 50$, and $K = 20$. All measures are computed based on 100 independent Monte Carlo replicates.

Table 2.1 reports the mean squared error (MSE) along with entry-wise and group-wise TPRs and FDRs for the (M1) setting. The results show that our method outperforms competitors in both estimation and selection accuracy across varying sample sizes n and numbers of sites d when signal strengths depend on distances. Similarly, as shown in Table A.1, our method also excels under (M2) when site connection probabilities are determined by their distances. These findings confirm the effectiveness of our method for accurately estimating coefficients when signal strengths or inter-site connections are physically distance-dependent. By incorporating distance information into parameter estimation, our method provides reliable estimates, even when the chosen $\mathcal{K}(D_{jr})$ only approximates the desired trend, as in (N2). This robustness supports the form in Equation (2.15) for physical distance-based group weights in practice. Additionally, estimation errors decrease and selection accuracy improves as n increases or d decreases, consistent with Theorem 2.3.1. Owing to space constraints, detailed results for (M2) and more settings with greater d and n, aligned with the real data in Section 2.6, are presented in Section A.3.2.

2.6 Protein Mutation Analysis and Fitness Landscape

To demonstrate our method for mutation analysis, we apply it to MSA datasets from 12 protein families, each accompanied by a representative protein structure and experimentally evaluated mutation effects.

In MSA data analysis, sequence sampling bias commonly arises from closely related species, leading to an overrepresentation of protein sequences with high similarities. This imbalance in taxonomic diversity skews mutation rate sampling and introduces spurious correlation signals (Hopf et al., 2017; Marks et al., 2011). Following Morcos et al. (2011), we used sample weights $\boldsymbol{\omega} = (\omega_1, ..., \omega_n)^{\mathrm{T}}$ to account for redundancies within a protein family by applying a threshold to

Table 2.1: Results in (M1) with $K = 20$

d	$\sum_{j<r} A_{jr}$	Methods	n	MSE	TPR	FDR	TPR_g	FDR_g
25	25	Our method with (N1)	1000	24.595	0.798	0.054	0.960	0.250
		Our method with (N2)		30.332	0.747	0.084	0.940	0.242
		Lasso		47.192	0.694	0.392	0.860	0.403
		Sparse Group Lasso		38.513	0.731	0.120	0.920	0.270
		Ridge		84.050	–	–	–	–
		Our method with (N1)	2000	17.368	0.842	0.051	0.980	0.197
		Our method with (N2)		25.357	0.790	0.080	0.980	0.210
		Lasso		36.148	0.724	0.352	0.900	0.365
		Sparse Group Lasso		29.429	0.757	0.116	0.940	0.254
		Ridge		60.954	–	–	–	–
		Our method with (N1)	4000	14.934	0.871	0.043	1.000	0.138
		Our method with (N2)		18.172	0.813	0.069	1.000	0.153
		Lasso		27.637	0.788	0.314	0.960	0.308
		Sparse Group Lasso		23.080	0.810	0.085	1.000	0.242
		Ridge		48.695	–	–	–	–
50	78	Our method with (N1)	1000	67.819	0.758	0.077	0.923	0.242
		Our method with (N2)		73.855	0.712	0.105	0.904	0.265
		Lasso		94.631	0.633	0.374	0.852	0.355
		Sparse Group Lasso		80.518	0.686	0.177	0.846	0.284
		Ridge		127.514	–	–	–	–
		Our method with (N1)	2000	50.145	0.803	0.073	0.944	0.216
		Our method with (N2)		57.254	0.762	0.093	0.936	0.239
		Lasso		73.953	0.677	0.293	0.878	0.327
		Sparse Group Lasso		62.148	0.724	0.132	0.913	0.253
		Ridge		88.367	–	–	–	–
		Our method with (N1)	4000	41.512	0.849	0.064	1.000	0.194
		Our method with (N2)		44.416	0.810	0.112	1.000	0.218
		Lasso		51.794	0.720	0.278	0.926	0.275
		Sparse Group Lasso		48.268	0.783	0.112	1.000	0.243
		Ridge		72.620	–	–	–	–

the normalized Hamming distances between sequences, where $\omega_i = \{\sum_{i' \neq i} \mathbb{1}\{D_{\text{HM}}(\mathbf{x}_{-j}^{(i)}, \mathbf{x}_{-j}^{(i')}) < 0.2\}\}^{-1}$ and D_{HM} is the normalized Hamming distance. In practice, sample weights can be computed from the same data or estimated from an independent source. Although our theoretical

results assume equal sample weights, they readily extend to cases with fixed or independently estimated sample weights.

2.6.1 Predicted energy changes versus experimental fitness

As discussed in Section 2.2, our method evaluates energy changes, reflecting the relative favorability of site-specific mutations. To validate our method, we compute the Spearman correlation between our estimated energy changes and experimentally determined mutation fitness. As shown in Section 2.5 later, our method is relatively robust when $\mathcal{K}(D_{jr})$ is of the form in Equation (2.15), reflecting the biological intuition that direct couplings between sites tend to decrease as inter-site distances increase. Thus, we adopt group weights Equation (2.14) with $\mathcal{K}(D_{jr})$ of the form in Equation (2.15) for analysis. We compare our results with predictions from other methods, including the widely-used EVM, which does not account for residue structural information or group sparsity. We also consider three sparse group Lasso estimators: no weights (SGL) with all $w_{jr} = 1$ in Equation (2.10), adaptive weights, and refitted weights. Both estimators with adaptive and refitted weights start with an initial estimate $\widehat{\boldsymbol{\gamma}}^{(g)}$ from group Lasso with $w_{jr} = (\sqrt{d_r/n} + \sqrt{2\log(d-1)/n})$. For refitted weights, we use the generalized additive model in `R` package `mgcv` to fit D_{jr} and $\|\widehat{\boldsymbol{\gamma}}^{(g)}_{j(r)}\|_2$, yielding $\|\widehat{\boldsymbol{\gamma}}^{(g)}_{j(r)}\|_2 = \widehat{f}(D_{jr})$, and then re-run the node-wise multinomial regression with a sparse group Lasso penalty and group weights $w_{jr} = [\widehat{f}(D_{jr})]^{-1}$. While refitted weights incorporate structural information between sites, they do not specifically model the decay of direct couplings with distance. For adaptive weights, we re-run the node-wise multinomial regression with group weights $w_{jr} = \|\widehat{\boldsymbol{\gamma}}^{(g)}_{j(r)}\|_2^{-1}$, which are adaptive (Huang et al., 2008) yet do not account for structural information.

Table 2.2 summarizes mutagenesis experiments of 12 protein families. These protein families encompass a diverse array of biological functions, including enzymatic activity in antibiotic resistance (e.g., Beta-lactamase TEM (BLAT), Aminoglycoside 3'-phosphotransferase (KKA2)), digestion (e.g., TRY2), molecular chaperoning under stress (e.g., yeast ortholog of heat shock protein 90 (HSP82)), and cellular signaling and regulation (e.g., Disks large homolog 4

Table 2.2: Spearman correlations between estimated energy change and experimental mutation fitness.

Protein Family	Sequences Number (n)	Sites (d)	Exp. Data	Mutation Feature	Our Method	EVM	Refitted Weights	Adaptive Weights	SGL
BLAT	8403	263	4611	T_m	**0.65**	0.57	0.35	0.42	0.37
DLG4	102410	101	1577	CRIPT	**0.55**	0.54	0.41	0.37	0.39
DYR	8494	158	16	abundance 37	**0.86**	0.75	0.81	0.83	0.73
FYN	115571	66	42	T_m	0.70	0.63	0.66	**0.73**	0.62
GAL4	17521	75	1196	SEL	**0.64**	0.59	0.52	0.60	0.48
HSP82	15329	240	4323	SEL	**0.57**	0.49	0.24	0.30	0.33
KKA2	12861	264	4385	$\text{Kan}_{1:8}$	0.62	0.49	0.39	**0.67**	0.42
PYP	124287	125	125	ΔG_U	**0.57**	0.52	0.40	0.35	0.42
YAP1	40302	36	363	linear	**0.63**	0.44	0.58	0.57	0.49
MTH3	14115	330	1957	W_{rel}	**0.52**	0.51	0.16	0.38	0.46
TRY2	47913	223	14	$\log(k_{cat}/K_m)$	**0.14**	-0.13	0.13	**0.14**	0.10
UBE4B	9172	104	900	$\log_2$ ratio	**0.47**	0.42	0.19	0.45	0.32

(DLG4), FYN, Galactose-responsive transcription factor 4 (GAL4), and Ubiquitination Factor E4B (UBE4B)). The column "Exp. Data" indicates the number of experiments conducted, each introducing a single mutation at one site on the wild-type sequence. The "Mutation Feature" column lists measures used to quantify mutation fitness, with various metrics applied in these studies. We denote T_m as the denaturation midpoint temperature, where 50% of proteins are folded, analogous to thermodynamic stability (ΔG_U). *CRIPT* refers to cysteine-rich interactor (McLaughlin Jr. et al., 2012), while *abundance 37* indicates intracellular abundance at $37°C$ (Bershtein et al., 2012). *SEL* represents functional selection coefficients (Kitzman et al., 2015). *Linear* and $\log_2$ *ratio* are different functions performing on the enrichment of variants (Araya et al., 2012; Starita et al., 2013), and W_{rel} denotes relative fitness effects (Rockah-Shmuel et al., 2015). Finally, k_{cat}/K_m represents the conversion rate at minimal substrate concentration, and $\textit{Kan}_{1:8}$ refers to kanamycin substrate with aminoglycosides at 1:8 dilutions (Melnikov et al., 2014).

Table 2.2 shows that the estimated energy changes from our method exhibit a strong correlation with experimental mutation fitness across all protein families. Particularly, our method outperforms all other methods for ten out of all twelve families, with slightly lower performance than the adaptive weights method in the remaining two. This confirms the advantages of incorporating structural information and group sparsity in predicting mutation fitness.

2.6.2 Mutation analysis for the Dihydrofolate reductase protein

We showcase our method for mutation analysis on two protein families: DYR in this section and Postsynaptic density protein 95 (PSD95) in Sec 2.6.3. Dysregulation of DYR activity is associated with various diseases and studying mutation patterns in the DYR family has broader implications for health and disease treatment (Baccanari et al., 1981; Schweitzer et al., 1990). Figure 2.2(a) presents the *fitness landscape*, where each block represents $\widehat{\Delta E}_{j,k}$, as defined in Equation (2.3), for the mutant of amino acid k from the wild-type at site j. The y-axis corresponds to selected sites, and the x-axis represents the 20 amino acid types. The color gradient (blue to white to red) reflects increasing $\widehat{\Delta E}_{j,k}$, with higher values (red) indicating favorable mutations and lower values (blue) indicating unfavorable ones. Wild-type amino acids at each site are shown in white. This landscape helps identify mutations favored by evolution. In this landscape, sites such as 40, 112, 115, and 133, shown in blue, represent highly conserved sites where most amino acid changes are unfavorable. Conversely, sites such as 12, 88, 127, and 145, shown in lighter red tones, exhibit greater tolerance to mutations.

Amino acid variations at coevolved site pairs exhibit strong mutation dependence, indicating constraints on changes at these sites and they often interact within close spatial proximity in the protein structure. Our approach calculates amino acid-wise dependencies between coevolved sites, which offers finer resolution of interactions between different amino acid types. Figure 2.2(b) presents a Sankey plot illustrating amino acid-wise dependencies between sites 12 and 127. Each side lists amino acid types observed at the MSA data, with connections between sites representing

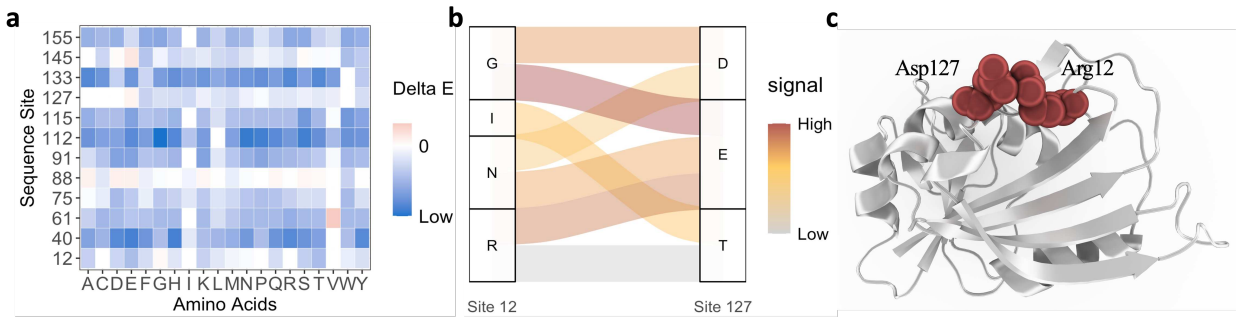


Figure 2.2: Predicted mutation fitness of DYR. (a) Landscape of estimated energy changes for different amino acids occurring at various sites. (b) Sankey plot showing amino acid-wise dependencies between sites 12 and 127. (c) Two coevolved residues, sites 12 (Arg12, right) and 127 (Asp127, left), forming a contact in the wild-type protein structure (DYR in E.coli). The two contacting residues are highlighted in red and shown in the sphere representation.

$\gamma_{jr,kl} + \theta_{jk} + \theta_{rl}$. Dark orange connections indicate a preference for co-occurrence of two amino acids, whereas lighter-colored connections indicate repulsion.

In Figure 2.2(b), we show two pairs of highly dependent amino acid types: (1) polar amino acids Asparagine (N) and Arginine (R) at site 12 with polar amino acids Aspartate (D) and Glutamate (E) at site 127, and (2) hydrophobic amino acids Glycine (G) and Isoleucine (I) at site 12 with polar amino acids Aspartate (D), Glutamate (E), and Threonine (T) at site 127. The amino acid composition at site 127 suggests conservation of polar amino acids, while site 12 shows greater tolerance for amino acids with different properties. The interaction between polar amino acids indicates potential charge compensation is preferred at those sites. In Figure 2.2(c), we show that the two residues corresponding to sites 12 and 127 are in close contact on the protein structure of DYR in E.coli. A covalent bond may form between the negatively charged Aspartate (D, left) and the positively charged Arginine (R, right).

2.6.3 Mutation analysis for the Postsynaptic density protein

PSD95, also known as Disks large homolog 4 (DLG4), is a postsynaptic scaffolding protein involved in synaptogenesis and synaptic plasticity (Prange et al., 2004; Gao et al., 2023). The mutation fitness landscape of the DLG4 family generated by our method is presented in Figure 2.3(a).

Figure 2.3(a) highlights framed regions covering sites 24-26 and 63-65, with low average mutation fitness values (mostly blue), suggesting weak tolerance to mutations at these sites. These regions are more conserved in the MSA data, often indicating functional or structural importance. Figure 2.3(b) shows amino acid-wise dependencies between two strongly dependent sites, 13 and 58. Similar to Figure 2.2(b), each side of the Sankey plot in Figure 2.3(b) lists amino acid types observed at the MSA data, with connections indicating the strength of the dependency between amino acid types.

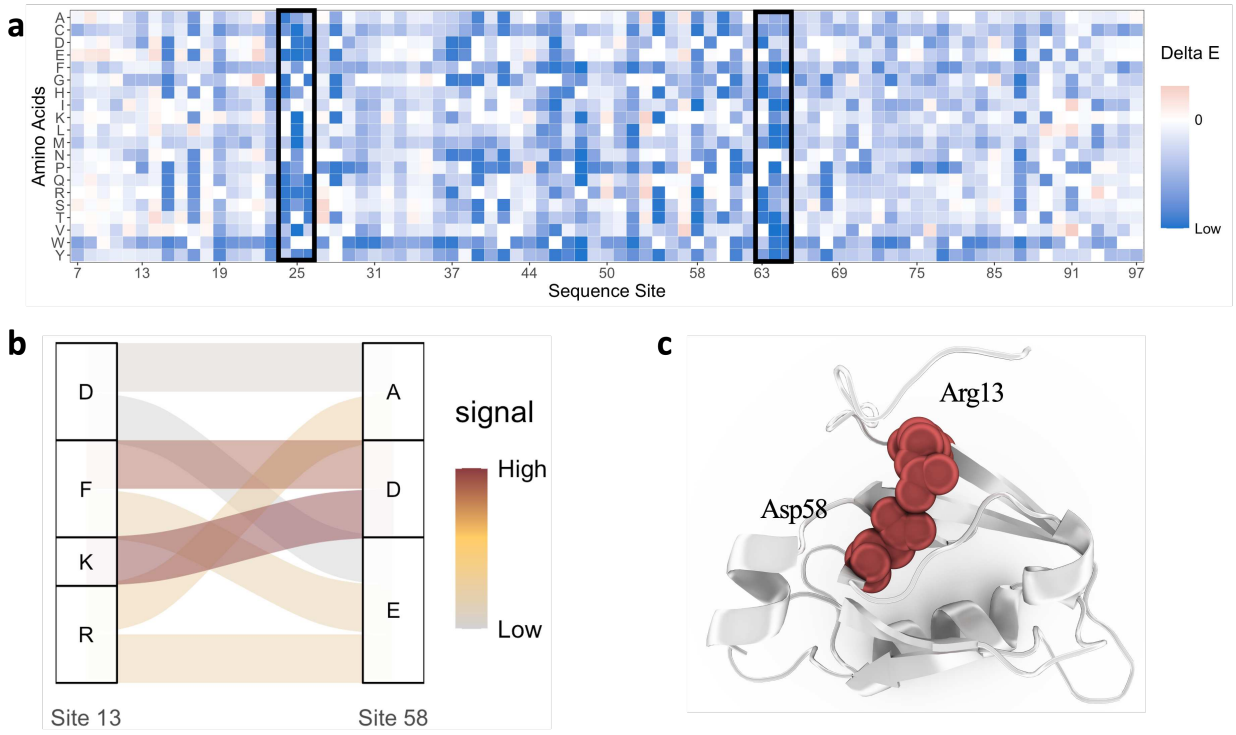


Figure 2.3: Predicted mutation fitness of DLG4. (a) Landscape of estimated energy change for different amino acids at each sites. The framed regions exhibit low mutation fitness. (b) The Sankey plot showing amino acid-wise dependence between sites 13 and 58. (c) Protein structure with a contact formed between sites 13 (Arg13, top) and 58 (Asp 58, bottom). The two contacting residues are highlighted in red and shown in the sphere representation.

We consider a pair of coevolved sites, 13 and 58, to demonstrate the mutation effect predicted by our method. The DLG4 protein structure from Rat is shown in Figure 2.3(c), where the interaction between the two corresponding residues is confirmed by their close proximity. A

covalent bond may form between a positively charged Arginine (R, top) and a negatively charged Aspartate (D, bottom), with their side chains pointing towards each other. Our method identifies strong mutation dependence between these two sites, with amino acid-wise dependencies shown in Figure 2.3(b). Strong connection between Lysine (K) at site 13 and Aspartate (D) at site 58, as well as Arginine (R) at site 13 and Glutamate (E) at site 58, is shown in dark orange connections, suggesting a clear pattern of charge compensation. In contrast, the co-occurrence of Aspartate (D) at site 13 and Glutamate (E) at site 58 in the sequence is unfavorable due to the same charges, as shown by the gray connection. Additionally, strong dependence is observed between Phenylalanine (F) and Aspartate (D) and between Alanine (A) and Arginine (R). While these pairs do not fit a clear property compensation pattern, they could be explained by higher-order dependencies involving multiple coevolved sites or false signals arising from misalignments in MSA data.

Chapter 3

Topic Modeling for sparse texts via high-dimensional zero inflated Poisson model with low-rank structure

3.1 Introduction

3.1.1 Motivation

As introduced in Chapter 1, traditional multinomial-based models such as LDA and pLSI face significant limitations to deal with the problem of topic analysis in sparse text corpora. Poisson-based topic models (Zhou et al., 2012; Gan et al., 2015; Airoldi and Bischof, 2016; Jiang et al., 2017; Henao et al., 2016; Carbonetto et al., 2022) offer a principled alternative by directly modeling word frequencies and allowing document length to arise naturally from the generative process. The Poisson framework further offers an advantage through zero-inflated Poisson (ZIP) models (Xu et al., 2015), which incorporate a zero-inflation component specifically designed to accommodate the abundant zeros.

To empirically evaluate the suitability of the ZIP framework, we perform a preliminary analysis on abstracts from the four top-tier statistics journals included in the MADStat, the AP corpus, and a subset of 2000 randomly selected words from the KOS Blog and Enron Email datasets, due to their large vocabulary sizes. For each word in the vocabulary, we conducted zero-inflated Poisson regression across documents (Lambert, 1992). Figure 3.1 shows the relationship between estimated zero-inflation probabilities and the logarithm of the estimated Poisson intensities. The scatterplots from all datasets except AP exhibit a clear and consistent relationship between the fitted zero-inflation probabilities and the logarithm of Poisson intensities. The lack of a discernible pattern in the AP dataset may be attributed to its pre-processing procedure, which involved removing the 50% least frequent words and the 5% shortest documents. This filtering likely eliminated informative low-frequency patterns that are essential for capturing the zero-inflation structure. These findings

provide strong empirical support for modeling the dependency between zero-inflation and intensity using a logit link function.

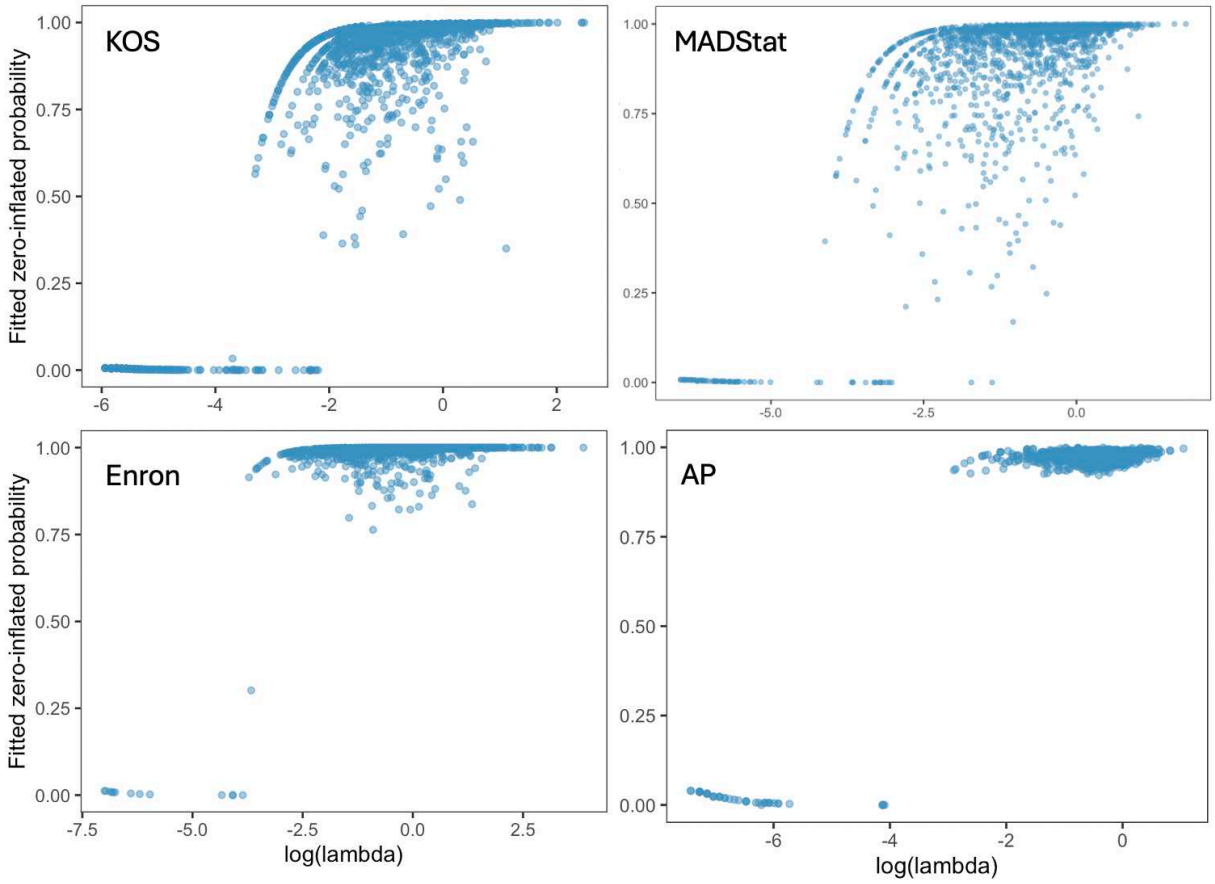


Figure 3.1: Estimated zero-inflation probabilities plotted against the logarithm of estimated Poisson intensities obtained from zero-inflated Poisson regression applied to four datasets.

3.1.2 Our contributions

We propose a zero-inflated Poisson (ZIP) topic model that explicitly models structural zeros through zero-inflation probabilities while allowing sampling zeros to arise naturally from the Poisson component. Guided by empirical evidence on the relationship between zero-inflation probabilities and Poisson intensities, we incorporate a logit link function to model this connection (Lambert, 1992; Xu et al., 2021). To further enhance model flexibility and better capture variability across documents, we introduce random effects into the Poisson intensity component. The full

model specification is provided in Section 3.2.2. This formulation enables more expressive modeling of real-world textual phenomena, including length-dependent sparsity, structural zeros, and thematic heterogeneity, which are particularly relevant in sparse or short-text corpora.

Similar to the pLSI model, we assume that the Poisson intensity matrix maintains a low-rank structure to preserve model interpretability. However, the incorporation of random effects renders traditional expectation-maximization (EM) algorithms (Lambert, 1992; Böhning et al., 1999; Xu et al., 2021) inapplicable. This departure from conventional topic modeling methods necessitates specialized optimization techniques. We develop an alternating optimization algorithm that leverages the low-rank structure of the intensity matrix. The procedure proceeds in two stages: we first estimate the Poisson intensity parameters $\boldsymbol{\lambda}$ via first-order moment matching, followed by maximum pseudo-likelihood estimation of the zero-inflation parameters α and β using iterative optimization. We then apply the vertex hunting algorithm of Ke and Wang (2024) to $\widehat{\boldsymbol{\lambda}}$ to recover the low-rank structure $\mathbf{A}$.

From an application perspective, we demonstrate the effectiveness of the proposed model through comprehensive simulation studies that examine performance across varying dimensions of (n, V, K), as well as different levels of sparsity and random effects. Comparisons with oracle settings and existing topic models reveal that our method achieves consistently superior performance. We further validate the model on the MADStat dataset (Ke et al., 2024), where our implementation uncovers 11 distinct topics that accurately capture the thematic structure of prominent statistics journals. The recovered temporal topic trends align closely with known developments in the field.

The remainder of this chapter is organized as follows. Section 3.2 provides background on the explicit form of our zero-inflated Poisson model. We then introduce two-step our estimation procedure with moment matching optimization for the Poisson intensity matrix and SCORE-based vertexing hunting for the topic-vocabulary matrix. In Section 3.2, we establish the relative ℓ_1 convergence rates of the proposed estimators. Section 3.4 and 3.5 evaluate our method through comprehensive simulation studies and application on the real-world corpus, both demonstrating the

superiority of the proposed method. Finally, Section 4.1 concludes with discussions and potential extensions.

The remainder of this chapter is organized as follows. Section 3.2 presents the formal specification of the proposed zero-inflated Poisson (ZIP) model and details the two-step estimation procedure, which combines moment-matching optimization for estimating the Poisson intensity matrix with a SCORE-based vertex hunting algorithm for recovering the topic-word matrix. Section 3.3 establishes theoretical guarantees for the proposed estimators, including relative ℓ_1 convergence rates for each column of the topic-word matrix. Sections 3.4 and 3.5 are dedicated to evaluating the proposed methodology through comprehensive simulation studies and an application to a real-world corpus, respectively. The simulation results demonstrate the superior performance of the method across a range of settings, while the real-data analysis reveals interpretable topic structures and meaningful temporal trends.

3.2 Methodology

3.2.1 Notations

Throughout the chapter, for an integer $K > 0$, we use $[K]$ to denote the set $\{1, 2, ..., K\}$, $I_n \in \mathbb{R}^{n\times n}$ to denote the identity matrix of size n and $\mathbf{1}_{n\times V}$ to denote the matrix with all entries equal to one. For a matrix $Q = (Q_{ij}) \in \mathbb{R}^{n\times n}$, $\mathrm{tr}(Q) = \sum_{i=1}^{n} Q_{ii}$ is the trace. For a vector u, $\|u\|$ is the Euclidean norm. For a matrix $A = (A_{ij}) \in \mathbb{R}^{K\times V}$ with $1 \le K \le V$ and the singular decomposition $A = \sum_{i=1}^{K} \sigma_i u_i v_i^\top$, $A_{\cdot j}$ is the j-th column of A and $A_{i\cdot}$ is the i-th row of A. $\|A\| = \max_{1\le i\le K} \sigma_i$, $\|A\|_* = \sum_{i=1}^{K} \sigma_i$, $\|A\|_F = \left(\sum_{i=1}^{k}\sum_{j=1}^{V} A_{ij}^2\right)^{1/2}$, $\|A\|_{2,\infty} = \max_{1\le i\le K} \|A_{i\cdot}\|$ and $\|\mathbf{A}\|_{1,1} = \sum_{i=1}^{K}\sum_{j=1}^{V} |A_{ij}|$ represent the spectral norm, the nuclear norm, the Frobenius norm, the two-to-infinity norm and the entry-wise ℓ_1 norm of A, respectively.

3.2.2 Model

Recall that we use $\mathbf{Z} = (Z_{ij}) \in \mathbb{R}^{n\times V}$ to denote the corpus matrix with n as number of documents and V as vocabulary size. Each entry Z_{ij} in the corpus matrix records the frequency of

word j in document i. We assume that each Z_{ij} follows a zero-inflated Poisson (ZIP) distribution, as motivated in Section 3.1.1.

$$Z_{ij} \sim \begin{cases} 0 & \text{with probability} \quad p_{ij} \\ \text{Poisson}(e_{ij} \cdot \lambda_{ij}) & \text{with probability} \quad 1 - p_{ij} \end{cases} \tag{3.1}$$

where $\mathbf{p} = (p_{ij}) \in \mathbb{R}^{n \times V}$ is the matrix of Bernoulli probabilities capturing structural zeros, with $0 \leq p_{ij} \leq 1$, and $\boldsymbol{\lambda} = (\lambda_{ij}) \in \mathbb{R}^{n \times V}$ is the matrix of Poisson intensities, with $\lambda_{ij} > 0$. p_{ij} and λ_{ij} jointly characterize both the occurrence and frequency of words across documents. Motivated by the positive association observed in Figure 3.1, we propose the logistic link with shape parameter α and location parameter β to associate p_{ij} and λ_{ij} by

$$\text{logit}(p_{ij}) = \alpha \log \lambda_{ij} + \beta, \tag{3.2}$$

where $\alpha \geq 0$. To account for additional variability not captured by λ_{ij}, we introduce a multiplicative random effect e_{ij} in the Poisson component, where $\mathbb{E}(e_{ij}) = 1$ and $\text{Var}(e_{ij}) < \infty$.

Consistent with the assumption in the probabilistic Latent Semantic Indexing (pLSI) model (Hofmann, 1999), we posit the existence of an unobserved number of latent topics $K \in \mathbb{Z}^+$ such that the Poisson intensity matrix $\boldsymbol{\lambda}$ admits a low-rank factorization of the form

$$\lambda_{ij} = \sum_{k=1}^{K} W_{ik} A_{kj}, \tag{3.3}$$

where W_{ik} are entries of the document-topic matrix $\mathbf{W} \in \mathbb{R}^{n \times K}$ and A_{kj} are entries of the topic-word matrix $\mathbf{A} \in \mathbb{R}^{K \times V}$ for all i, j and k. Each row $A_{k\cdot} := (A_{k1}, \ldots, A_{kV})^\intercal$ defines a probability mass function (PMF) over the vocabulary, and W_{ik} represents the contribution of topic k to document i.

To formalize the statistical structure of our zero-inflated Poisson topic model, we introduce several assumptions. These assumptions, stemming from both practical modeling considerations

and theoretical requirements, are designed to preserve interpretability, ensure identifiability, and maintain coherence with established topic modeling frameworks of pLSI discussed in Chapter 1.

Assumption 3 (Nonnegativity)**.** *All entries of the document-topic matrix* $\mathbf{W} \in \mathbb{R}^{n\times K}$ *and the topic-word matrix* $\mathbf{A} \in \mathbb{R}^{K\times V}$ *are nonnegative. That is,*

$$W_{ik} \geq 0 \quad \textit{for all } i \in [n],\, k \in [K], \quad \textit{and} \quad A_{kj} \geq 0 \quad \textit{for all } k \in [K],\, j \in [V].$$

In addition, the parameter α *in the linear model Equation* (3.2) *is assumed to satisfy* $\alpha \geq 0$*.*

Assumption 4 (Simplex Constraint)**.** *Each row of the topic-word matrix* $\mathbf{A} \in \mathbb{R}^{K\times V}$ *defines a probability distribution over the vocabulary. Specifically, for each topic* $k \in [K]$*, the entries of the* k*-th row satisfy*

$$\sum_{j=1}^{V} A_{kj} = 1. \tag{3.4}$$

It is important to note that the matrix factorization $\boldsymbol{\lambda} = \mathbf{WA}$ is inherently non-unique without additional structural constraints. In general, both $\mathbf{W}$ and $\mathbf{A}$ are only identifiable up to a nonsingular transformation, which poses challenges for interpreting the latent topics. To achieve identifiability of the topic-word matrix $\mathbf{A}$, we adopt the anchor word assumption, a widely used condition in topic modeling, which imposes a minimal and interpretable structural constraint that aligns with the separability condition introduced in Chapter 1.

Definition 1. *The word* j *is defined as an anchor word for topic* k *if it appears exclusively in that topic and has zero probability under all other topics. Formally,*

$$A_{kj} > 0 \quad \textit{and} \quad A_{k'j} = 0 \quad \textit{for all } k' \neq k. \tag{3.5}$$

Assumption 5 (Anchor Word)**.** *For each topic* $k \in [K]$*, there exists at least one anchor word* $j_k \in [V]$ *that appears exclusively in topic* k*.*

3.2.3 Estimation and Algorithm

As introduced in Chapter 1, topic modeling seeks to uncover latent semantic structures in textual data by recovering the underlying topic-word matrix $\mathbf{A}$ from the observed document-word count matrix $\mathbf{Z}$ with $K \ll \min(n, V)$. In the context of the proposed ZIP model with random effects, we now outline the model fitting procedure designed to estimate the topic-word matrix $\mathbf{A}$. Assuming the number of topics K is known, the estimation is carried out in two primary stages. First, the Poisson intensity matrix $\widehat{\boldsymbol{\lambda}}$ is estimated from the observed data using moment-based and likelihood-based techniques. Second, the topic-word matrix $\widehat{\mathbf{A}}$ is recovered from $\widehat{\boldsymbol{\lambda}}$ by leveraging its assumed low-rank structure.

Alternating estimation of $\boldsymbol{\lambda}$. We begin by jointly estimating the Poisson intensity matrix $\boldsymbol{\lambda}$ and the parameters α and β in the zero-inflation link function Equation (3.2), using an alternating optimization procedure. To simplify notation, we reparameterize β as $\eta := e^{\beta}$. The optimization proceeds iteratively by alternately updating $\boldsymbol{\lambda}$ and (α, η) until convergence.

At each iteration, we first fix (α, η) and update $\boldsymbol{\lambda}$. To mitigate the random effects, we leverage the expectation under the ZIP model defined in Equation (3.1), which satisfies

$$E(\mathbf{Z}) = \frac{\boldsymbol{\lambda}}{\mathbf{1}_{n\times V} + \boldsymbol{\lambda}^{\alpha}\eta}$$

where the operations represent the entry-wise calculation. Based on this, we estimate $\boldsymbol{\lambda}$ by minimizing the squared Frobenius norm between the observed document-word matrix $\mathbf{Z}$ and its model-based expectation. Specifically, we solve the following constrained optimization problem

$$\begin{aligned} \min_{\boldsymbol{\lambda}>0} \quad & L(\boldsymbol{\lambda}) = \frac{1}{2}\|\mathbf{Z} - \frac{\boldsymbol{\lambda}}{\mathbf{1}_{n\times V} + \boldsymbol{\lambda}^{\alpha}\eta}\|_F^2 \\ \text{s.t.} \quad & \boldsymbol{\lambda} \in \mathcal{C}_K = \{\boldsymbol{\lambda} : \text{rank}(\boldsymbol{\lambda}) \leq K\}. \end{aligned} \tag{3.6}$$

To solve the optimization problem Equation (3.6), we apply projected gradient descent. The entry-wise gradient of the objective function $L(\boldsymbol{\lambda})$ with respect to each entry of $\boldsymbol{\lambda}$ is given by

$$\nabla L(\boldsymbol{\lambda}) = \left(\frac{\boldsymbol{\lambda}}{\mathbf{1}_{n\times V} + \boldsymbol{\lambda}^{\alpha}\eta} - \mathbf{Z} \right) \circ \left(\frac{\mathbf{1}_{n\times V}}{\mathbf{1}_{n\times V} + \boldsymbol{\lambda}^{\alpha}\eta} - \frac{\alpha \boldsymbol{\lambda}^{\alpha}\eta}{(\mathbf{1}_{n\times V} + \boldsymbol{\lambda}^{\alpha}\eta)^2} \right), \tag{3.7}$$

where $\circ$ represents the entry-wise production in matrices. To enforce the low-rank constraint and nonnegativity, we apply singular value projection (SVP) (Jain et al., 2010) at each iteration by truncating the singular value decomposition of the updated matrix to rank K and further threshold any negative entries to a small positive constant (e.g., 10^{-5}).

With the updated estimate of $\boldsymbol{\lambda}$ fixed, we proceed to estimate the parameters (α, η) by maximizing a pseudo-likelihood. Due to the presence of unobserved random effects, the exact likelihood function of the observed data is not directly tractable. To address this, we adopt a Poisson-based pseudo-likelihood approach, which, while not modeling the full data-generating process, preserves the correct expectation of the model. This approximation facilitates a tractable approach to estimating the parameters (α, η) by maximizing a pseudo-likelihood that correctly captures the first moment of the ZIP model under a Poisson approximation.

$$\begin{aligned} (\hat{\alpha}, \hat{\eta}) &= \arg\max_{\alpha,\eta>0} \ell(\alpha, \eta | \mathbf{Z}, \boldsymbol{\lambda}) \\ &= \arg\max_{\alpha,\eta>0} \sum_{i,j} Z_{ij} \log\left(\frac{\lambda_{ij}}{1 + \lambda_{ij}^{\alpha}\eta} \right) - \frac{\lambda_{ij}}{1 + \lambda_{ij}^{\alpha}\eta} \end{aligned} \tag{3.8}$$

The optimization in Equation (3.8) can be efficiently performed using the `optim` function in `R`, which provides a flexible framework for nonlinear optimization. The complete alternating estimation procedure for jointly estimating the Poisson intensity matrix $\boldsymbol{\lambda}$ and the zero-inflation parameters (α, η) is summarized in Algorithm 2.

Estimation of $\mathbf{A}$ *and* $\mathbf{W}$. Once $\widehat{\boldsymbol{\lambda}}$ is estimated via the alternating optimization procedure, we proceed to recover the topic-word matrix $\mathbf{A}$ using the SCORE method with vertex hunting, implemented via the Sketched Vertex Search (SVS) algorithm introduced by Jin et al. (2024). Specifically, we begin by normalizing the estimated Poisson intensity matrix $\widehat{\boldsymbol{\lambda}}$ by dividing each

Algorithm 2: Alternating Estimation of $\boldsymbol{\lambda}$, α, and η

Require: Corpus matrix $\mathbf{Z} \in \mathbb{R}^{n \times V}$, number of topics K, step size t

Initialization iteration index $r \leftarrow 1$, $\boldsymbol{\lambda}^{(0)} \leftarrow \mathbf{0}$, $\alpha^{(0)} \leftarrow 0.5$, $\eta^{(0)} \leftarrow 0$

while not converged **do**

Update $\boldsymbol{\lambda}$ with fixed $\alpha^{(r-1)}$, $\eta^{(r-1)}$

repeat

Compute gradient $\nabla L(\boldsymbol{\lambda}^{(r-1)})$ using Equation (3.7)

Perform gradient step: $\tilde{\boldsymbol{\lambda}}^{(r)} \leftarrow \boldsymbol{\lambda}^{(r-1)} - t \cdot \nabla L(\boldsymbol{\lambda}^{(r-1)})$

Compute rank-K SVD: $\tilde{\boldsymbol{\lambda}}^{(r)} = UDV^\top$

Project to nonnegative space: $\boldsymbol{\lambda}^{(r)} \leftarrow \max(UDV^\top, 10^{-5})$

until $\boldsymbol{\lambda}$ converges

Update $\alpha^{(r)}$, $\eta^{(r)}$ with fixed $\boldsymbol{\lambda}^{(r)}$

Solve Equation (3.8) using `optim` on (α, η) with fixed $\boldsymbol{\lambda}^{(r)}$

Increment $r \leftarrow r + 1$

end while

return Final estimates: $\boldsymbol{\lambda}^{(r)}$, $\alpha^{(r)}$, $\eta^{(r)}$

row by its row sum to mitigate the effects of row-wise heterogeneity in $\widehat{\boldsymbol{\lambda}}$, thereby reducing the influence of differing total word frequencies across documents on the estimation of topic proportions. We then perform spectral decomposition on the normalized matrix to extract its right singular vectors, which provide a low-dimensional embedding of the vocabulary. These embeddings define a point cloud in a reduced-dimensional space, from which the Sketched Vertex Search (SVS) algorithm identifies K vertices corresponding to anchor words via a K-means clustering step. The coordinates of these anchor words are subsequently used to reconstruct the right singular vector matrix, yielding the final estimate $\widehat{\mathbf{A}}$ of the topic-word matrix.

As a byproduct, each row of the document-topic matrix $\mathbf{W}$ is estimated via a weighted least-squares procedure, where the weights are derived from the normalization factors used in the pre-processing step of the vertex hunting algorithm. Any negative entries in $\mathbf{W}$ are set to zero to ensure nonnegativity.

3.3 Theoretical Guarantee

In this section, we present the main theoretical results demonstrating that the proposed projected gradient algorithm consistently recovers the low-rank Poisson intensity matrix $\boldsymbol{\lambda}$, and that the subsequent SCORE-based procedure yields a provable error bound for the estimation of the topic-word matrix $\mathbf{A}$.

Although document length is not parametrized explicitly, we have the length of each document i as a random variable with expectation $\mathbb{E}(N_i) = \sum_{j=1}^{V}(1 - p_{ij})\lambda_{ij}$. We denote the corpus-wide average of these expected lengths by N.

To analyze the theoretical properties of $\widehat{\boldsymbol{\lambda}}$, we begin by recalling that the projected gradient descent procedure used for its estimation

$$\widehat{\boldsymbol{\lambda}}^{(r+1)} = \mathcal{P}_{\mathcal{C}_K}\left(\tilde{\boldsymbol{\lambda}}^{(r+1)}\right) = \mathcal{P}_{\mathcal{C}_K}\left[\widehat{\boldsymbol{\lambda}}^{(r)} - t\nabla\psi(\boldsymbol{\lambda}^{(r)}) \circ \left(\psi(\boldsymbol{\lambda}^{(r)}) - \mathbf{Z}\right)\right],$$

where $\mathcal{P}_{\mathcal{C}_K}$ denotes the projection onto the rank-K constraint set $\mathcal{C}_K$, $\psi : \mathbb{R}^{n\times V} \to \mathbb{R}^{n\times V}$ is a nonlinear, entry-wise mapping as $\psi(\boldsymbol{\lambda}) = \frac{\boldsymbol{\lambda}}{1+\boldsymbol{\lambda}^{\alpha}\eta}$, $\nabla\psi(\boldsymbol{\lambda})$ denotes the corresponding entry-wise derivative and $\circ$ denotes the Hadamard product.

We adopt standard assumptions from the optimization literature (e.g., Chen and Wainwright, 2015; Barber and Ha, 2018), beginning with the following definitions on the objective function .

Definition 2 (Restricted strong convexity and restricted smoothness)**.** *For curvature parameters c_a,c_b and statistical tolerance $\epsilon_{\mathcal{L}} \geq 0$, we say the objective function $\mathcal{L}$ satisfies a restricted strong convexity (RSC) and restricted smoothness (RSM) over $\mathcal{C}$ if for any $\boldsymbol{\lambda}_1, \boldsymbol{\lambda}_2 \in \mathcal{C}$,*

$$\mathcal{L}(\boldsymbol{\lambda}_1) \geq \mathcal{L}(\boldsymbol{\lambda}_2) + < \nabla\mathcal{L}(\boldsymbol{\lambda}_2), \boldsymbol{\lambda}_1 - \boldsymbol{\lambda}_2 > + \frac{c_a}{2}\|\boldsymbol{\lambda}_1 - \boldsymbol{\lambda}_2\|_F^2 - \frac{c_a}{2}\epsilon_{\mathcal{L}}^2 \tag{3.9}$$

$$\mathcal{L}(\boldsymbol{\lambda}_1) \leq \mathcal{L}(\boldsymbol{\lambda}_2) + < \nabla\mathcal{L}(\boldsymbol{\lambda}_2), \boldsymbol{\lambda}_1 - \boldsymbol{\lambda}_2 > + \frac{c_b}{2}\|\boldsymbol{\lambda}_1 - \boldsymbol{\lambda}_2\|_F^2 - \frac{c_a}{2}\epsilon_{\mathcal{L}}^2 \tag{3.10}$$

As is common in the low-rank factorized optimization literature, we conduct our analysis within a local neighborhood of the target $\boldsymbol{\lambda}^*$ by assuming that the initialization lies within a

radius ρ of $\boldsymbol{\lambda}^*$. This local assumption permits the required regularity conditions on the objective function $\mathcal{L}$ to be imposed only in a neighborhood around $\boldsymbol{\lambda}^*$. The term $\epsilon_{\mathcal{L}}$ is introduced to quantify a vanishingly small level of statistical error, which, in the high-dimensional statistics literature, characterizes the best attainable accuracy in recovering the underlying parameter. Specifically, $\boldsymbol{\lambda}^*$ may denote a global minimizer that lies within an $\epsilon_{\mathcal{L}}$-neighborhood of the true parameter. Consequently, convergence to $\boldsymbol{\lambda}^*$ within an error of order $\epsilon_{\mathcal{L}}$ implies that the estimated solution achieves near-optimal accuracy in approximating the true parameter.

The following lemma establishes that the objective function satisfies the necessary curvature and smoothness properties within a local neighborhood, and that the projection operator used in the algorithm behaves in a geometrically stable manner.

Lemma 3.3.1. *The objective function $\mathcal{L}(\boldsymbol{\lambda})$ defined via the projected gradient updates satisfies the restricted strong convexity and restricted smoothness in Equation* (3.9) *and Equation* (3.10)*, respectively, with statistical tolerance $\epsilon_{\mathcal{L}} \lesssim \sqrt{\frac{(n+V)\log(n+V)}{NnV}}$. Furthermore, the projection operator $\mathcal{P}_{\mathcal{C}_K}$ that retains the top-K singular vectors naturally satisfies $\|\boldsymbol{\lambda}_1 - \mathcal{P}_{\mathcal{C}_K}(\boldsymbol{\lambda}_1)\|_F \leq \min_{\boldsymbol{\lambda}\in\mathcal{C}_k} \|\boldsymbol{\lambda}_1 - \boldsymbol{\lambda}\|_F$.*

We now establish the convergence behavior of the projected gradient algorithm under the local geometric and statistical conditions outlined above. To guarantee that the sub-exponential tail bounds required for the analysis hold, we assume the random effects are bounded, thereby ensuring the applicability of concentration inequalities used in our theoretical guarantees.

Theorem 3.3.2 (Convergence of projected gradient algorithm)**.** *With the constraint set $\mathcal{C}_K$ and $\boldsymbol{\lambda}^* \in \mathcal{C}_K$ defined above, suppose Lemma 3.3.1 holds, for any initialization $\boldsymbol{\lambda}^0$ with $d(\boldsymbol{\lambda}^0, \boldsymbol{\lambda}^*) \leq \sigma_1(\boldsymbol{\lambda}^*)$, where $\sigma_1(\boldsymbol{\lambda}^*)$ is the largest singular value, the projected gradient iterates $\{\boldsymbol{\lambda}^t\}_{t=1}^{\infty}$ satisfies*

$$d(\widehat{\boldsymbol{\lambda}}^{(r+1)}, \boldsymbol{\lambda}^*)^2 \leq \left(1 - \frac{2\log(n+V)}{\eta\alpha nV}\right)^r d(\widehat{\boldsymbol{\lambda}}^{(0)}, \boldsymbol{\lambda}^*)^2 \frac{c_\alpha K(n+V)\log(n+V)}{NnV}, \tag{3.11}$$

where $d(\widehat{\boldsymbol{\lambda}}^{(r+1)}, \boldsymbol{\lambda}^*)^2 = \frac{1}{nV}\|\boldsymbol{\lambda}^{(r+1)} - \boldsymbol{\lambda}^*\|_F^2$, $c_\alpha = 1$ when $0 < \alpha < 1$ or $c_\alpha = \max\{1, \frac{(\alpha-1)^2}{4\alpha}\}$ otherwise.

Wheen turning to the second step of the procedure, where vertex hunting is applied to $\widehat{\boldsymbol{\lambda}}$ to recover the topic-word matrix $\mathbf{A}$, let $D_{\boldsymbol{\lambda}}$ be the diagonal matrix of average word intensities, with $(D_{\boldsymbol{\lambda}})_{jj} = \frac{1}{n}\sum_{i=1}^{n}\lambda_{ij}$, vertex hunting is performed on the normalized matrix $\boldsymbol{\lambda}^*(D_{\boldsymbol{\lambda}^*})^{-1/2}$. To ensure stability of the right singular vectors, we impose the following assumptions.

Assumption 6 ($\mathbf{A}$ and $\mathbf{W}$ are well-conditioned)**.** *For some constant $c \in (0,1)$ and $h > 0$,*

$$\sigma_K(\mathbf{A}) \geq c\sqrt{K},\ \sigma_K(\frac{\mathbf{W}^{\mathrm{T}}\mathbf{W}}{n}) \geq h.$$

Assumption 7 (Regular topic-topic correlation)**.** *The entries of $\mathbf{A}\mathbf{A}^{\mathrm{T}}$ satisfy the following for some constant $c > 0$:*

$$\min_{1\leq k,l\leq K}(\mathbf{A}\mathbf{A}^{\mathrm{T}})_{kl} \geq c.$$

Both assumptions ensure that topic vectors in $\mathbf{A}$ are not too correlated and no single word dominates the embedding space, promoting the stability of vertex estimation and ultimately of $\widehat{\mathbf{A}}$.

We establish an entry-wise error bound for the estimated topic-word matrix obtained via the SCORE-based procedure applied to $\widehat{\boldsymbol{\lambda}}$ after normalization.

Theorem 3.3.3 (Error bound for topic-word matrix estimation)**.** *Suppose Assumptions 3–7 hold. Then, as $r \to \infty$, with probability $1 - o(n^{-1})$, the estimated topic-word matrix satisfies:*

$$\|A_{\cdot j} - \hat{A}_{\cdot j}\|_1 \lesssim c_\alpha \|A_{\cdot j}\|_1 \sqrt{\frac{V\log(n+V)}{Nn}}.$$

Therefore, with probability $1 - o(n^{-1})$, we also have

$$\sum_{j=1}^{V}\|A_{\cdot j} - \hat{A}_{\cdot j}\|_1 \lesssim c_\alpha K\sqrt{\frac{V\log(n+V)}{Nn}}.$$

3.4 Simulations

3.4.1 Data generation mechanism

In this section, we assess the numerical performance of our proposed method through a series of simulation studies conducted under various controlled settings. To construct the topic-to-vocabulary matrix $\mathbf{A} \in \mathbb{R}^{K \times V}$, we assume that each topic contains exactly one anchor word (i.e., $n_{\text{anchor}} = 1$). For topic k, we assign a small baseline loading to its corresponding anchor word, setting $A_{kj} = 0.001$ for the anchor word j associated with topic k. For the non-anchor words, we follow the simulation framework introduced in Tran et al. (2023), where word frequencies are designed to follow Zipf's law (Zipf, 1999). This results in a distribution of word loadings that mimics empirical patterns commonly observed in real-world text corpora, as illustrated in Figure 1.2(a). Specifically, for each topic k, the loading of the jth most frequent non-anchor word, denoted $A_{k(j)}$, is generated according to

$$A_{k(j)} \propto \frac{1}{(j + b_{\text{zipf}})^{a_{\text{zipf}}}}$$

where $a_{\text{zipf}} = 1$, $b_{\text{zipf}} = 2.7$. Each row of the topic-word matrix $\mathbf{A}$ is normalized to ensure that its entries sum to one, thereby defining valid probability mass functions over the vocabulary. The document-topic matrix $\mathbf{W} \in \mathbb{R}^{n \times K}$ is generated by independently sampling each row from a scaled Dirichlet distribution, $\tau, \text{Dir}(\mathbf{1}_K)$, where $\tau = 500$ serves as a concentration parameter selected to yield an average document length of approximately 300.

To introduce moderate variability around the structured Poisson intensity matrix, we generate a random effect matrix $\mathbf{e} \in \mathbb{R}^{n \times V}$ with entries independently drawn from the uniform distribution $\text{Unif}(0, 2)$. In Section 3.4.2, we also consider oracle scenarios where the random effects are not entirely latent, and partial prior information is available.

Given $\mathbf{A}$ and $\mathbf{W}$, we compute the Poisson intensity matrix as $\boldsymbol{\lambda} = \mathbf{WA}$, and then derive the zero-inflation probability matrix $\mathbf{p}$ via the logistic link function in Equation (3.2), parameterized by (α, η). We investigate a range of (α, η) values across experimental settings, as detailed in the subsequent subsections, chosen to ensure that the average of the resulting zero-inflation

probabilities remains approximately constant as the vocabulary size V increases. With $(\mathbf{p}, \boldsymbol{\lambda}, \mathbf{e})$ specified, the observed corpus matrix $\mathbf{Z}$ is generated according to the zero-inflated Poisson (ZIP) model defined in Equation (3.1).

We evaluate the numerical performance of our method across a range of document counts $n \in \{2000, 4000, 6000, 8000, 10000\}$ and vocabulary sizes $V \in \{2000, 4000, 6000, 8000\}$, where the vocabulary sizes are chosen to reflect the typical number of commonly used words in real-world corpora. In addition, we assess the estimators under different numbers of topics, with $K \in \{3, 5, 10\}$. For each setting, we assume that the true number of topics K is known and report the average relative estimation errors over 100 independent repetitions.

The model's performance is evaluated in Section 3.4.2 using the logarithm of the relative ℓ_1 estimation error of the topic-word matrix $\mathbf{A}$, denoted as $\log(L_1(\mathbf{A} - \widehat{\mathbf{A}}))$, defined by $\log(L_1(\mathbf{A} - \widehat{\mathbf{A}}))$, which is the same measurement in Tran et al. (2023); Ke and Wang (2024), Since $\mathbf{A}$ is identifiable only up to a permutation of its rows, we evaluate the RE by considering all possible row permutations of the estimated matrix. Let $S([K])$ denote the permutation group on $[K]$, and let $\widehat{\mathbf{A}}_{\mathbf{r}}$ represent the matrix obtained by permuting the rows of $\widehat{\mathbf{A}}$ according to $\mathbf{r} \in S([K])$. The relative ℓ_1 estimation error $L_1(\mathbf{A} - \widehat{\mathbf{A}})$ is then defined as

$$L_1(\mathbf{A} - \widehat{\mathbf{A}}) = \min_{\mathbf{r} \in S([K])} \frac{\|\mathbf{A} - \widehat{\mathbf{A}}_{\mathbf{r}}\|_{1,1}}{K}.$$

3.4.2 Comparison with oracle scenarios across varying number of documents and vocabulary size

In this section, we evaluate our model's performance across a range of (n, V) with $K = 3$ and $(\alpha, \eta) = (1, \exp(V/2000))$ and compare it with two oracle scenarios (S1) and (S2) that assume different levels of knowledge about the random effect.

(S1) We assume the distribution from which each e_{ij} is drawn is known, and aim to integrate out e_{ij} from the ZIP model to facilitate estimation on the model parameters. Since $E(Z_{ij})$ remains the same, the moment matching step for estimating $\boldsymbol{\lambda}$ in Equation (3.6) still applies.

Moreover, knowing the random effect distribution yields an explicit likelihood function, allowing us to apply maximum likelihood estimation (MLE) similar to (S1) to estimate α and η. Assuming $e_{ij} \sim \text{Unif}(0, 2)$, the marginal distribution of Z_{ij} is also a zero inflated distribution

$$Z_{ij} \sim \begin{cases} 0 & \text{with probability} \quad p_{ij} \\ f(Z_{ij}; \lambda_{ij}) = \frac{\gamma(Z_{ij}+1, 2\lambda_{ij})}{2\lambda_{ij} Z_{ij}!} & \text{with probability} \quad 1 - p_{ij} \end{cases} \tag{3.12}$$

where $\gamma(s, x)$ is the lower incomplete gamma function defined as $\gamma(s, x) = \int_0^x t^{s-1} e^{-t} dt$.

(S2) We assume the exact values of the error terms e_{ij} are known. Therefore the conditional expectation simplifies to $E(Z_{ij}|e_{ij}) = (1 - p_{ij}) e_{ij} \lambda_{ij}$. Consequently, the moment-matching criterion for estimating $\boldsymbol{\lambda}$ becomes

$$\min_{\boldsymbol{\lambda}} \frac{1}{2} \|\mathbf{Z} - \frac{\mathbf{e}\boldsymbol{\lambda}}{1 + \boldsymbol{\lambda}^{\alpha} \eta}\|_F^2.$$

Moreover, having exact error terms allows us to leverage the full likelihood, rather than a pseudo-likelihood. Specifically, the probability mass function becomes

$$\begin{aligned} \mathbb{P}(Z_{ij} = 0) &= p_{ij} + (1 - p_{ij}) e^{-e_{ij}\lambda_{ij}} \\ \mathbb{P}(Z_{ij} = z) &= (1 - p_{ij}) \frac{e^{-e_{ij}\lambda_{ij}} (e_{ij}\lambda_{ij})^{Z_{ij}}}{Z_{ij}!}, \quad z > 0. \end{aligned}$$

Thus, to estimate α and η, we maximize the following log-likelihood

$$\max_{\alpha, \eta > 0} \sum_{ij} \left[\mathbb{I}\{Z_{ij} = 0\} \log \left(\mathbb{P}(Z_{ij}) \right) + \mathbb{I}\{Z_{ij} \neq 0\} \log \left(\mathbb{P}(Z_{ij}) \right) \right]$$

Figure 3.2 demonstrates that the REs of $\mathbf{A}$ in our model decrease as the number of documents n grows, but increase with larger vocabulary sizes V. When examining $\log(L_1(\mathbf{A} - \widehat{\mathbf{A}}))$, VK corresponds to the number of parameters to be estimated, whereas n represents the sample

size. Consequently, the observed trends from our model indicate that increasing the sample size improves estimation accuracy of $\mathbf{A}$, while a higher number of parameters makes the estimation problem more challenging.

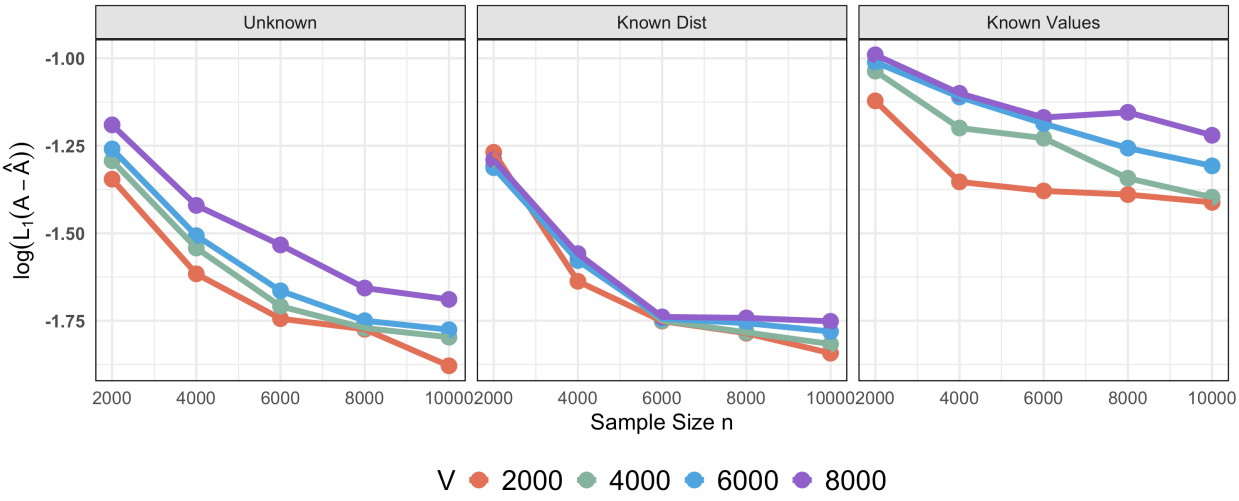


Figure 3.2: $\log(L_1(\mathbf{A}-\widehat{\mathbf{A}}))$ with $K = 3$: Comparison among three scenarios—(i) e unknown (our model), (ii)known distribution with $\mathrm{Unif}(0, 2)$ in (S1), and (iii) known values in (S2). Each scenario is labeled accordingly in the plot.

When compared to the oracle scenarios, our model exhibits only minor discrepancies in the estimation of $\mathbf{A}$, indicating that the variability introduced by the random effect is largely absorbed at the topic-to-vocabulary level. Interestingly, the scenario in which the random effects are fully observed performs worse than the other two, potentially due to overfitting to the observed noise. These results underscore the robustness of our estimator under increasing data dimensionality and across varying assumptions regarding latent noise.

3.4.3 Connections and comparisons with pLSI

In this section, we present a comprehensive evaluation of our proposed model through systematic comparisons with three existing approaches: Topic-SCORE (Ke and Wang, 2024), the Sparse Topic Model solver (STM-TOP) (Bing et al., 2020b), and the thresholded Topic-SCORE (TTS) (Tran et al., 2023), where the latter two are pLSI-based models specifically

designed to handle sparse data. For TTS, we follow the authors' recommendations and set the threshold parameter to 10, which ensures that approximately 10–40% of the vocabulary is removed. The design of our model incorporates three essential components: (i) random effects accounting for unobserved heterogeneity, (ii) zero inflation to address excess zeros in the data accounting for the high sparsity in the data, and (iii) a functional relationship linking the zero-inflation probability to the Poisson intensity accounting for informative structural zeros patterns. To rigorously evaluate the individual and joint contributions of these components, we consider all possible combinations of their inclusion, resulting in a set of distinct experimental settings. Table 3.1 outlines the experimental settings and summarizes the performance of our method relative to existing approaches under each setting. Detailed simulation results are presented in the following subsections.

Table 3.1: Experimental settings and comparative performance of our method relative to existing approaches.

Highly Sparse	Heterogeneity	Informative Zeros	Performance Summary
✗	✗	–	Our model, Topic-SCORE, and TTS exhibit comparable performance, while all three outperform STM-TOP.
✗	✓	–	
✓	✗	✓	Our model and Topic-SCORE perform similarly under mild zero inflation, while the other approaches already exhibit failure and instability. Under strong zero inflation, our model outperforms Topic-SCORE.
✓	✓	✓	
✓	✗	✗	When the sample size is small, our model outperforms Topic-SCORE. Moreover, both methods consistently outperform the remaining baseline approaches.
✓	✓	✗	All the methods fail to estimate.

Settings without zero inflation

We begin by evaluating the performance of our method under data-generating settings that do not include zero inflation with $(\alpha, \eta) = (0, 0)$. As shown in Figure 3.3, in both the presence of random effects and their absence, our method, Topic-SCORE, and TTS yield consistent estimation results, all of which outperform STM-TOP. The similarity in performance between our method and Topic-SCORE is consistent with the theoretical equivalence between the Poisson and multinomial models (Carbonetto et al., 2022) which proves in Section B.1. While TTS is designed to enhance the identifiability by removing low-frequency words, we observe that it performs slightly worse than our method and Topic-SCORE. This may be due to the unintended removal of informative rare words, which can reduce the resolution of $\mathbf{A}$ and impact estimation accuracy. Moreover, for our model, Topic-SCORE and TTS, we observe a clear trend in which the estimation error increases with the vocabulary size V, reflecting the greater difficulty of accurate estimation as the number of parameters grows. When random effects are present, the estimation error tends to be higher at smaller values of V, likely due to the increased variability introduced by the unobserved heterogeneity. However, as V becomes large, the impact of random effects diminishes, resulting in similar performance with or without random effects. In contrast, STM-TOP exhibits instability as n and V increases. The presence of random effects does not exhibit a consistent or interpretable pattern on STM-TOP.

Settings with zero-inflation

We examine the complex scenario, where zero-inflation probabilities are linked to Poisson intensity via a logit function defined in Equation (3.2). Since the parameters (α, η) directly control the magnitude of zero inflation, we consider both mild and strong levels of zero-inflation probability in the following experiments. Due to instability in estimation, STM-TOP is excluded from the following comparisons. Furthermore, because bounded random effects with light-tailed distributions introduce only limited distortion to the estimation process, we include random effects in all scenarios to consistently account for document-level heterogeneity.

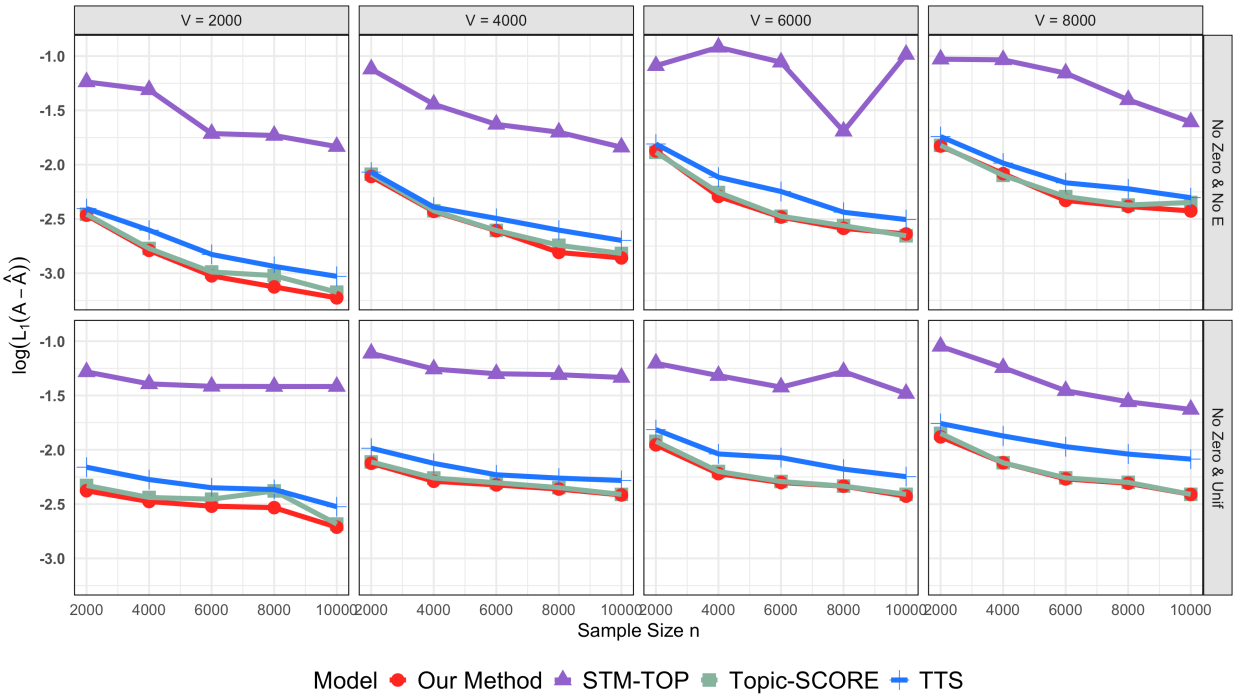


Figure 3.3: $\log(L_1(\mathbf{A} - \widehat{\mathbf{A}}))$ under zero-inflation-free settings with $K = 3$. Top: (a) No random effects ($e_{ij} = 1$). Bottom: (b) Random effects drawn from $\mathrm{Unif}(0, 2)$.

Mild zero inflation. We consider a setting with $(\alpha, \eta) = (1, \exp(V/4000))$ with random effects, which increases the average zero-inflation probability from approximately 0.25 to 0.27 as V ranges from 2000 to 8000. Correspondingly, the proportion of observed zeros in the data ranges from 86.5% to 96.5%. Under this setting, we further evaluate model performance across different numbers of topics with $K \in \{3, 5, 10\}$. As shown in Figure 3.4, our method substantially outperforms TTS, which fails to produce reliable estimates when $V > 2000$ with relative estimation errors exceeding 1. Our model also consistently outperforms Topic-SCORE across all settings, with particularly pronounced advantages when K is large and n is small. The estimation error decreases as n increases, demonstrating consistency. In contrast, as either V or K increases, the estimation error tends to rise due to the growing number of parameters to be estimated. Although our theoretical guarantees in Section 3.3 are established under the assumption of bounded random effects, we further investigate the impact of alternative random effect distributions in Section B.2.

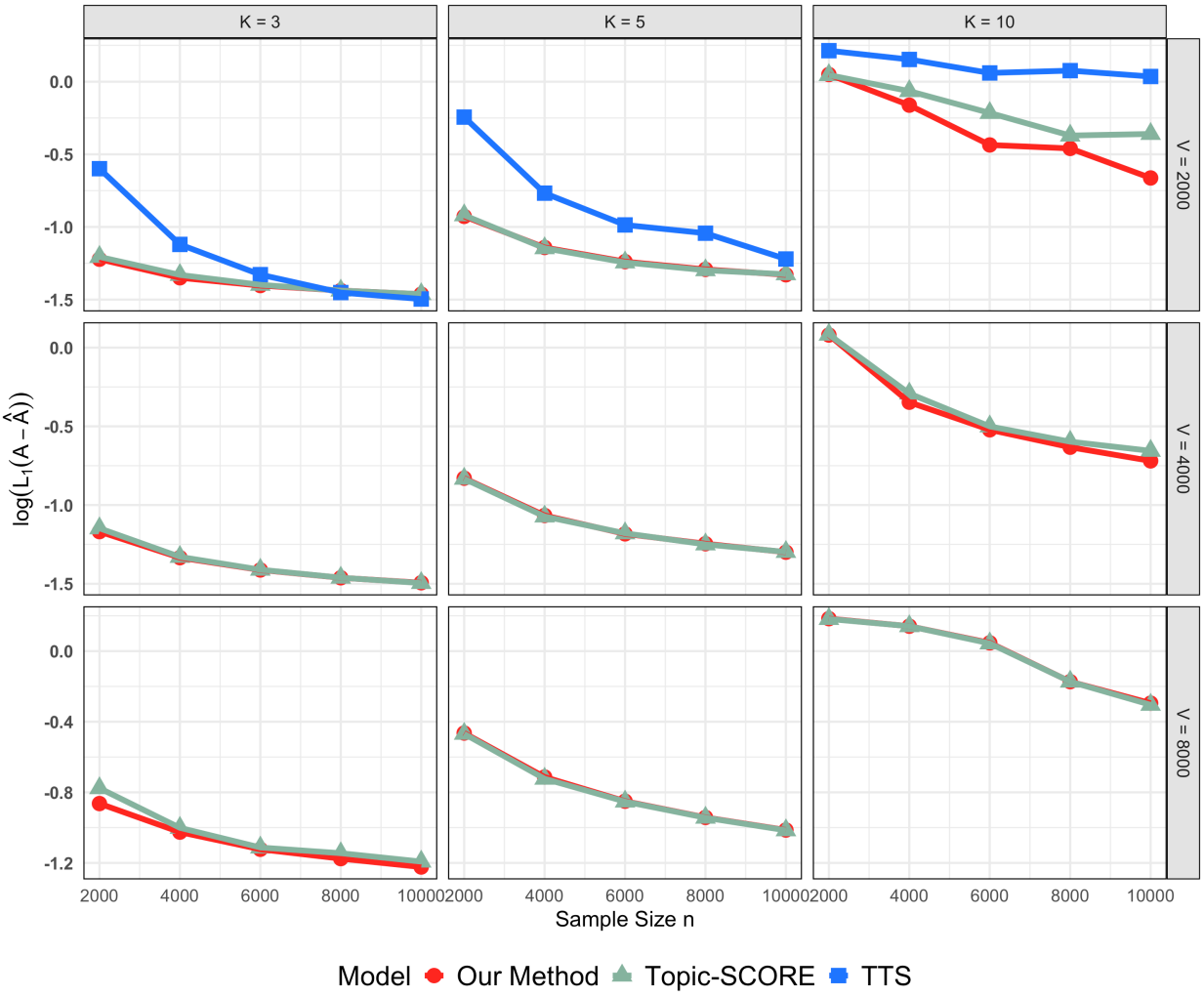


Figure 3.4: $\log(L_1(\mathbf{A} - \widehat{\mathbf{A}}))$ with random effects, evaluated across varying (n, V, K), comparing our method with Topic-SCORE and TTS.

Strong zero inflation. We next consider a setting characterized by strong zero inflation, with parameters $(\alpha, \eta) = (0.5, \exp(V/2000))$. Under this specification, the average proportion of observed zeros increases from approximately 92.2% to 99.6% as the vocabulary size V ranges from 2000 to 6000, aligning closely with the sparsity observed in real-world datasets. The results, summarized in Figure 3.5, show that our model benefits from increasing sample size, thereby confirming its consistency. Moreover, it maintains stable performance across varying values of V and K, while capturing the increasing estimation challenge posed by larger vocabulary sizes and a greater number of topics. Overall, our method consistently outperforms both Topic-

SCORE and TTS in terms of estimation accuracy across most settings, with especially pronounced improvements observed when the vocabulary size and the number of topics are small.

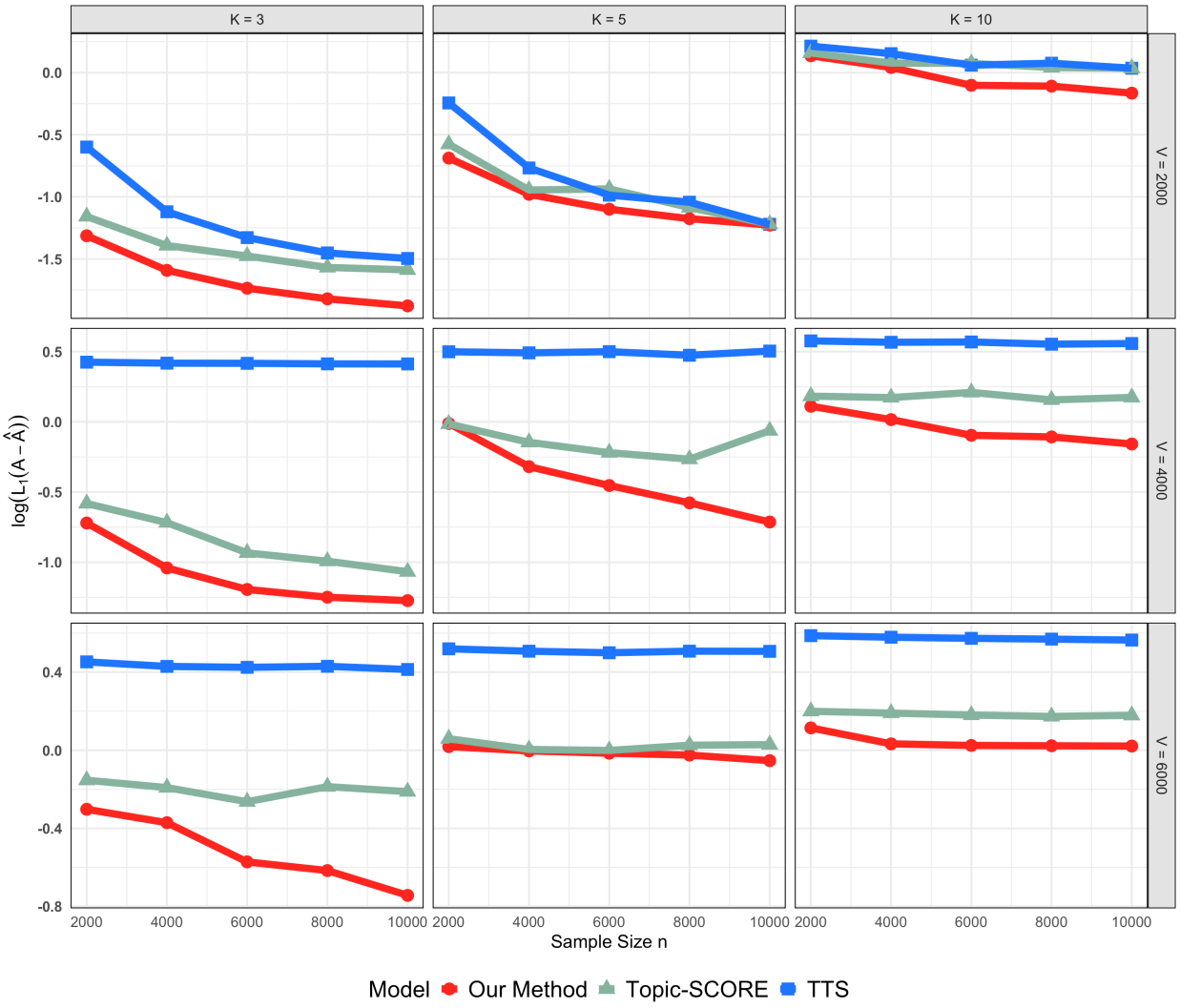


Figure 3.5: $\log(L_1(\mathbf{A} - \widehat{\mathbf{A}}))$ from our model, Topics-SCORE and TTS with strong zero inflation and random effects generated from $\mathrm{Unif}(0, 2)$, evaluated across varying (n, V, K)

Our method demonstrates strong performance across diverse parameter regimes, particularly excelling in strong zero inflation containing informative structural zeros with large V. Notably, it maintains comparable performance even when our zero inflation assumption is relaxed. This versatility makes our approach well-suited for most real-world text datasets, which typically exhibit sparsity and Zipf's law decay. While alternative methods like Topic-SCORE and TTS may remain

competitive in certain scenarios, especially those with minimal zero inflation, our method offers broader applicability across common text analysis challenges.

Furthermore, we consider the presence of structural zeros and evaluate settings without the functional link defined in Equation (3.2), examining both the presence and absence of random effects. In this case, we set $(\alpha, \eta) = (0, \exp(0.5))$ which results in zero proportion remain constant of 0.92 across documents and vocabulary. This setting corresponds to a missing-completely-at-random (MCAR) mechanism, in which structural zeros arise independently of the underlying topic intensities. Given that previous results consistently show that LDA and STM-TOP yield higher estimation errors than the other three models, we omit them from the comparisons in all subsequent analyses.

Figure 3.6 shows that all methods exhibit decreasing estimation error as n increases, reflecting consistency in the absence of random effects. Our method performs better at smaller n, while Topic-SCORE slightly outperforms our method across all vocabulary sizes when n is large. Our method also consistently outperforms TTS. These trends suggest that the convergence rate of our method is slower compared to Topic-SCORE and TTS under this setting. Nevertheless, the overall differences in estimation error between the three methods remain modest. This outcome suggests that, under a MCAR mechanism where the zero-inflation pattern is independent of the underlying structure, estimation becomes relatively easier due to the absence of informative zeros, and our method is able to provide competitively reliable results.

To further illustrate the effect of informative zeros on estimation, we vary the value of α under the setting $n = V = 2000$ and $K = 3$. The results in Figure 3.7 show that, even as the overall zero proportion decreases, estimation becomes more difficult when zeros are more informative—that is, when α is larger. This trend is understandable for Topic-SCORE, which is known to be limited in handling sparse data and lacks a mechanism to account for informative zeros. However, for our model, although the zero-inflation mechanism is correctly specified, we still observe that estimation accuracy deteriorates as α increases. This is because high-intensity entries in $\boldsymbol{\lambda}$, which are the most informative for vertex hunting, are increasingly masked as zeros. While the zero

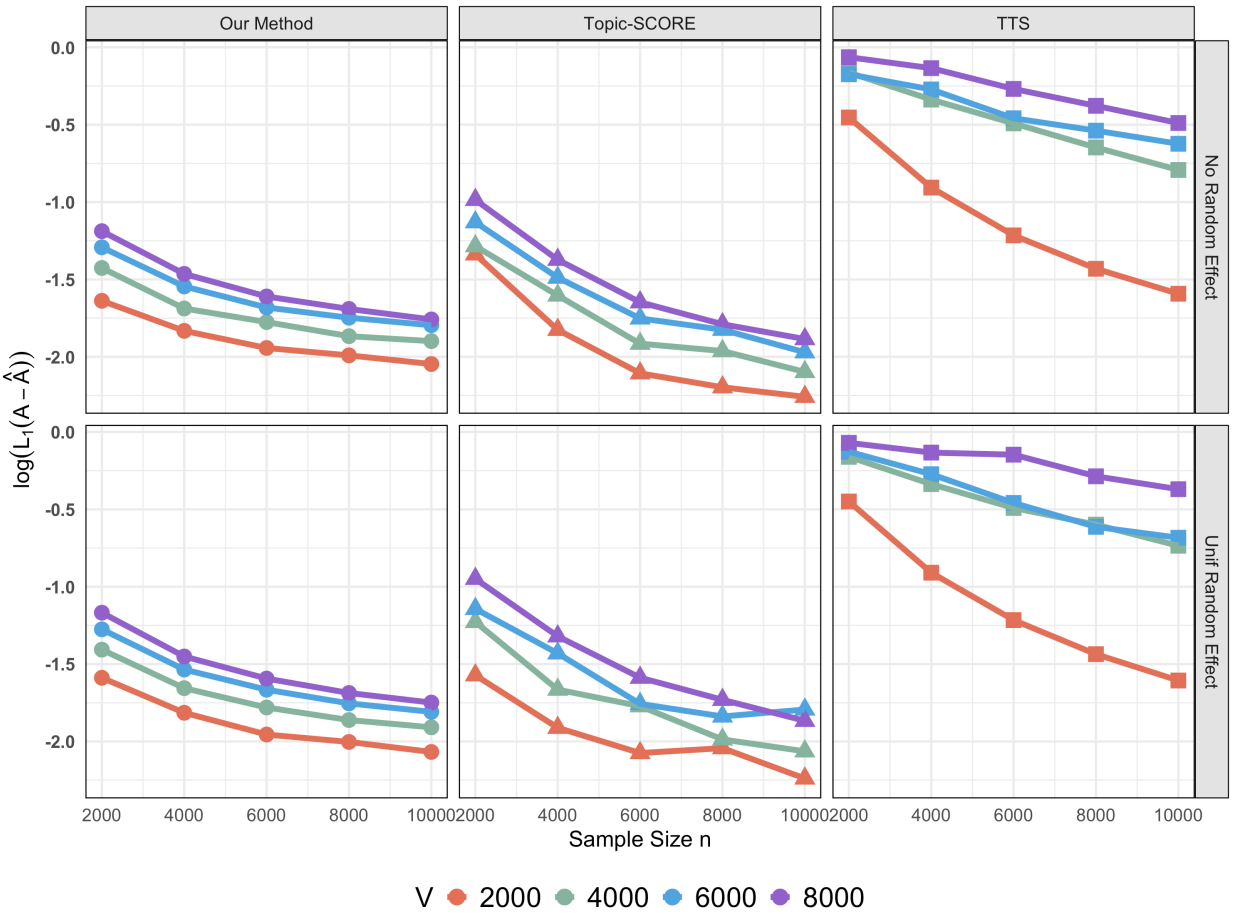


Figure 3.6: $\log(L_1(\mathbf{A} - \widehat{\mathbf{A}}))$ with $K = 3$, constant zero inflation probability and the absence (top) and presence (bottom) of random effects, evaluated across varying (n, V), comparing our method with Topic-SCORE and TTS.

pattern remains informative in theory, it provides weaker and more indirect information, leading to increased estimation variance in finite samples. As a result, even under correct specification, estimation becomes more challenging when the zero-inflation pattern is strongly linked to the underlying signal.

3.5 Real Data Analysis

In this section, we apply our ZIP model to the Multi-Attribute Dataset on Statisticians (MADStat) (Ji et al., 2022), which contains bibliographic information (e.g., author, title, abstract, journal, year, references) for 83,331 papers authored by 47,311 individuals over a 41-year period

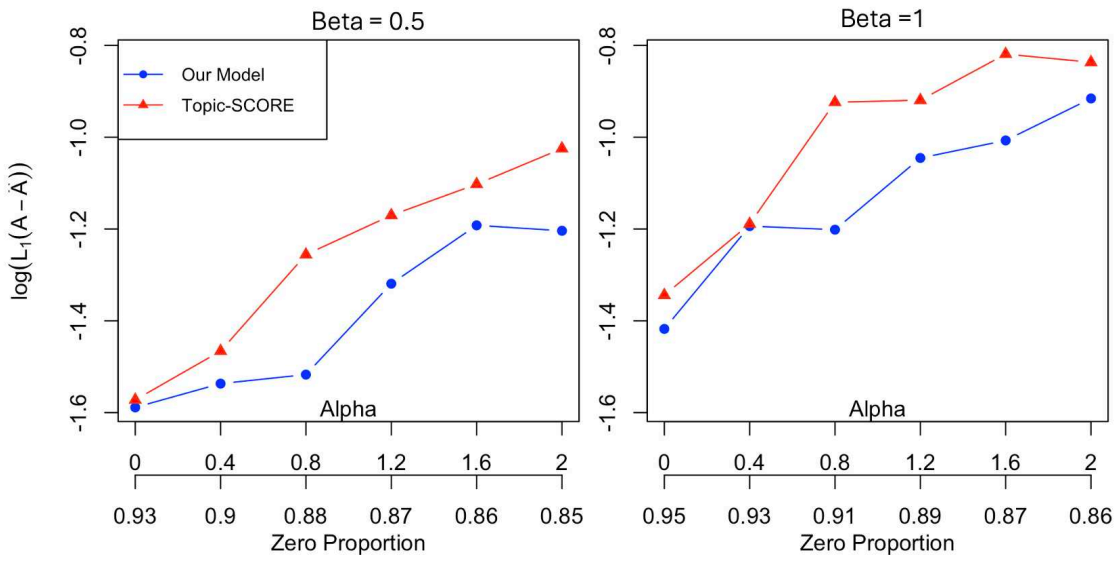


Figure 3.7: $\log(L_1(\mathbf{A} - \widehat{\mathbf{A}}))$ with $n = V = 2000$ and $K = 3$ under different values of α with relative average zero proportions of data.

(1975–2015). For our analysis, we focus on 14,000 abstracts from four leading statistical journals: *Annals of Statistics* (AoS), *Biometrika* (Bka), *Journal of the American Statistical Association* (JASA), and *Journal of the Royal Statistical Society: Series B* (JRSSB). The abstracts in MADStat are preprocessed through tokenization, stemming, and the removal of numbers, punctuation, and stop words. The text corpus consists of $(n, V) = (14000, 2106)$ after the pre-processing. By focusing on these high-impact journals, we explore the topics identified by our model and analyze topic trends over time and across journals.

3.5.1 Selection of the number of topics

The analysis so far has assumed a known K. However, selecting an appropriate K is crucial for real data applications and remains a challenging problem (Ke et al., 2024). To systematically analyze the data and identify topics along with their corresponding representative words, we combine the scree plot from singular value decomposition (SVD) with substantial manual evaluation to determine a suitable number of topics. As shown in Figure 3.8, the scree plot reveals noticeable drops in singular values, suggesting a valid range for K between 4 and 16.

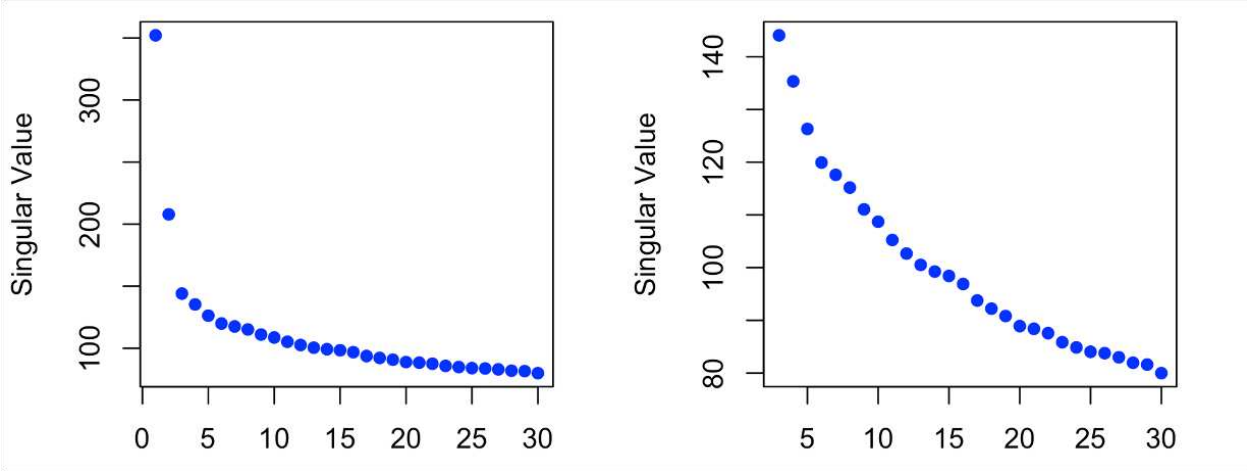


Figure 3.8: Left: the top 30 singular values of $\mathbf{Z}$; Right: excluding the first two singular values, a discernible drop in magnitude becomes evident starting from the 17th singular value.

The analysis has the assumption of known K so far. But it's necessary to consider the estimation of K to ensure applicability to real data analysis. Since we have the pseudo-likelihood of our model and the low rank structure of $\boldsymbol{\lambda}$, we consider the Bayesian information criterion (BIC) based on the factor model (Bai and Ng, 2002; Alessi et al., 2010):

$$\mathrm{BIC}(K) = -\ell(\widehat{\boldsymbol{\lambda}}, \hat{\alpha}, \hat{\eta}|\mathbf{Z}) + cK(n+V)\log(\frac{nV}{n+V}). \tag{3.13}$$

where $\ell(\widehat{\boldsymbol{\lambda}}, \hat{\alpha}, \hat{\eta}|\mathbf{Z})$ follows the definition in Equation (3.8), and $c > 0$ is a tuning parameter to adjust the penalty.

For each value of K within the range of $[4, 16]$, we estimate the model parameters using Algorithm 2 and calculate the Bayesian Information Criterion (BIC) as defined in Equation (3.13) with $c = 0.1$. The results depicted in 3.9 indicate that $K = 11$ is the optimal number of topics which matches the choice in Ke et al. (2024).

3.5.2 Frequent words for the identified topics

With the number of topics $K = 11$, we apply our method to get the estimated word-topic matrix $\widehat{\mathbf{A}}$. Theoretically, each topic should have at least one anchor word. However, in practice, it is uncommon to find a word j with a clear assignment, i.e., a case where $\hat{A}_{jk} = 1$ and $\hat{A}_{jk'} =$

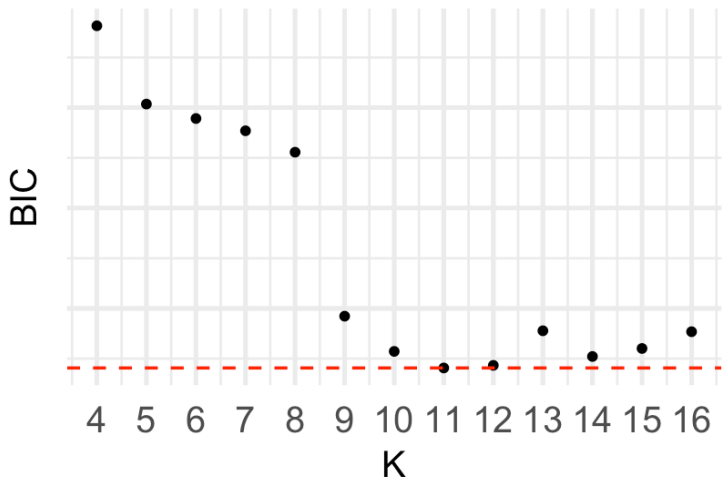


Figure 3.9: BIC criteria under different K.

$0, \forall k' \neq k$. Instead, we identify the top frequent 20 words for each topic based on their topic loading scores, denoted as $a_j(k)$, which quantify the association of word j with topic k. Given $\widehat{\mathbf{A}}$, we follow the rule defined in Ke et al. (2024) to compute $a_j(k) = \hat{A}_{kj} / \left[\sum_{l=1}^{K} \hat{A}_{lj}\right]$ and define the most frequent word j^* for topic k when $j^* = \arg\max_j \{a_j(k) : 1 \leq j \leq V\}$. Figure 3.10 displays the 20 most frequent words for each of the 11 estimated topics. Based on our understanding of the statistical community, we assign topic names by interpreting the frequent words of each topic. For example, "optimal" and "D-optimality" are indicative of experimental design, while "prior" and "posterior" suggest Bayesian statistics. Following this approach, we propose possible topic names, as summarized in Table 3.2.

We evaluate the quality of the estimated topics using the topic coherence Coherence(k) defined in Mimno et al. (2011). It assesses how semantically related the top frequent words within each topic are by measuring their co-occurrence frequency in the corpus. Given the estimated topics from our model and those from Ke et al. (2024), we compute coherence by

$$\text{Coherence}(k) = \sum_{i=2}^{20} \sum_{j=1}^{i-1} \log \frac{D(v_{ki}, v_{kj}) + 1}{D(v_{kj})} \tag{3.14}$$

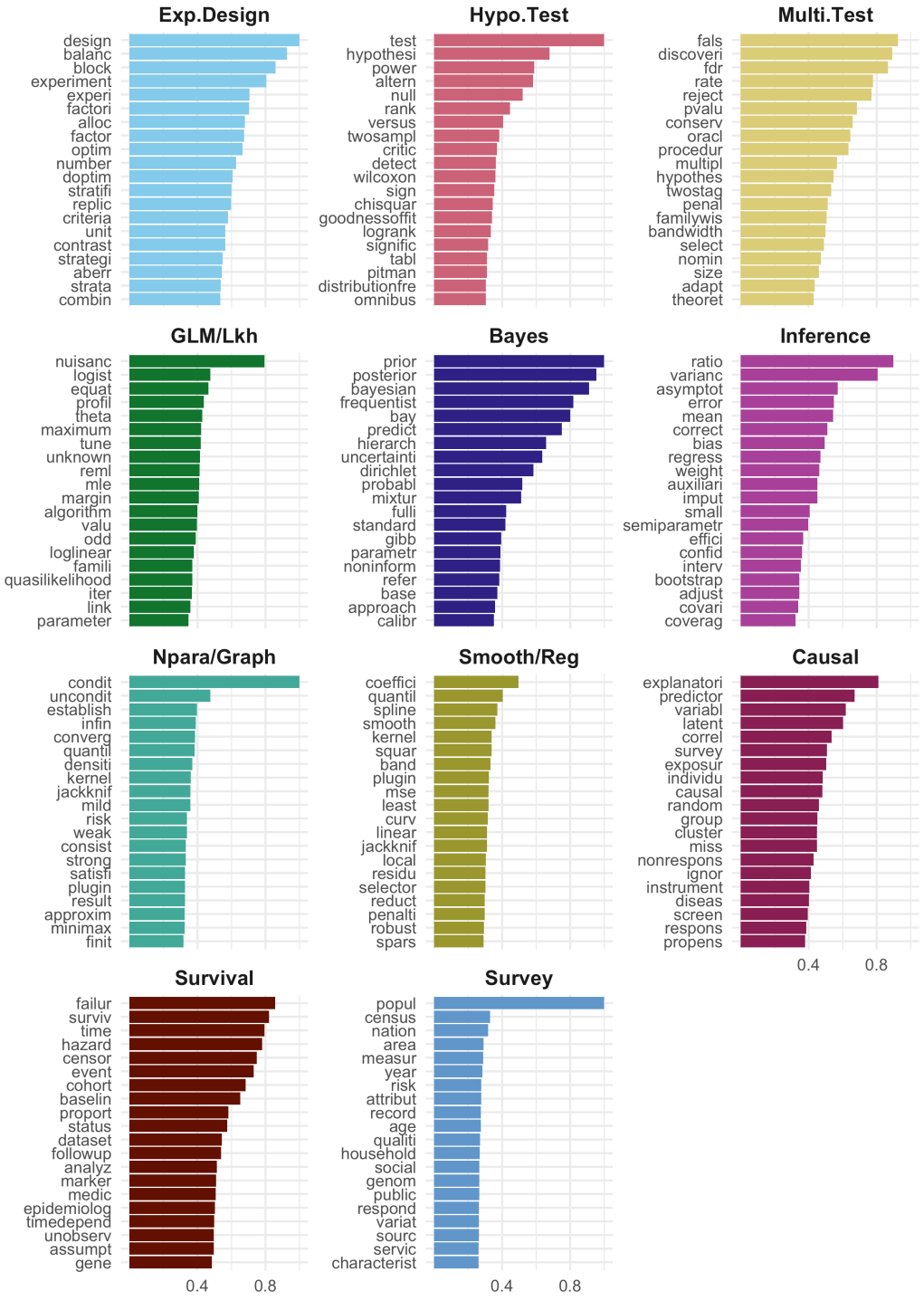


Figure 3.10: For $1 \leq k \leq K$ with $K = 11$, Panel k displays a bar chart of the 20 words with the largest weights $a_j(k)$ among all words, where the length of each bar corresponds to the values.

No.	Label	Topic
1	Exp.Design	Experimental design
2	Hypo.Test	Hypothesis testing
3	Multi.Test	Multiple testing
4	GLM/Lkh	Generalized linear models and likelihood methods
5	Bayes	Bayesian statistics
6	Inference	Statistical inference
7	Npara/Graph	Nonparametric estimation and graphical models
8	Smooth/Reg	Smoothing and regularization
9	Causal	Causal inference
10	Survival	Survival analysis
11	Survey	Survey statistics

Table 3.2: Topics Identified with Labels

where v_{ki} represents i^{th} representative words in topic k, $D(v_{kj})$ is the number of documents containing word v_{kj}, and $D(v_{ki}, v_{kj})$ is co-document frequency of i^{th} and j^{th} representative words (the number of documents containing both words). The measure provides an interpretability-focused assessment of topic quality and has been demonstrated to align more closely with human judgment. Figure 3.11 presents the topic coherence scores for both our method and the results from the pLSI model in Ke et al. (2024). The results indicate that our method consistently produces topics with higher coherence scores, suggesting that the extracted topics are more semantically meaningful and better clustered.

3.5.3 Topic prevalence over time

Given the estimated word-topic information, we obtain the estimation of $\mathbf{W}$ by the weighed least square discussed in Sec 3.2.3. Each row $\widehat{\mathbf{W}}_{i\cdot}$ represents the topic allocation for an individual document i, enabling us to analyze the topic prevalence over time across different journals. We standardize the allocation for each document by $\widetilde{\mathbf{W}}_{i\cdot} = \widehat{\mathbf{W}}_{i\cdot} / \sum_k \widehat{\mathbf{W}}_{ik}$. We then compute the yearly average of $\widetilde{\mathbf{W}}_{i\cdot}$ by averaging standardized topic allocation across all documents published in the same year from the same journal. To enhance stability and reveal long-term trends, we apply a moving average with a 5-year window for smoothing.

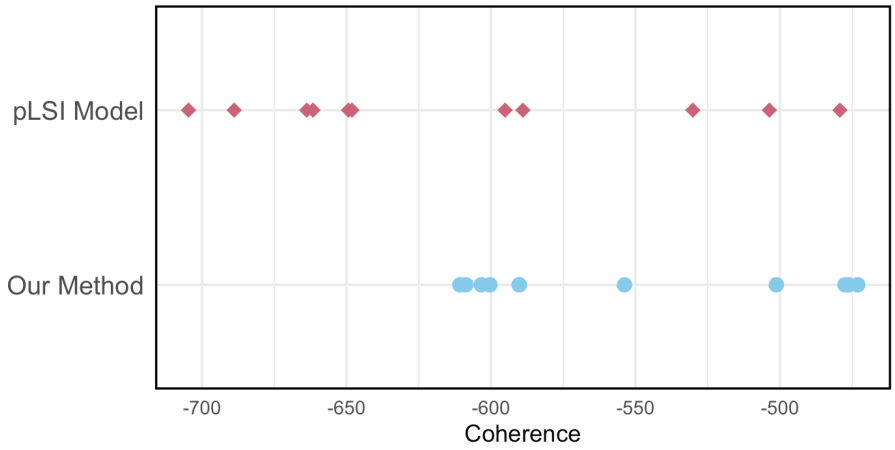


Figure 3.11: Coherence in different topics under different methods.

With the average standardized weights over time across journals, we identify six topics (experiment design, hypothesis testing, multiple testing, Bayesian statistics, asymptotic theory and causal inference) with particularly notable trends for further discussion, as they reflect meaningful shifts in statistical research over time.

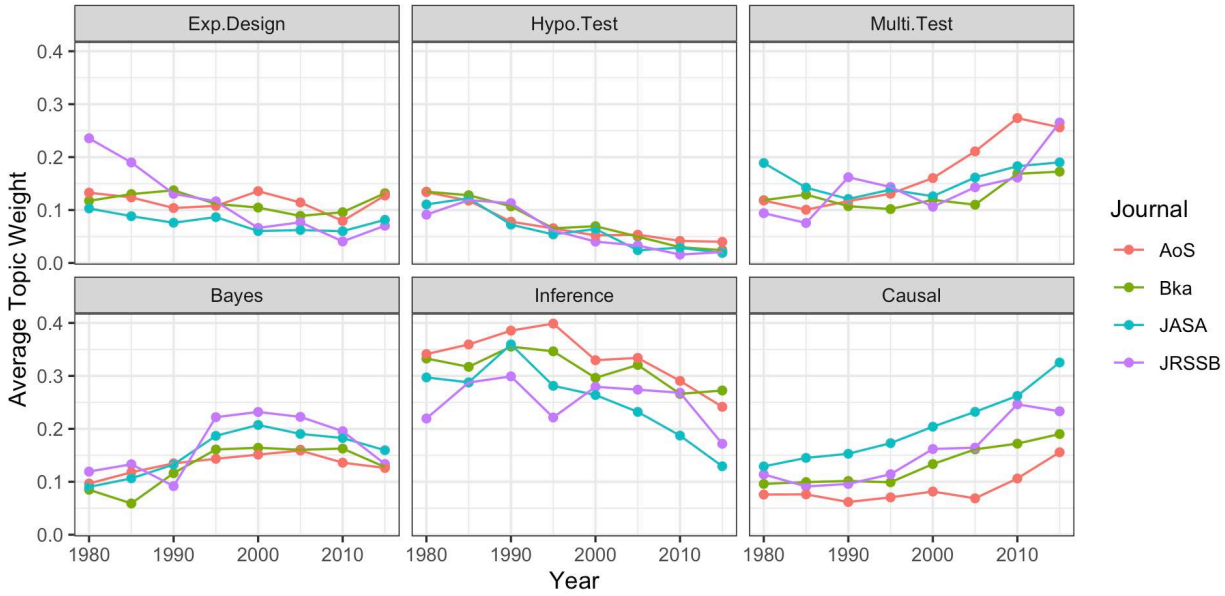


Figure 3.12: Topic prevalent over time across journals.

The trends depicted in Figure 3.12 closely mirror the historical evolution of statistical research across distinct time periods. The prominence of **experimental design** during the 20th century reflects its foundational role in the development of modern statistical methodology. While it remains a fundamental area in statistics, as reflected in the stabilization of its prevalence in the 2000s, modern applications have increasingly shifted toward causal inference and machine learning-driven experimentation, which have built upon and extended classical design principles. The observed decline in **hypothesis testing** suggests a move away from traditional null hypothesis significance testing as the dominant paradigm for statistical inference. This trend is likely influenced by increasing awareness of the limitations and potential misinterpretation of p-values, leading to a reevaluation of their role in scientific practice. In contrast, interest in **multiple testing** has grown markedly since the early 2000s, in parallel with the rise of high-dimensional data analysis. This growth corresponds with the development and widespread application of techniques for controlling the false discovery rate (FDR), particularly in journals such as AoS and JRSSB. The increasing relevance of **Bayesian statistics** from the 1990s through the early 2000s reflects its resurgence driven by advances in computational techniques, most notably MCMC methods. The slight decline in its prevalence after 2010 may be less indicative of reduced interest and more reflective of its integration into broader methodological frameworks, where Bayesian tools are often embedded rather than presented as standalone contributions. The declining focus on **statistical inference** as a general topic over the past two decades underscores a broader disciplinary shift. As predictive modeling and machine learning continue to gain prominence in applied settings, the emphasis has increasingly shifted from explanatory modeling and inference toward predictive accuracy and algorithmic performance. Finally, the steady rise of **causal inference**, particularly after 2000, corresponds with the growing adoption of the potential outcomes framework across disciplines such as economics, epidemiology, and the social sciences. This trend highlights the expanding role of causal reasoning in empirical research and its centrality to data-driven decision-making. The heightened representation of causal inference in JASA further reflects the journal's interdisciplinary scope and its strong emphasis on applied methodology.

Chapter 4

Conclusion and future work

4.1 Potts model with structure information for mutation effects prediction

Chapter 2 studies the long-standing challenge of predicting protein mutation fitness by using the high-dimensional Potts model. To estimate model parameters, we adopt node-wise multinomial regression with a sparse group Lasso penalty, capturing both site-wise and element-wise sparsity in protein sequences. Our method balances selecting significant site pairs with identifying key amino acid-level interactions, and incorporating protein structural data via group weights enhances the estimation accuracy of parameters. Theoretically, we extend recent oracle inequalities for high-dimensional linear models (Bellec et al., 2018) to derive error bounds for our estimator, underscoring its validity. These error bounds match the minimax lower bound for double sparse linear regression, up to a factor specific to the multinomial structure of the Potts model. Experimentally, our method outperforms existing competitors, particularly the widely-used EVM in predicting mutation fitness, validated by high-throughput mutagenesis experiments across multiple protein families. This suggests the potential of our method to advance the understanding of protein functionality in an evolutionary context.

Limitations and Challenges. While our method shows promising results, several areas warrant further exploration. Our method assumes that spatial distances between sites is the primary factor that determines the strength of direct couplings. Including additional factors, such as solubility and solvent accessibility, could enhance predictions and merit investigations. Furthermore, our approach focuses on pairwise effects, but higher-order interactions involving three or more sites may also significantly impact protein functionality. Efficient algorithms to address these higher-order dependencies remain a key challenge for future work.

While parallel computing with node-wise multinomial regressions mitigates some of the demanding computational costs of our method, the high dimensionality of the categorical design matrix, caused by many sites and categories, makes the algorithm slow to converge. This limitation underscores the need for more efficient optimization strategies, and recent advances in nonconvex optimization may offer promising directions for improvement (Qi and Li, 2024). In addition, incorporating complex structural dependencies into data generation from the Potts model remains a difficult task. Accurately capturing such structure is essential for realistic benchmarking and evaluation. Future work is needed to develop more sophisticated generative procedures that better reflect the characteristics of real multiple sequence alignment (MSA) data, analogous to efforts such as scDesign3 (Song et al., 2024) in the context of single-cell RNA sequencing.

Our theoretical guarantees focus on estimators from multinomial regression. While node-wise multinomial regressions provide a solid foundation for estimating the Potts model, the lower bound for estimation in multinomial regression over the double sparse parameter class remains unknown. Consequently, it is unclear whether the derived error bounds fully capture the complexities inherent in multinomial regressions, which is left for future studies.

Extensions and Future Directions. While focused on protein mutations, the proposed method has broader applications in computational biology, such as detecting gene regulatory networks or metabolic pathways, where hierarchical and spatial dependencies are prevalent. The model can also be refined to account for dynamic aspects of protein interactions, incorporating temporal changes in coevolutionary patterns. Our theoretical results can readily extend to accommodate fixed or independently estimated sample weights (for sequence sampling bias) and group weights (for structural information integration) in the penalty. However, in practice, these weights may be derived from the same dataset, a challenge that could be addressed through dedicated analysis or data splitting techniques. Moreover, while this work emphasizes the estimation of the Potts model, statistical inference for the Potts model remains largely unexplored yet holds significant values for understanding uncertainties in predicting protein mutation fitness. We plan to address these directions in future research.

4.2 Topic Modeling by using ZIP

Chapter 3 proposes a zero-inflated Poisson (ZIP) topic modeling framework tailored for analyzing sparse and short-text corpora. By modeling word counts directly using a Poisson distribution and accounting for excess zeros through a zero-inflation component, the framework offers a more expressive alternative to traditional multinomial-based approaches. To capture heterogeneity across documents and words, the model incorporates entry-wise random effects, while a low-rank structure is imposed to preserve interpretability of the latent topic space. Estimation proceeds in two stages: the Poisson intensity matrix and zero-inflation parameters are estimated via a moment-matching and pseudo-likelihood optimization procedure, followed by recovery of the topic-word matrix through spectral decomposition and vertex hunting, using the Sketched Vertex Search (SVS) algorithm. Theoretical guarantees are provided in the form of relative ℓ_1 convergence rates for the recovered topic-word matrix, ensuring statistical consistency. Extensive simulation studies demonstrate the effectiveness of the proposed method, showing improved estimation accuracy across a wide range of sparsity levels, dimensional settings, and random effect magnitudes compared with existing methods. Application to the MADStat dataset further illustrates the practical utility of the approach, revealing underlying topics and interpretable temporal trends that align with well-known developments in statistics.

Limitations and Challenges. Despite these advantages, several challenges remain. The incorporation of both zero-inflation and random effects introduces additional model complexity, which increases sensitivity to initialization and makes the optimization landscape more intricate. While empirical performance remains reliable even under mildly heavy-tailed unbounded random effects, the lack of formal theoretical guarantees in such settings warrants further investigation. Additionally, although the number of topics was selected using the Bayesian Information Criterion (BIC) in the real data analysis, selecting the optimal number of topics remains an open problem that calls for more principled and scalable model selection strategies.

Extensions and Future Directions. Several methodological extensions are possible. Integrating document-level covariates, such as author seniority, collaboration networks, and journal metadata,

could enhance both the interpretability and predictive accuracy of the model. For example, authors who frequently collaborate may develop shared research interests, leading to recurring focus on similar topics across their publications. Similarly, articles published in specialized journals may consistently highlight domain-specific themes, such as biostatistics in a medical journal or machine learning in a computer science journal, reflecting editorial priorities and audience expectations. Furthermore, the model could be extended to accommodate temporal dynamics, which enables the analysis of topic evolution over time to offer insights into shifting thematic trends. Incorporating such external variables may necessitate reformulating the underlying intensity structure to accommodate covariate-dependent effects and potential interactions, thereby enabling more nuanced modeling of topic prevalence across documents.

Beyond text analysis, the proposed ZIP modeling framework holds promise for other domains characterized by sparse count data, notably genomics. Single-cell RNA sequencing data, for instance, exhibits extreme sparsity due to both technical limitations and biological variability. The zero-inflation component naturally models this duality, distinguishing between technical zeros (arising from measurement error) and biological zeros (true absence of gene expression). Topics in this setting can correspond to latent gene modules or cellular pathways, offering biologically interpretable insights. The random effects component can accommodate batch effects, cell-type variability, and experimental conditions, thereby improving robustness to noise and enhancing biological relevance. Applying this framework to such data could yield novel insights into cellular heterogeneity and gene regulatory mechanisms, while providing a flexible statistical tool that accounts for the unique challenges of high-throughput sequencing technologies.

4.3 Discussion

This dissertation presents two complementary contributions to the statistical modeling of high-dimensional categorical data. These contributions are unified by the overarching aim of developing interpretable and theoretically grounded models for complex, sparse, and structured observations. Although the two projects are motivated by different applications in statistical genomics and

natural language processing, both emphasize the importance of structural assumptions, low-rank representations, and specialized estimation procedures for addressing the challenges associated with high-dimensional discrete data.

These methodological advances offer contributions to the fields of statistical genomics and natural language processing, providing a flexible foundation for future research. Potential extensions include the integration of auxiliary covariates, the incorporation of temporal or spatial dependencies, and improvements in computational scalability through advanced optimization techniques. More broadly, this work highlights the critical importance of domain-informed model design and opens promising avenues for adapting these frameworks to a wider array of data-intensive scientific and technological problems.

Bibliography

Abramovich, F., Grinshtein, V., and Levy, T. (2021). Multiclass classification by sparse multinomial logistic regression. *IEEE Trans. Inform. Theory*, 67(7):4637–4646.

Adzhubei, I. A., Schmidt, S., Peshkin, L., Ramensky, V. E., Gerasimova, A., Bork, P., Kondrashov, A. S., and Sunyaev, S. R. (2010). A method and server for predicting damaging missense mutations. *Nat. Methods.*, 7(4):248–249.

Airoldi, E. M. and Bischof, J. M. (2016). Improving and evaluating topic models and other models of text. *Journal of the American Statistical Association*, 111(516):1381–1403.

Alessi, L., Barigozzi, M., and Capasso, M. (2010). Improved penalization for determining the number of factors in approximate factor models. *Statistics & Probability Letters*, 80(23-24):1806–1813.

Araya, C. L., Fowler, D. M., Chen, W., Muniez, I., Kelly, J. W., and Fields, S. (2012). A fundamental protein property, thermodynamic stability, revealed solely from large-scale measurements of protein function. *Proc. Natl. Acad. Sci.*, 109(42):16858–16863.

Arora, S., Ge, R., and Moitra, A. (2012). Learning topic models–going beyond svd. In *2012 IEEE 53rd annual symposium on foundations of computer science*, pages 1–10. IEEE.

Baccanari, D., Stone, D., and Kuyper, L. (1981). Effect of a single amino acid substitution on escherichia coli dihydrofolate reductase catalysis and ligand binding. *J. Biol. Chem.*, 256(4):1738–1747.

Bai, J. and Ng, S. (2002). Determining the number of factors in approximate factor models. *Econometrica*, 70(1):191–221.

Balakrishnan, S., Kamisetty, H., Carbonell, J. G., Lee, S.-I., and Langmead, C. J. (2011). Learning generative models for protein fold families. *Proteins.*, 79(4):1061–1078.

Barber, R. F. and Ha, W. (2018). Gradient descent with non-convex constraints: local concavity determines convergence. *Information and Inference: A Journal of the IMA*, 7(4):755–806.

Becht, E., McInnes, L., Healy, J., Dutertre, C.-A., Kwok, I. W., Ng, L. G., Ginhoux, F., and Newell, E. W. (2019). Dimensionality reduction for visualizing single-cell data using UMAP. *Nat. Biotechnol.*, 37(1):38–44.

Bellec, P. C., Lecué, G., and Tsybakov, A. B. (2018). Slope meets Lasso: improved oracle bounds and optimality. *Ann. Statist.*, 46(6B):3603–3642.

Bershtein, S., Mu, W., and Shakhnovich, E. I. (2012). Soluble oligomerization provides a beneficial fitness effect on destabilizing mutations. *Proc. Natl. Acad. Sci.*, 109(13):4857–4862.

Bing, X., Bunea, F., and Wegkamp, M. (2020a). A fast algorithm with minimax optimal guarantees for topic models with an unknown number of topics. *Bernoulli*, 26(3):1765 – 1796.

Bing, X., Bunea, F., and Wegkamp, M. (2020b). Optimal estimation of sparse topic models. *Journal of Machine Learning Research*, 21(177):1–45.

Blei, D. M. and Lafferty, J. D. (2006). Dynamic topic models. In *Proceedings of the 23rd international conference on Machine learning*, pages 113–120.

Blei, D. M., Ng, A. Y., and Jordan, M. I. (2003). Latent dirichlet allocation. *Journal of machine Learning research*, 3(Jan):993–1022.

Böhning, D., Dietz, E., Schlattmann, P., Mendonca, L., and Kirchner, U. (1999). The zero-inflated poisson model and the decayed, missing and filled teeth index in dental epidemiology. *Journal of the Royal Statistical Society Series A: Statistics in Society*, 162(2):195–209.

Bollobás, B. and Riordan, O. (2011). Sparse graphs: metrics and random models. *Random Struct. Algorithms*, 39(1):1–38.

Boriah, S., Chandola, V., and Kumar, V. (2008). Similarity measures for categorical data: a comparative evaluation. In *Proceedings of the 2008 SIAM International Conference on Data Mining*, pages 243–254.

Boucheron, S., Lugosi, G., and Massart, P. (2013). *Concentration Inequalities: A Nonasymptotic Theory of Independence*. Oxford University Press.

Bühlmann, P. and van de Geer, S. (2011). *Statistics for High-Dimensional Data: Methods, Theory and Applications*. Springer Series in Statistics. Springer Berlin, Heidelberg, first edition.

Cai, T. T., Li, H., Ma, J., and Xia, Y. (2019). Differential markov random field analysis with an application to detecting differential microbial community networks. *Biometrika*, 106(2):401–416.

Cai, T. T., Zhang, A. R., and Zhou, Y. (2022). Sparse group lasso: optimal sample complexity, convergence rate, and statistical inference. *IEEE Trans. Inform. Theory*, 68(9):5975–6002.

Carbonetto, P., Sarkar, A., Wang, Z., and Stephens, M. (2022). Non-negative matrix factorization algorithms greatly improve topic model fits.

Carlo, C. M. (2004). Markov chain Monte Carlo and Gibbs sampling. Available at http://membres-timc.imag.fr/Olivier.Francois/mcmc_gibbs_sampling.pdf. Lecture Notes for EEB 581.

Casella, G. and George, E. I. (1992). Explaining the Gibbs sampler. *Am. Stat.*, 46(3):167–174.

Chen, Y. and Wainwright, M. J. (2015). Fast low-rank estimation by projected gradient descent: General statistical and algorithmic guarantees. *arXiv preprint arXiv:1509.03025*.

Consortium, T. U. (2020). Uniprot: the universal protein knowledgebase in 2021. *Nucleic. Acids. Res.*, 49(D1):D480–D489.

Deek, R. A. and Li, H. (2021). A zero-inflated latent dirichlet allocation model for microbiome studies. *Frontiers in Genetics*, 11:602594.

Di Nardo, A. A., Larson, S. M., and Davidson, A. R. (2003). The relationship between conservation, thermodynamic stability, and function in the SH3 domain hydrophobic core. *J. Mol. Biol.*, 333(3):641–655.

Donoho, D. and Stodden, V. (2003). When does non-negative matrix factorization give a correct decomposition into parts? In *Advances in Neural Information Processing Systems*, volume 16. MIT Press.

Edwards, R. G. and Sokal, A. D. (1988). Generalization of the Fortuin-Kasteleyn-Swendsen-Wang representation and Monte Carlo algorithm. *Phys. Rev. D.*, 38(6):2009–2012.

Friedman, J., Hastie, T., and Tibshirani, R. (2010). Regularization paths for generalized linear models via coordinate descent. *J. Stat. Softw.*, 33(1):1–22.

Gan, Z., Chen, C., Henao, R., Carlson, D., and Carin, L. (2015). Scalable deep poisson factor analysis for topic modeling. In Bach, F. and Blei, D., editors, *Proceedings of the 32nd International Conference on Machine Learning*, volume 37 of *Proceedings of Machine Learning Research*, pages 1823–1832, Lille, France. PMLR.

Gao, J., Pang, X., Zhang, L., Li, S., Qin, Z., Xie, X., and Liu, J. (2023). Transcriptome analysis reveals the neuroprotective effect of dlg4 against fastigial nucleus stimulation-induced ischemia/reperfusion injury in rats. *BMC Neurosci.*, 24(1):40.

Gelfand, A. E. (2000). Gibbs sampling. *J. Amer. Statist. Assoc.*, 95(452):1300–1304.

Grootendorst, M. (2022). Bertopic: Neural topic modeling with a class-based tf-idf procedure. *arXiv preprint arXiv:2203.05794*.

Guo, Z., Rakshit, P., Herman, D. S., and Chen, J. (2021). Inference for the case probability in high-dimensional logistic regression. *J. Mach. Learn. Res.*, 22:Paper No. [254], 54.

Henao, R., Lu, J. T., Lucas, J. E., Ferranti, J., and Carin, L. (2016). Electronic health record analysis via deep poisson factor models. *Journal of Machine Learning Research*, 17(186):1–32.

Hofmann, T. (1999). Probabilistic latent semantic indexing. In *Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval*, pages 50 – 57.

Hopf, T. A., Ingraham, J. B., Poelwijk, F. J., Schärfe, C. P., Springer, M., Sander, C., and Marks, D. S. (2017). Mutation effects predicted from sequence co-variation. *Nat. Biotechnol.*, 35(2):128–135.

Hsieh, C.-J., Chiang, K.-Y., and Dhillon, I. S. (2012). Low rank modeling of signed networks. In *Proceedings of the 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '12, page 507–515.

Huang, J., Ma, S., and Zhang, C.-H. (2008). Adaptive lasso for sparse high-dimensional regression models. *Statist. Sinica*, 18(4):1603–1618.

Ising, E. (1925). Beitrag zur theorie des ferromagnetismus. *Z. Physik*, 31(1):253–258.

Izenman, A. J. (2021). Sampling algorithms for discrete Markov random fields and related graphical models. *J. Amer. Statist. Assoc.*, 116(536):2065–2086.

Jain, P., Meka, R., and Dhillon, I. (2010). Guaranteed rank minimization via singular value projection. In Lafferty, J., Williams, C., Shawe-Taylor, J., Zemel, R., and Culotta, A., editors, *Advances in Neural Information Processing Systems*, volume 23. Curran Associates, Inc.

Jernigan, R., Jia, K., Ren, Z., and Zhou, W. (2021). Large-scale multiple inference of collective dependence with applications to protein function. *Ann. Appl. Stat.*, 15(2):902–924.

Ji, P., Jin, J., Ke, Z. T., and Li, W. (2022). Co-citation and co-authorship networks of statisticians. *Journal of Business & Economic Statistics*, 40(2):469–485.

Jiang, H., Zhou, R., Zhang, L., Wang, H., and Zhang, Y. (2017). A topic model based on poisson decomposition. In *Proceedings of the 2017 ACM on Conference on Information and Knowledge Management*, pages 1489–1498.

Jin, J., Ke, Z. T., and Luo, S. (2024). Mixed membership estimation for social networks. *Journal of Econometrics*, 239(2):105369.

Jones, D. T., Buchan, D. W., Cozzetto, D., and Pontil, M. (2012). Psicov: precise structural contact prediction using sparse inverse covariance estimation on large multiple sequence alignments. *Bioinformatics*, 28(2):184–190.

Ke, Z. T., Ji, P., Jin, J., and Li, W. (2024). Recent advances in text analysis. *Annual Review of Statistics and Its Application*, 11(1):347–372.

Ke, Z. T. and Wang, M. (2024). Using svd for topic modeling. *Journal of the American Statistical Association*, 119(545):434–449.

Kitzman, J. O., Starita, L. M., Lo, R. S., Fields, S., and Shendure, J. (2015). Massively parallel single-amino-acid mutagenesis. *Nat. Methods.*, 12(3):203–206.

Koltchinskii, V. (2011). *Oracle Inequalities in Empirical Risk Minimization and Sparse Recovery Problems*, volume 2033 of *Lecture Notes in Mathematics*. Springer Berlin, Heidelberg, first edition.

Lambert, D. (1992). Zero-inflated poisson regression, with an application to defects in manufacturing. *Technometrics*, 34(1):1–14.

Ledoux, M. and Talagrand, M. (1991). *Probability in Banach Spaces: Isoperimetry and Processes*, volume 23 of *Classics in Mathematics*. Springer Berlin, Heidelberg, first edition.

Lek, M., Karczewski, K. J., Minikel, E. V., Samocha, K. E., Banks, E., Fennell, T., O'Donnell-Luria, A. H., Ware, J. S., Hill, A. J., Cummings, B. B., et al. (2016). Analysis of protein-coding genetic variation in 60,706 humans. *Nature*, 536(7616):285–291.

Levy, R. M., Haldane, A., and Flynn, W. F. (2017). Potts hamiltonian models of protein co-variation, free energy landscapes, and evolutionary fitness. *Curr. Opin. Struct. Biol.*, 43:55–62.

Levy, T. and Abramovich, F. (2023). Generalization error bounds for multiclass sparse linear classifiers. *J. Mach. Learn. Res.*, 24:Paper No. [151], 35.

Li, H. and Luan, Y. (2003). Kernel cox regression models for linking gene expression profiles to censored survival data. *Pac. Symp. Biocomput.*, pages 65–76.

Li, Q., Wang, X., Liang, F., Yi, F., Xie, Y., Gazdar, A., and Xiao, G. (2019). A Bayesian hidden Potts mixture model for analyzing lung cancer pathology images. *Biostatistics*, 20(4):565–581.

Li, Q., Yi, F., Wang, T., Xiao, G., and Liang, F. (2017). Lung cancer pathological image analysis using a hidden potts model. *Cancer Inform.*, 16:1176935117711910.

Li, Y., Zhu, R., Qu, A., Ye, H., and and, Z. S. (2021). Topic modeling on triage notes with semiorthogonal nonnegative matrix factorization. *Journal of the American Statistical Association*, 116(536):1609–1624.

Li, Z., Zhang, Y., and Yin, J. (2023). Sharp minimax optimality of LASSO and SLOPE under double sparsity assumption. Available at https://arxiv.org/abs/2308.09548.

Lin, D. (1998). An information-theoretic definition of similarity. In *Proceedings of the Fifteenth International Conference on Machine Learning*, pages 296–304.

Lounici, K., Pontil, M., van de Geer, S., and Tsybakov, A. B. (2011). Oracle inequalities and optimal inference under group sparsity. *Ann. Statist.*, 39(4):2164–2204.

Ma, R., Guo, Z., Cai, T. T., and Li, H. (2022). Statistical inference for genetic relatedness based on high-dimensional logistic regression. Available at https://arxiv.org/abs/2202.10007.

Marks, D. S., Colwell, L. J., Sheridan, R., Hopf, T. A., Pagnani, A., Zecchina, R., and Sander, C. (2011). Protein 3d structure computed from evolutionary sequence variation. *PLOS ONE*, 6(12):e28766.

Marks, D. S., Hopf, T. A., and Sander, C. (2012). Protein structure prediction from sequence variation. *Nat. Biotechnol.*, 30(11):1072–1080.

McLaughlin Jr., R. N., Poelwijk, F. J., Raman, A., Gosal, W. S., and Ranganathan, R. (2012). The spatial architecture of protein function and adaptation. *Nature*, 491(7422):138–142.

McMurdie, P. J. and Holmes, S. (2014). Waste not, want not: why rarefying microbiome data is inadmissible. *PLoS computational biology*, 10(4):e1003531.

Melamed, D., Young, D. L., Gamble, C. E., Miller, C. R., and Fields, S. (2013). Deep mutational scanning of an rrm domain of the saccharomyces cerevisiae poly(a)-binding protein. *RNA*, 19(11):1537–1551.

Melnikov, A., Rogov, P., Wang, L., Gnirke, A., and Mikkelsen, T. S. (2014). Comprehensive mutational scanning of a kinase in vivo reveals substrate-dependent fitness landscapes. *Nucleic. Acids. Res.*, 42(14):e112.

Mimno, D., Wallach, H., Talley, E., Leenders, M., and McCallum, A. (2011). Optimizing semantic coherence in topic models. In *Proceedings of the 2011 conference on empirical methods in natural language processing*, pages 262–272.

Morcos, F., Pagnani, A., Lunt, B., Bertolino, A., Marks, D. S., Sander, C., Zecchina, R., Onuchic, J. N., Hwa, T., and Weigt, M. (2011). Direct-coupling analysis of residue coevolution captures native contacts across many protein families. *Proc. Natl. Acad. Sci.*, 108(49):E1293–E1301.

Møller, J., Pettitt, A. N., Reeves, R., and Berthelsen, K. K. (2006). An efficient markov chain monte carlo method for distributions with intractable normalising constants. *Biometrika*, 93(2):451–458.

Newman, D. (2008). Bag of Words. UCI Machine Learning Repository. DOI: https://doi.org/10.24432/C5ZG6P.

Ng, P. C. and Henikoff, S. (2003). Sift: predicting amino acid changes that affect protein function. *Nucleic. Acids. Res.*, 31(13):3812–3814.

Nibbering, D. and Hastie, T. J. (2022). Multiclass-penalized logistic regression. *Comput. Statist. Data Anal.*, 169:Paper No. 107414, 16.

Park, J. and Haran, M. (2018). Bayesian inference in the presence of intractable normalizing functions. *J. Amer. Statist. Assoc.*, 113(523):1372–1390.

Potts, R. B. (1952). Some generalized order-disorder transformations. *Proc. Cambridge Philos. Soc.*, 48:106–109.

Prange, O., Wong, T. P., Gerrow, K., Wang, Y. T., and El-Husseini, A. (2004). A balance between excitatory and inhibitory synapses is controlled by psd-95 and neuroligin. *Proc. Natl. Acad. Sci.*, 101(38):13915–13920.

Qi, M. and Li, T. (2024). The non-overlapping statistical approximation to overlapping group lasso. *J. Mach. Learn. Res.*, 25(115):1–70.

Ravikumar, P., Wainwright, M. J., and Lafferty, J. D. (2010). High-dimensional ising model selection using ℓ_1-regularized logistic regression. *Ann. Statist.*, 38(3):1287–1319.

Razaee, Z. and Amini, A. (2020). The potts-ising model for discrete multivariate data. In *Adv. Neural Inf. Process. Syst.*, volume 33, pages 13727–13737. Curran Associates, Inc.

Rockah-Shmuel, L., Tóth-Petróczy, Á., and Tawfik, D. S. (2015). Systematic mapping of protein mutational space by prolonged drift reveals the deleterious effects of seemingly neutral mutations. *PLoS. Comput. Biol.*, 11(8):e1004421.

Schnell, J. R., Dyson, H. J., and Wright, P. E. (2004). Structure, dynamics, and catalytic function of dihydrofolate reductase. *Annu. Rev. Biophys. Biomol. Struct.*, 33:119–140.

Schweitzer, B. I., Dicker, A. P., and Bertino, J. R. (1990). Dihydrofolate reductase as a therapeutic target. *FASEB J.*, 4(8):2441–2452.

Simon, N., Friedman, J., Hastie, T., and Tibshirani, R. (2013). A sparse-group lasso. *J. Comput. Graph. Statist.*, 22(2):231–245.

Song, D., Wang, Q., Yan, G., Liu, T., Sun, T., and Li, J. J. (2024). scDesign3 generates realistic in silico data for multimodal single-cell and spatial omics. *Nat. Biotechnol.*, 42(2):247–252.

Starita, L. M., Pruneda, J. N., Lo, R. S., Fowler, D. M., Kim, H. J., Hiatt, J. B., Shendure, J., Brzovic, P. S., Fields, S., and Klevit, R. E. (2013). Activity-enhancing mutations in an e3 ubiquitin ligase identified by high-throughput mutagenesis. *Proc. Natl. Acad. Sci.*, 110(14):E1263–E1272.

Tian, Y. and Feng, Y. (2023). Transfer learning under high-dimensional generalized linear models. *J. Amer. Statist. Assoc.*, 118(544):2684–2697.

Tian, Y., Rusinek, H., Masurkar, A. V., and Feng, Y. (2024). ℓ_1-penalized multinomial regression: estimation, inference, and prediction, with an application to risk factor identification for different dementia subtypes. Available at https://arxiv.org/abs/2302.02310.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *J. Roy. Statist. Soc. Ser. B*, 58(1):267–288.

Tran, H., Liu, Y., and Donnat, C. (2023). Sparse topic modeling via spectral decomposition and thresholding.

Tutz, G. and Gertheiss, J. (2016). Regularized regression for categorical data. *Stat. Model.*, 16(3):161–200.

van de Geer, S., Bühlmann, P., Ritov, Y., and Dezeure, R. (2014). On asymptotically optimal confidence regions and tests for high-dimensional models. *Ann. Statist.*, 42(3):1166–1202.

Vincent, M. and Hansen, N. R. (2014). Sparse group lasso and high dimensional multinomial classification. *Comput. Statist. Data Anal.*, 71:771–786.

Wahle, J. P., Ruas, T., Mohammad, S. M., and Gipp, B. (2022). D3: A massive dataset of scholarly metadata for analyzing the state of computer science research. In *Proceedings of The*

13th Language Resources and Evaluation Conference, Marseille, France. European Language Resources Association.

Wainwright, M. J. (2019). *High-Dimensional Statistics: A Non-Asymptotic Viewpoint*, volume 48 of *Cambridge Series in Statistical and Probabilistic Mathematics*. Cambridge University Press, Cambridge, first edition.

Wang, J., Lu, H., Plataniotis, K. N., and Lu, J. (2009). Gaussian kernel optimization for pattern classification. *Pattern Recognit.*, 42(7):1237–1247.

Wang, L. L., Lo, K., Chandrasekhar, Y., Reas, R., Yang, J., Burdick, D., Eide, D., Funk, K., Katsis, Y., Kinney, R., Li, Y., Liu, Z., Merrill, W., Mooney, P., Murdick, D., Rishi, D., Sheehan, J., Shen, Z., Stilson, B., Wade, A., Wang, K., Wang, N. X. R., Wilhelm, C., Xie, B., Raymond, D., Weld, D. S., Etzioni, O., and Kohlmeier, S. (2020). Cord-19: The covid-19 open research dataset.

Wu, R., Zhang, L., and Cai, T. T. (2023). Sparse topic modeling: Computational efficiency, near-optimal algorithms, and statistical inference. *Journal of the American Statistical Association*, 118(543):1849–1861. PMID: 37771513.

Xia, L., Lee, C., and Li, J. J. (2024). Statistical method scDEED for detecting dubious 2D single-cell embeddings and optimizing t-SNE and UMAP hyperparameters. *Nat. Commun.*, 15(1):1753.

Xu, L., Paterson, A. D., Turpin, W., and Xu, W. (2015). Assessment and selection of competing models for zero-inflated microbiome data. *PloS one*, 10(7):e0129606.

Xu, T., Demmer, R. T., and Li, G. (2021). Zero-inflated poisson factor model with application to microbiome read counts. *Biometrics*, 77(1):91–101.

Yuan, M. and Lin, Y. (2006). Model selection and estimation in regression with grouped variables. *J. R. Stat. Soc. Ser. B Stat. Methodol.*, 68(1):49–67.

Zhang, J. and Li, Y. (2023). High-dimensional gaussian graphical regression models with covariates. *J. Amer. Statist. Assoc.*, 118(543):2088–2100.

Zhou, M., Hannah, L., Dunson, D., and Carin, L. (2012). Beta-negative binomial process and poisson factor analysis. In *Artificial Intelligence and Statistics*, pages 1462–1471. PMLR.

Zipf, G. (1999). *The Psycho-Biology of Language: An Introduction to Dynamic Philology*. Cognitive psychology. Routledge.

Appendix A

Supplemental materials for Chapter 2

In Section A.1, we provide the detailed proof of Theorem 2.3.1, with auxiliary results presented in Section A.2. In Section A.3, we provide details of the data generation procedure that mimics the real data, as well as additional results from simulation studies.

A.1 Proof of Theorem 2.3.1

A.1.1 Notation and Preliminaries

For simplicity, for each site $j \in [d]$, we suppress the subscript j from the related notations. Specifically, for each $k \in [K]$ and $i \in [n]$, let $\boldsymbol{\theta} = \boldsymbol{\theta}_j \in \mathbb{R}^K$, $\boldsymbol{\gamma} = \boldsymbol{\gamma}_j \in \mathbb{R}^D$ with $D = (d-1)K^2$, $y_k^{(i)} = y_{jk}^{(i)} \in \{0,1\}$ and $\mathbf{x}^{(i)} = \mathbf{x}_{-j}^{(i)} \in \{0,1\}^{(d-1)K}$. Moreover, for components in $\boldsymbol{\gamma}$, we denote $\boldsymbol{\gamma}_{(r)} = \boldsymbol{\gamma}_{j(r)} \in \mathbb{R}^{K^2}$ and $\boldsymbol{\gamma}_k = \boldsymbol{\gamma}_{j\bullet,k\bullet} \in \mathbb{R}^{(d-1)K}$, for $r \neq j$ and $k \in [K]$. As the vector of dependency $\boldsymbol{\gamma} \in \mathbb{R}^D$ consists of $d-1$ groups of parameters, with a slightly abuse of notation, we denote the $d-1$ groups as $\{\boldsymbol{\gamma}_{(r)} : r \in [d-1]\}$ in this chapter.

Set $\boldsymbol{\gamma}^\circ = (\boldsymbol{\theta}^{\mathrm{T}}, \boldsymbol{\gamma}^{\mathrm{T}})^{\mathrm{T}}$ to be a general parameter vector, $\boldsymbol{\gamma}^{\circ *} = (\boldsymbol{\theta}^{*\mathrm{T}}, \boldsymbol{\gamma}^{*\mathrm{T}})^{\mathrm{T}}$ to be the true parameter vector, and $\widehat{\boldsymbol{\gamma}}^\circ$ to be the estimated parameter vector. Since $\boldsymbol{\gamma}^*$ is assumed to be (s, s_g)-sparse, $\boldsymbol{\gamma}^{\circ *}$ is then (s°, s_g°)-sparse, where $s^\circ = s + K$ and $s_g^\circ = s_g + 1$ by treating $\boldsymbol{\theta}^*$ as an additional group. In our analysis, unless otherwise specified, the notations c, C, c_1, C_1, and so on denote positive absolute constants, which may vary from context to context.

With these notations, our estimator $\widehat{\boldsymbol{\gamma}}^\circ$ can be equivalently obtained from the following penalized optimization:

$$\widehat{\boldsymbol{\gamma}}^\circ = \underset{\boldsymbol{\gamma}^\circ}{\operatorname{argmin}}\, \ell(\boldsymbol{\gamma}^\circ, \mathbf{Y}, \mathbf{X}) + h^\circ(\boldsymbol{\gamma}^\circ),$$

where the negative log-likelihood is

$$\ell(\boldsymbol{\gamma}^\circ) := \ell(\boldsymbol{\gamma}^\circ, \mathbf{Y}, \mathbf{X}) = \frac{1}{n}\sum_{i=1}^{n}\left(\log\left[1 + \sum_{k=1}^{K}\exp(\boldsymbol{\gamma}_k^{\circ\mathrm{T}}\mathbf{x}^{(i)})\right] - \sum_{k=1}^{K} y_k^{(i)}(\boldsymbol{\gamma}_k^{\circ\mathrm{T}}\mathbf{x}^{(i)})\right),$$

where $\mathbf{X} = ((\mathbf{x}^{(1)})^{\mathrm{T}}, \ldots, (\mathbf{x}^{(n)})^{\mathrm{T}})^{\mathrm{T}} \in \mathbb{R}^{n \times (1+(d-1)K)}$ is the design matrix by a slightly abuse of notation such that $\mathbf{x}^{(i)} = (1, (\mathbf{x}_1^{(i)})^{\mathrm{T}}, \ldots, (\mathbf{x}_{j-1}^{(i)})^{\mathrm{T}}, (\mathbf{x}_{j+1}^{(i)})^{\mathrm{T}}, \ldots, (\mathbf{x}_d^{(i)})^{\mathrm{T}})^{\mathrm{T}} \in \{0,1\}^{1+(d-1)K}$ includes 1 for the intercept,

$$\boldsymbol{\gamma}_k^{\circ} = (\theta_k, \boldsymbol{\gamma}_k^{\mathrm{T}})^{\mathrm{T}} \in \mathbb{R}^{1+(d-1)K}, \tag{A.1}$$

and $h^{\circ}(\boldsymbol{\gamma}^{\circ})$ is the sparse group penalty, which is convex in $\boldsymbol{\gamma}^{\circ}$, given by

$$h^{\circ}(\boldsymbol{\gamma}^{\circ}) := h(\boldsymbol{\gamma}) = \lambda \sum_{k=1}^{K} \|\boldsymbol{\gamma}_k\|_1 + \lambda_g \sum_{r=1}^{d-1} \|\boldsymbol{\gamma}_{(r)}\|_2.$$

The corresponding gradient of ℓ at $\boldsymbol{\gamma}^{\circ}$ is

$$\nabla \ell(\boldsymbol{\gamma}^{\circ}) = (\dot{\ell}^{\mathrm{T}}(\boldsymbol{\gamma}_1^{\circ}), \ldots, \dot{\ell}^{\mathrm{T}}(\boldsymbol{\gamma}_K^{\circ}))^{\mathrm{T}} \in \mathbb{R}^{K+(d-1)K^2},$$

where, for each $k \in [K]$,

$$\dot{\ell}(\boldsymbol{\gamma}_k^{\circ}) = \frac{1}{n} \sum_{i=1}^{n} \left(\frac{\exp(\boldsymbol{\gamma}_k^{\circ \mathrm{T}} \mathbf{x}^{(i)})}{1 + \sum_{l=1}^{K} \exp(\boldsymbol{\gamma}_l^{\circ \mathrm{T}} \mathbf{x}^{(i)})} - y_k^{(i)} \right) \mathbf{x}^{(i)} = \frac{1}{n} \sum_{i=1}^{n} \left(p_k^{(i)}(\boldsymbol{\gamma}^{\circ}) - y_k^{(i)} \right) \mathbf{x}^{(i)} \in \mathbb{R}^{1+(d-1)K},$$

and

$$p_k^{(i)}(\boldsymbol{\gamma}^{\circ}) := P(y_k^{(i)} = 1 | \mathbf{x}^{(i)}) = \frac{\exp(\boldsymbol{\gamma}_k^{\circ \mathrm{T}} \mathbf{x}^{(i)})}{1 + \sum_{l=1}^{K} \exp(\boldsymbol{\gamma}_l^{\circ \mathrm{T}} \mathbf{x}^{(i)})}.$$

Moreover, we introduce the pseudo error $\epsilon_k^{(i)} = y_k^{(i)} - p_k^{(i)}(\boldsymbol{\gamma}^{\circ *})$ for each $k \in [K]$ and $i \in [n]$. Recalling that $\sigma^2 = \max_{i,k} \mathrm{Var}(y_k^{(i)} | \mathbf{x}^{(i)})$, by Assumption 2, we have $\sigma^2 = \max_{i,k} p_k^{(i)}(\boldsymbol{\gamma}^{\circ *})[1 - p_k^{(i)}(\boldsymbol{\gamma}^{\circ *})]$.

A.1.2 From Zero Sub-gradient Condition

By the zero sub-gradient condition, we have

$$\nabla \ell(\widehat{\boldsymbol{\gamma}}^{\circ}) + v = 0, \tag{A.2}$$

where v is a sub-gradient of h° at $\widehat{\boldsymbol{\gamma}}^\circ$. Since h° is convex, by the definition of a sub-gradient, we further have $v^{\mathrm{T}}(\boldsymbol{\gamma}^{\circ *} - \widehat{\boldsymbol{\gamma}}^\circ) \leq h^\circ(\boldsymbol{\gamma}^{\circ *}) - h^\circ(\widehat{\boldsymbol{\gamma}}^\circ)$, which, together with Equation (A.2), results in

$$[\nabla \ell(\widehat{\boldsymbol{\gamma}}^\circ)]^{\mathrm{T}} (\widehat{\boldsymbol{\gamma}}^\circ - \boldsymbol{\gamma}^{\circ *}) + h(\widehat{\boldsymbol{\gamma}}^\circ) \leq h(\boldsymbol{\gamma}^{\circ *}). \tag{A.3}$$

For the first term on the left-hand side of Equation (A.3), we have

$$\begin{aligned}
[\nabla \ell(\widehat{\boldsymbol{\gamma}}^\circ)]^{\mathrm{T}} (\widehat{\boldsymbol{\gamma}}^\circ - \boldsymbol{\gamma}^{\circ *}) &= \sum_{k=1}^{K} \left\{ \frac{1}{n} \sum_{i=1}^{n} \left(p_k^{(i)}(\widehat{\boldsymbol{\gamma}}^\circ) - y_k^{(i)} \right) (\mathbf{x}^{(i)})^{\mathrm{T}} (\widehat{\boldsymbol{\gamma}}_k^\circ - \boldsymbol{\gamma}_k^{\circ *}) \right\} \\
&= \sum_{k=1}^{K} \left\{ \frac{1}{n} \sum_{i=1}^{n} \left(p_k^{(i)}(\widehat{\boldsymbol{\gamma}}^\circ) - p_k^{(i)}(\boldsymbol{\gamma}^{\circ *}) - \epsilon_k^{(i)} \right) (\mathbf{x}^{(i)})^{\mathrm{T}} (\widehat{\boldsymbol{\gamma}}_k^\circ - \boldsymbol{\gamma}_k^{\circ *}) \right\} \\
&= \sum_{k=1}^{K} \left\{ \frac{1}{n} \sum_{i=1}^{n} \delta_k^{(i)} (\mathbf{x}^{(i)})^{\mathrm{T}} (\widehat{\boldsymbol{\gamma}}_k^\circ - \boldsymbol{\gamma}_k^{\circ *}) - \frac{1}{n} \epsilon_k^{\mathrm{T}} \mathbf{X} (\widehat{\boldsymbol{\gamma}}_k^\circ - \boldsymbol{\gamma}_k^{\circ *}) \right\},
\end{aligned} \tag{A.4}$$

where, for each $k \in [K]$, $\widehat{\boldsymbol{\gamma}}_k^\circ$ and $\boldsymbol{\gamma}_k^{*\circ}$ are defined similarly to Equation (A.1), $\delta_k^{(i)} = p_k^{(i)}(\widehat{\boldsymbol{\gamma}}^\circ) - p_k^{(i)}(\boldsymbol{\gamma}^{\circ *})$, and

$$\epsilon_k = \begin{pmatrix} \epsilon_k^{(1)} \\ \vdots \\ \epsilon_k^{(n)} \end{pmatrix} \in \mathbb{R}^n, \quad \mathbf{X} = \begin{pmatrix} (\mathbf{x}^{(1)})^{\mathrm{T}} \\ \vdots \\ (\mathbf{x}^{(n)})^{\mathrm{T}} \end{pmatrix} \in \mathbb{R}^{n \times (1 + (d-1)K)}. \tag{A.5}$$

Inserting Equation (A.4) into Equation (A.3) leads to

$$\sum_{k=1}^{K} \left\{ \frac{1}{n} \sum_{i=1}^{n} \delta_k^{(i)} (\mathbf{x}^{(i)})^{\mathrm{T}} (\widehat{\boldsymbol{\gamma}}_k^\circ - \boldsymbol{\gamma}_k^{\circ *}) \right\} + h(\widehat{\boldsymbol{\gamma}}^\circ) \leq h(\boldsymbol{\gamma}^{\circ *}) + \sum_{k=1}^{K} \frac{1}{n} \epsilon_k^{\mathrm{T}} \mathbf{X} (\widehat{\boldsymbol{\gamma}}_k^\circ - \boldsymbol{\gamma}_k^{\circ *}). \tag{A.6}$$

Next, we turn to analyze the first term $\Delta := n^{-1} \sum_{i=1}^{n} \sum_{k=1}^{K} \delta_k^{(i)} (\mathbf{x}^{(i)})^{\mathrm{T}} (\widehat{\boldsymbol{\gamma}}_k^\circ - \boldsymbol{\gamma}_k^{\circ *})$ on the left-hand side of Equation (A.6). For each $i \in [n]$, let $\delta^{(i)} = (\delta_1^{(i)}, \ldots, \delta_K^{(i)})^{\mathrm{T}} \in \mathbb{R}^K$ and $\mathbf{b}^{(i)} = ((\mathbf{x}^{(i)})^{\mathrm{T}} (\widehat{\boldsymbol{\gamma}}_k^\circ - \boldsymbol{\gamma}_1^{\circ *}, \ldots, (\mathbf{x}^{(i)})^{\mathrm{T}} (\widehat{\boldsymbol{\gamma}}_k^\circ - \boldsymbol{\gamma}_K^{\circ *})))^{\mathrm{T}} \in \mathbb{R}^K$. Then, we have $\Delta = n^{-1} \sum_{i=1}^{n} (\delta^{(i)})^{\mathrm{T}} \mathbf{b}^{(i)}$. For

each $i \in [n]$ and $k \in [K]$, by the mean-value theorem, one obtains

$$\delta^{(i)} = \begin{pmatrix} p_1^{(i)}(\widehat{\boldsymbol{\gamma}}^\circ) \\ \vdots \\ p_K^{(i)}(\widehat{\boldsymbol{\gamma}}^\circ) \end{pmatrix} - \begin{pmatrix} p_1^{(i)}(\boldsymbol{\gamma}^{\circ *}) \\ \vdots \\ p_K^{(i)}(\boldsymbol{\gamma}^{\circ *}) \end{pmatrix} = \begin{pmatrix} \int_0^1 [\nabla p_1^{(i)}(\boldsymbol{\gamma}_t^\circ)]^{\mathrm{T}} (\widehat{\boldsymbol{\gamma}}^\circ - \boldsymbol{\gamma}^{\circ *}) dt \\ \vdots \\ \int_0^1 [\nabla p_K^{(i)}(\boldsymbol{\gamma}_t^\circ)]^{\mathrm{T}} (\widehat{\boldsymbol{\gamma}}^\circ - \boldsymbol{\gamma}^{\circ *}) dt \end{pmatrix} = \int_0^1 B_i(\boldsymbol{\gamma}_t^\circ) \mathbf{b}^{(i)} dt,$$

where $\boldsymbol{\gamma}_t^\circ = \boldsymbol{\gamma}^{\circ *} - t(\widehat{\boldsymbol{\gamma}}^\circ - \boldsymbol{\gamma}^{\circ *})$ for any $t \in [0,1]$, $\nabla p_k^{(i)}(\boldsymbol{\gamma}_t^\circ)$ is the gradient of $p_k^{(i)}$ evaluated at $\boldsymbol{\gamma}_t^\circ$, and

$$B_i(\boldsymbol{\gamma}^\circ) = \begin{pmatrix} p_1^{(i)}(\boldsymbol{\gamma}^\circ)[1 - p_1^{(i)}(\boldsymbol{\gamma}^\circ)] & -p_1^{(i)}(\boldsymbol{\gamma}^\circ) p_2^{(i)}(\boldsymbol{\gamma}^\circ) & \cdots & -p_1^{(i)}(\boldsymbol{\gamma}^\circ) p_K^{(i)}(\boldsymbol{\gamma}^\circ) \\ -p_2^{(i)}(\boldsymbol{\gamma}^\circ) p_1^{(i)}(\boldsymbol{\gamma}^\circ) & p_2^{(i)}(\boldsymbol{\gamma}^\circ)[1 - p_2^{(i)}(\boldsymbol{\gamma}^\circ)] & \cdots & -p_2^{(i)}(\boldsymbol{\gamma}^\circ) p_K^{(i)}(\boldsymbol{\gamma}^\circ) \\ \vdots & \vdots & \ddots & \vdots \\ -p_K^{(i)}(\boldsymbol{\gamma}^\circ) p_1^{(i)}(\boldsymbol{\gamma}^\circ) & -p_K^{(i)}(\boldsymbol{\gamma}^\circ) p_2^{(i)}(\boldsymbol{\gamma}^\circ) & \cdots & p_K^{(i)}(\boldsymbol{\gamma}^\circ)[1 - p_K^{(i)}(\boldsymbol{\gamma}^\circ)] \end{pmatrix},$$

for any $\boldsymbol{\gamma}^\circ$ and each $i \in [n]$. From Lemma A.2.5, for each $i \in [n]$, we have $\lambda_{\min}(B_i(\boldsymbol{\gamma}_t^\circ)) \geq \min_{k \in [K]} p_k^{(i)}(\boldsymbol{\gamma}_t^\circ) \cdot p_0^{(i)}(\boldsymbol{\gamma}_t^\circ)$, where $p_0^{(i)}(\boldsymbol{\gamma}_t^\circ) = \left[1 + \sum_{l=1}^K \exp(\boldsymbol{\gamma}_{t,l}^{\circ \mathrm{T}} \mathbf{x}^{(i)})\right]^{-1} \geq c_*$ for c_* given in Assumption 2. Therefore,

$$\Delta = \frac{1}{n} \sum_{i=1}^n [\mathbf{b}^{(i)}]^{\mathrm{T}} \int_0^1 B_i(\boldsymbol{\gamma}_t^\circ) dt \mathbf{b}^{(i)} \geq \frac{1}{n} \sum_{i=1}^n [\mathbf{b}^{(i)}]^{\mathrm{T}} \mathbf{b}^{(i)} \int_0^1 \min_{k \in [K]} p_k^{(i)}(\boldsymbol{\gamma}_t^\circ) \cdot p_0^{(i)}(\boldsymbol{\gamma}_t^\circ) dt. \tag{A.7}$$

For each $k \in [K]$, by Lemma A.2.7, we can bound $p_k^{(i)}(\boldsymbol{\gamma}_t^\circ) \cdot p_0^{(i)}(\boldsymbol{\gamma}_t^\circ)$ from below as

$$
\begin{aligned}
p_k^{(i)}(\boldsymbol{\gamma}_t^\circ) \cdot p_0^{(i)}(\boldsymbol{\gamma}_t^\circ) &= \frac{\exp\left((\mathbf{x}^{(i)})^{\mathrm{T}}\boldsymbol{\gamma}_{t,k}^\circ\right)}{\left[1+\sum_{l=1}^K \exp\left((\mathbf{x}^{(i)})^{\mathrm{T}}\boldsymbol{\gamma}_{t,l}^\circ\right)\right]^2} \\
&\geq \frac{\exp\left((\mathbf{x}^{(i)})^{\mathrm{T}}\boldsymbol{\gamma}_k^{\circ *}\right)}{\left[1+\sum_{l=1}^K \exp\left((\mathbf{x}^{(i)})^{\mathrm{T}}\boldsymbol{\gamma}_l^{\circ *}\right)\right]^2} \exp\left(-2t \sum_{l=1}^K \left|(\mathbf{x}^{(i)})^{\mathrm{T}}(\widehat{\boldsymbol{\gamma}}_l^\circ - \boldsymbol{\gamma}_l^{\circ *})\right|\right) \\
&\geq \frac{\exp\left((\mathbf{x}^{(i)})^{\mathrm{T}}\boldsymbol{\gamma}_k^{\circ *}\right)}{\left[1+\sum_{l=1}^K \exp\left((\mathbf{x}^{(i)})^{\mathrm{T}}\boldsymbol{\gamma}_l^{\circ *}\right)\right]^2} \exp\left(-2t \max_{i\in[n]} \sum_{l=1}^K \left|(\mathbf{x}^{(i)})^{\mathrm{T}}(\widehat{\boldsymbol{\gamma}}_l^\circ - \boldsymbol{\gamma}_l^{\circ *})\right|\right) \\
&\geq \min_{i,k} \frac{\exp\left((\mathbf{x}^{(i)})^{\mathrm{T}}\boldsymbol{\gamma}_k^{\circ *}\right)}{\left[1+\sum_{l=1}^K \exp\left((\mathbf{x}^{(i)})^{\mathrm{T}}\boldsymbol{\gamma}_l^{\circ *}\right)\right]^2} \exp\left(-2t \max_{i\in[n]} \sum_{l=1}^K \left|(\mathbf{x}^{(i)})^{\mathrm{T}}(\widehat{\boldsymbol{\gamma}}_l^\circ - \boldsymbol{\gamma}_l^{\circ *})\right|\right) \\
&\geq c_*^2 \exp\left(-2t \max_{i\in[n]} \sum_{l=1}^K \left|(\mathbf{x}^{(i)})^{\mathrm{T}}(\widehat{\boldsymbol{\gamma}}_l^\circ - \boldsymbol{\gamma}_l^{\circ *})\right|\right),
\end{aligned}
$$

where the last inequality is due to Assumption 2. Inserting the above result into Equation (A.7), we obtain a further lower bound for Δ:

$$
\begin{aligned}
\Delta &\geq \frac{1}{n}\sum_{i=1}^n [\mathbf{b}^{(i)}]^{\mathrm{T}}\mathbf{b}^{(i)} \cdot c_*^2 \cdot \int_0^1 \exp\left(-2t \max_{i\in[n]} \sum_{l=1}^K \left|(\mathbf{x}^{(i)})^{\mathrm{T}}(\widehat{\boldsymbol{\gamma}}_l^\circ - \boldsymbol{\gamma}_l^{\circ *})\right|\right) dt \\
&= c_*^2 F_{\max} \cdot \left(\frac{1}{n}\sum_{i=1}^n \sum_{k=1}^K [(\mathbf{x}^{(i)})^{\mathrm{T}}(\widehat{\boldsymbol{\gamma}}_k^\circ - \boldsymbol{\gamma}_k^{\circ *})]^2\right) \\
&= c_*^2 F_{\max} \cdot \left(\sum_{k=1}^K \|\mathbf{X}(\widehat{\boldsymbol{\gamma}}_k^\circ - \boldsymbol{\gamma}_k^{\circ *})\|_n^2\right),
\end{aligned}
$$

where

$$
F_{\max} = \frac{1-\exp\left(-2\max_{i\in[n]}\sum_{l=1}^K \left|(\mathbf{x}^{(i)})^{\mathrm{T}}(\widehat{\boldsymbol{\gamma}}_l^\circ - \boldsymbol{\gamma}_l^{\circ *})\right|\right)}{2\max_{i\in[n]}\sum_{l=1}^K \left|(\mathbf{x}^{(i)})^{\mathrm{T}}(\widehat{\boldsymbol{\gamma}}_l^\circ - \boldsymbol{\gamma}_l^{\circ *})\right|},
$$

and $\|x\|_n^2 = n^{-1}\sum_{j=1}^p x_j^2$ is the empirical norm for any p-dimensional vector $x = (x_1, \ldots, x_p)^{\mathrm{T}} \in \mathbb{R}^p$. Combining above result with Equation (A.6), one obtains

$$
c_*^2 F_{\max} \sum_{k=1}^K \|\mathbf{X}(\widehat{\boldsymbol{\gamma}}_k^\circ - \boldsymbol{\gamma}_k^{\circ *})\|_n^2 + h^\circ(\widehat{\boldsymbol{\gamma}}^\circ) \leq h^\circ(\boldsymbol{\gamma}^{\circ *}) + \sum_{k=1}^K \frac{1}{n}\epsilon_k^{\mathrm{T}}\mathbf{X}(\widehat{\boldsymbol{\gamma}}_k^\circ - \boldsymbol{\gamma}_k^{\circ *}). \tag{A.8}
$$

To proceed, we divide our analysis into two regions: $2\max_{i\in[n]}\sum_{l=1}^{K}\left|(\mathbf{x}^{(i)})^{\mathrm{T}}(\widehat{\boldsymbol{\gamma}}_l^{\circ}-\boldsymbol{\gamma}_l^{\circ *})\right|\leq C$ and $2\max_{i\in[n]}\sum_{l=1}^{K}\left|(\mathbf{x}^{(i)})^{\mathrm{T}}(\widehat{\boldsymbol{\gamma}}_l^{\circ}-\boldsymbol{\gamma}_l^{\circ *})\right|> C$ for some absolute constant $C>0$. In the first region, referred to as the convergence rate region, we derive the ℓ_q, with $q=1,2$, error bounds for the proposed estimator. While in the second region, referred to as the contradiction region, we aim to find a contradiction.

A.1.3 Convergence Rate Region

Assume

$$2\max_{i\in[n]}\sum_{l=1}^{K}\left|(\mathbf{x}^{(i)})^{\mathrm{T}}(\widehat{\boldsymbol{\gamma}}_l^{\circ}-\boldsymbol{\gamma}_l^{\circ *})\right|\leq C, \tag{A.9}$$

for some absolute constant $C>0$. Given Equation (A.9), for each $k\in[K]$,

$$F_{\max}=\int_0^1\exp\left(-2t\max_{i\in[n]}\sum_{l=1}^{K}\left|(\mathbf{x}^{(i)})^{\mathrm{T}}(\widehat{\boldsymbol{\gamma}}_l^{\circ}-\boldsymbol{\gamma}_l^{\circ *})\right|\right)dt\geq\int_0^1\exp\left(-tC\right)\mathrm{d}t=\frac{1-\exp(-C)}{C}.$$

Inserting the above inequality into Equation (A.8), we find

$$C_1c_*^2\sum_{k=1}^{K}\|\mathbf{X}(\widehat{\boldsymbol{\gamma}}_k^{\circ}-\boldsymbol{\gamma}_k^{\circ *})\|_n^2\leq\sum_{k=1}^{K}\frac{1}{n}\epsilon_k^{\mathrm{T}}\mathbf{X}(\widehat{\boldsymbol{\gamma}}_k^{\circ}-\boldsymbol{\gamma}_k^{\circ *})+h^{\circ}(\boldsymbol{\gamma}^{\circ *})-h^{\circ}(\widehat{\boldsymbol{\gamma}}^{\circ}), \tag{A.10}$$

for some absolute constant $C_1>0$. Without loss of generality, we assume $C_1=1$ in our analysis.

Given a constant $\rho\in[0,1/2)$, from Equation (A.10), one has

$$\begin{aligned}
&\rho\lambda\|\widehat{\boldsymbol{\gamma}}^{\circ}-\boldsymbol{\gamma}^{\circ *}\|_1+c_*^2\sum_{k=1}^{K}\|\mathbf{X}(\widehat{\boldsymbol{\gamma}}_k^{\circ}-\boldsymbol{\gamma}_k^{\circ *})\|_n^2\\
&\quad\leq\sum_{k=1}^{K}\frac{1}{n}\epsilon_k^{\mathrm{T}}\mathbf{X}(\widehat{\boldsymbol{\gamma}}_k^{\circ}-\boldsymbol{\gamma}_k^{\circ *})+\rho\lambda\|\widehat{\boldsymbol{\gamma}}^{\circ}-\boldsymbol{\gamma}^{\circ *}\|_1+\sum_{k=1}^{K}\left(\lambda\|\boldsymbol{\gamma}_k^{*}\|_1-\lambda\|\widehat{\boldsymbol{\gamma}}_k\|_1\right)+\lambda_g\sum_{r=1}^{d-1}\left(\|\boldsymbol{\gamma}_{(r)}^{*}\|_2-\|\widehat{\boldsymbol{\gamma}}_{(r)}\|_2\right)\\
&\quad=B_1+B_2,
\end{aligned}$$

where

$$B_1=\sum_{k=1}^{K}\frac{1}{n}\epsilon_k^{\mathrm{T}}\mathbf{X}(\widehat{\boldsymbol{\gamma}}_k^{\circ}-\boldsymbol{\gamma}_k^{\circ *}), \tag{A.11}$$

and

$$
\begin{aligned}
B_2 &= \rho\lambda\|\widehat{\boldsymbol{\gamma}}^{\circ} - \boldsymbol{\gamma}^{\circ *}\|_1 + \sum_{k=1}^{K}\left(\lambda\|\boldsymbol{\gamma}_k^*\|_1 - \lambda\|\widehat{\boldsymbol{\gamma}}_k\|_1\right) + \lambda_g \sum_{r=1}^{d-1}\left(\|\boldsymbol{\gamma}_{(r)}^*\|_2 - \|\widehat{\boldsymbol{\gamma}}_{(r)}\|_2\right) \\
&= \rho\lambda\|\widehat{\boldsymbol{\gamma}}^{\circ} - \boldsymbol{\gamma}^{\circ *}\|_1 + \lambda(\|\boldsymbol{\gamma}^*\|_1 - \|\widehat{\boldsymbol{\gamma}}\|_1) + \lambda_g(\|\boldsymbol{\gamma}^*\|_{1,2} - \|\widehat{\boldsymbol{\gamma}}\|_{1,2}) \\
&= 2\lambda_\sharp\left[(\rho\|\widehat{\boldsymbol{\gamma}}^{\circ} - \boldsymbol{\gamma}^{\circ *}\|_1 + \|\boldsymbol{\gamma}^*\|_1 - \|\widehat{\boldsymbol{\gamma}}\|_1) + \sqrt{s_0}(\|\boldsymbol{\gamma}^*\|_{1,2} - \|\widehat{\boldsymbol{\gamma}}\|_{1,2})\right], \qquad \text{(A.12)}
\end{aligned}
$$

where $\|\boldsymbol{\gamma}^*\|_{1,2} = \sum_{r=1}^{d-1}\|\boldsymbol{\gamma}_{(r)}^*\|_2$ and similarly for $\|\widehat{\boldsymbol{\gamma}}\|_{1,2}$, $s_0 = s/s_g$, and $\lambda_\sharp$ is defined in Equation (2.11) for a given absolute constant $\eta \in (0, 1/2)$.

For any vector $x \in \mathbb{R}^p$, denote $x_{j\sharp}$ as the jth largest element in x for $j \in [p]$. Let $u = \widehat{\boldsymbol{\gamma}}^{\circ} - \boldsymbol{\gamma}^{\circ *} \in \mathbb{R}^{K+(d-1)K^2}$. Note that u consists of d groups, denoted as $\{u_{(r)} : r \in [d-1]_0\}$. In these groups, $u_{(0)} \in \mathbb{R}^K$ corresponds to the intercept vector $\boldsymbol{\theta}$ and $u_{(r)} \in \mathbb{R}^{K^2}$ corresponds to the coefficient vector $\boldsymbol{\gamma}_{(r)}$ of group r for each $r \in [d-1]$. Augment $u_{(0)}$ with zeros such that $\tilde{u}_{(0)} = (u_{(0)}^{\mathrm{T}}, 0, \ldots, 0) \in \mathbb{R}^{K^2}$. Then, combining d groups of vectors $\{\tilde{u}_{(0)}, u_{(r)} : r \in [d-1]\}$ into $\tilde{u} \in \mathbb{R}^{\tilde{D}}$ with $\tilde{D} = dK^2$, we have $\|\tilde{u}\|_q = \|u\|_q$ for $q = 1, 2$. Since $\tilde{u}_{(0)}$ and all $u_{(r)}$ for $r \in [d-1]_0$ are of the same length, we can set $U \in \mathbb{R}^{K^2 \times d}$, referred to as the group matrix of $\tilde{u}$, to be the matrix whose columns are formed by d groups in $\tilde{u}$. Let $\|U_{j\sharp}\|_2$ be the jth largest element of ℓ_2-norms of columns in U for $j \in [d]$. Moreover, recall that $\tilde{u}_{j\sharp}$ is the jth largest element in $\tilde{u}$ for $j \in [\tilde{D}]$. For each $k \in [K]$, we set

$$
\tilde{u}_k = \begin{pmatrix} \mathbf{0}_{K-1} \\ \widehat{\boldsymbol{\gamma}}_k^{\circ} \end{pmatrix} - \begin{pmatrix} \mathbf{0}_{K-1} \\ \boldsymbol{\gamma}_k^{\circ *} \end{pmatrix} \in \mathbb{R}^{dK}, \qquad \text{(A.13)}
$$

where $\mathbf{0}_{K-1}$ is the zero vector in $\mathbb{R}^{K-1}$.

With Lemma A.2.1 and by applying the same arguments in the proof of Theorem 3 in Li et al. (2023), it follows that, given $u = \widehat{\boldsymbol{\gamma}}^{\circ} - \boldsymbol{\gamma}^{\circ *}$ and the corresponding augmented vector $\tilde{u}$ defined via Equation (A.13), we have

$$
B_1 \leq \max\{F(u), G(u)\}, \qquad \text{(A.14)}
$$

where

$$F(u) = 40\eta\lambda_\sharp \left[2\sqrt{s^\circ}\|\tilde{u}\|_2 + \left(\sum_{j=s^\circ+1}^{\tilde{D}} \tilde{u}_{j^\sharp} + \sum_{j=s_g^\circ+1}^{d} \sqrt{s_0^\circ}\|U_{j^\sharp}\|_2 \right) \right],$$

and

$$G(u) = 40\sigma \left(\sum_{k=1}^{K} \|\tilde{\mathbf{X}}\tilde{u}_k\|_n^2 \right)^{1/2} \frac{2\sqrt{\log(1/\delta_0)}}{\sqrt{n}} = 2\eta\lambda_\sharp\sqrt{s^\circ}\sqrt{\frac{\log(1/\delta_0)}{s^\circ \log(1/\delta(\lambda_\sharp))}} \left(\sum_{k=1}^{K} \|\tilde{\mathbf{X}}\tilde{u}_k\|_n^2 \right)^{1/2}$$
$$= 2\eta\lambda_\sharp\sqrt{s^\circ}\sqrt{\frac{\log(1/\delta_0)}{s^\circ \log(1/\delta(\lambda_\sharp))}} \left(\sum_{k=1}^{K} \|\mathbf{X}u_k\|_n^2 \right)^{1/2},$$

with $\tilde{\mathbf{X}} = [\mathbf{0}_{n\times(K-1)}, \mathbf{X}] \in \mathbb{R}^{n\times dK}$ for the zero matrix $\mathbf{0}_{n\times(K-1)}$ of the shape $n \times (K-1)$, $u_k = \widehat{\boldsymbol{\gamma}}_k^\circ - \boldsymbol{\gamma}_k^{\circ *}$ and

$$\delta(\lambda_\sharp) \triangleq \exp\left(-\left(\frac{\eta\lambda_\sharp\sqrt{n}}{40} \right)^2 \right) \quad \text{so that} \quad \lambda_\sharp = \frac{40}{\eta}\sqrt{\frac{\log(1/\delta(\lambda_\sharp))}{n}}.$$

for any given absolute constant $\eta \in (0, 1/2)$.

For B_2 in Equation (A.12), letting $o = \widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*$ and $v = \widehat{\boldsymbol{\gamma}} - \boldsymbol{\gamma}^* \in \mathbb{R}^D$ with the corresponding group matrix $V \in \mathbb{R}^{K^2\times(d-1)}$, and recalling that $\tilde{u}$ is augmented from $\widehat{\boldsymbol{\gamma}}^\circ - \boldsymbol{\gamma}^{\circ *}$ by adding zeros,

one has

$$
\begin{aligned}
B_2 &= 2\lambda_\sharp \left[(\rho\|\widehat{\boldsymbol{\gamma}}^\circ - \boldsymbol{\gamma}^{\circ *}\|_1 + \|\boldsymbol{\gamma}^*\|_1 - \|\widehat{\boldsymbol{\gamma}}\|_1) + \sqrt{s_0}(\|\boldsymbol{\gamma}^*\|_{1,2} - \|\widehat{\boldsymbol{\gamma}}\|_{1,2})\right] \\
&= 2\lambda_\sharp \left[\rho\|\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*\|_1 + (\rho\|\widehat{\boldsymbol{\gamma}} - \boldsymbol{\gamma}^*\|_1 + \|\boldsymbol{\gamma}^*\|_1 - \|\widehat{\boldsymbol{\gamma}}\|_1) + \sqrt{s_0}(\|\boldsymbol{\gamma}^*\|_{1,2} - \|\widehat{\boldsymbol{\gamma}}\|_{1,2})\right] \\
&\leq 2\lambda_\sharp \left[\rho\sqrt{K}\|o\|_2 + \left((1+\rho)\sqrt{s}\|v\|_2 - (1-\rho)\sum_{j=s+1}^{D} v_{j^\sharp}\right) + \sqrt{s_0}(\|\boldsymbol{\gamma}^*\|_{1,2} - \|\widehat{\boldsymbol{\gamma}}\|_{1,2})\right] \\
&\leq 2\lambda_\sharp \left[\rho\sqrt{K}\|o\|_2 + \left((1+\rho)\sqrt{s}\|v\|_2 - (1-\rho)\sum_{j=s+1}^{D} v_{j^\sharp}\right) + \left(\sqrt{s_g s_0}\|v\|_2 - \sum_{j=s_g+1}^{d-1}\sqrt{s_0}\|V_{j^\sharp}\|_2\right)\right] \\
&= 2\lambda_\sharp \left[\left(\rho\sqrt{K}\|o\|_2 + (1+\rho)\sqrt{s}\|v\|_2 + \sqrt{s_g s_0}\|v\|_2\right) - \left((1-\rho)\sum_{j=s+1}^{D} v_{j^\sharp} + \sum_{j=s_g+1}^{d-1}\sqrt{s_0}\|V_{j^\sharp}\|_2\right)\right] \\
&\leq 2\lambda_\sharp \left[\rho\sqrt{K}\|o\|_2 + (2+\rho)\sqrt{s}\|v\|_2 - \left((1-\rho)\sum_{j=s+1}^{D} v_{j^\sharp} + \sum_{j=s_g+1}^{d-1}\sqrt{s_0}\|V_{j^\sharp}\|_2\right)\right] \\
&\leq 2\lambda_\sharp \left[(2+\rho)\sqrt{s^\circ}\|\tilde{u}\|_2 - \left((1-\rho)\sum_{j=s+1}^{D} v_{j^\sharp} + \sum_{j=s_g+1}^{d-1}\sqrt{s_0}\|V_{j^\sharp}\|_2\right)\right], \qquad \text{(A.15)}
\end{aligned}
$$

where the first inequality is due to Lemma A.1 in Bellec et al. (2018), the second inequality is from Lemma 3 in Li et al. (2023), and the last inequality is from a basic inequality that

$$\sqrt{K}\|o\|_2 + \sqrt{s}\|v\|_2 \leq \sqrt{K+s}\|(o^{\mathrm{T}}, v^{\mathrm{T}})^{\mathrm{T}}\|_2 = \sqrt{s^\circ}\|\tilde{u}\|_2.$$

Therefore, combining Equation (A.14) and Equation (A.15), we have

$$
\begin{aligned}
&\rho\lambda\|\widehat{\boldsymbol{\gamma}}^\circ - \boldsymbol{\gamma}^{\circ *}\|_1 + c_*^2\sum_{k=1}^{K}\|\mathbf{X}(\widehat{\boldsymbol{\gamma}}_k^\circ - \boldsymbol{\gamma}_k^{\circ *})\|_n^2 \\
&\quad\leq \max\{F(u), G(u)\} + 2\lambda_\sharp \left[(2+\rho)\sqrt{s^\circ}\|\tilde{u}\|_2 - \left((1-\rho)\sum_{j=s+1}^{D} v_{j^\sharp} + \sum_{j=s_g+1}^{d-1}\sqrt{s_0}\|V_{j^\sharp}\|_2\right)\right]. \qquad \text{(A.16)}
\end{aligned}
$$

(Case $F(u) \leq G(u)$). In this case, we have

$$\|u\|_2 = \|\tilde{u}\|_2 \leq \sqrt{\frac{\log(1/\delta_0)}{s^\circ \log(1/\delta(\lambda_\sharp))}} \left(\sum_{k=1}^{K} \|\mathbf{X} u_k\|_n^2 \right)^{1/2}. \tag{A.17}$$

Inserting Equation (A.17) into Equation (A.16) gives

$$\left(\sum_{k=1}^{K} \|\mathbf{X} u_k\|_n^2 \right)^{1/2} \leq \frac{2}{c_*^2}(2+\rho+\eta)\lambda_\sharp \sqrt{s^\circ} \sqrt{\frac{\log(1/\delta_0)}{s^\circ \log(1/\delta(\lambda_\sharp))}},$$

which further implies that

$$\begin{aligned} \|u\|_2 &\leq \frac{2}{c_*^2}(2+\rho+\eta)\frac{\log(1/\delta_0)}{s^\circ \log(1/\delta(\lambda_\sharp))}\sqrt{s^\circ}\lambda_\sharp \\ &\leq \frac{C_1}{c_*^2}\frac{\log(1/\delta_0)}{s^\circ \log(1/\delta(\lambda_\sharp))}\sqrt{s^\circ}\lambda_\sharp, \end{aligned} \tag{A.18}$$

where the first inequality is from Equation (A.17). Furthermore, from Equation (A.16), one also has

$$\begin{aligned} \rho\lambda\|\widehat{\boldsymbol{\gamma}}^\circ - \boldsymbol{\gamma}^{\circ *}\|_1 &\leq 2\lambda_\sharp(2+\rho+\eta)\sqrt{s^\circ}\sqrt{\frac{\log(1/\delta_0)}{s^\circ \log(1/\delta(\lambda_\sharp))}} \left(\sum_{k=1}^{K} \|\mathbf{X} u_k\|_n^2 \right)^{1/2} \\ &\leq \frac{(2\lambda_\sharp)^2}{c_*^2}(2+\rho+\eta)^2 s^\circ \frac{\log(1/\delta_0)}{s^\circ \log(1/\delta(\lambda_\sharp))}. \end{aligned}$$

We then deduce that

$$\begin{aligned} \|\widehat{\boldsymbol{\gamma}}^\circ - \boldsymbol{\gamma}^{\circ *}\|_1 &\leq \frac{2}{\rho c_*^2}(2+\rho+\eta)^2 \frac{\log(1/\delta_0)}{s^\circ \log(1/\delta(\lambda_\sharp))} s^\circ \lambda_\sharp \\ &\leq \frac{C_2}{c_*^2}\frac{\log(1/\delta_0)}{s^\circ \log(1/\delta(\lambda_\sharp))} s^\circ \lambda_\sharp. \end{aligned} \tag{A.19}$$

(Case $F(u) > G(u)$). With $D = (d-1)K^2$ and $\tilde{D} = dK^2$, note that $v = \widehat{\boldsymbol{\gamma}} - \boldsymbol{\gamma}^* \in \mathbb{R}^D$ is contained in $\tilde{u} \in \mathbb{R}^{\tilde{D}}$ which is augmented from $\widehat{\boldsymbol{\gamma}}^\circ - \boldsymbol{\gamma}^{\circ *} \in \mathbb{R}^{K+(d-1)K^2}$ by adding zeros, and $s^\circ = s + K$ and $d - s_g^\circ = (d-1) - s_g$. Hence, $\sum_{j=s^\circ+1}^{\tilde{D}} \tilde{u}_{j^\sharp} \leq \sum_{j=s+1}^{D} v_{j^\sharp}$ and $\sum_{j=s_g^\circ+1}^{d} \|U_{j^\sharp}\|_2 \leq \sum_{j=s_g+1}^{d-1} \|V_{j^\sharp}\|_2$.

Consequently, when $F(u) > G(u)$, we have

$$\rho\lambda\|\widehat{\boldsymbol{\gamma}}^{\circ} - \boldsymbol{\gamma}^{\circ *}\|_1 + c_*^2 \sum_{k=1}^{K} \|\mathbf{X}(\widehat{\boldsymbol{\gamma}}_k - \boldsymbol{\gamma}_k^*)\|_n^2 \\ \leq \lambda_\sharp \left[2(2+\rho+\eta)\sqrt{s^\circ}\|\tilde{u}\|_2 - (2-\eta-2\rho)\left(\sum_{j=s+1}^{D} v_{j^\sharp}\right) - (2-\eta)\left(\sum_{j=s_g+1}^{d-1} \sqrt{s_0}\|V_{j^\sharp}\|_2\right)\right]. \tag{A.20}$$

Therefore, since $\rho \in [0, 1/2)$ and $\eta \in (0, 1/2)$, one obtains

$$\frac{2+\rho+\eta}{1-\rho-\eta}\sqrt{s^\circ}\|\tilde{u}\|_2 \geq \sum_{j=s+1}^{D} v_{j^\sharp} + \sum_{j=s_g+1}^{d-1} \sqrt{s_0}\|V_{j^\sharp}\|_2 \geq \sum_{j=s^\circ+1}^{\tilde{D}} \tilde{u}_{j^\sharp}.$$

This implies that $\tilde{u}$ belongs to a cone $\mathcal{C}_{SRE}(s^\circ, c')$, as defined right before Lemma A.2.3, for an absolute constant $c' > 0$. Consequently, by Lemma A.2.3 and Equation (A.20), with probability at least $1 - c((d-1)K+1)^{-1}$, we have

$$\|u\|_2^2 = \|\tilde{u}\|_2^2 = \sum_{k=1}^{K} \|\tilde{u}_k\|_2^2 \leq \frac{\sum_{k=1}^{K} \|\tilde{\mathbf{X}}\tilde{u}_k\|_n^2}{\lambda_{\min}(\boldsymbol{\Sigma})} = \frac{\sum_{k=1}^{K} \|\mathbf{X}u_k\|_n^2}{\lambda_{\min}(\boldsymbol{\Sigma})} \leq \frac{2(2+\eta+\rho)\sqrt{s^\circ}\|u\|_2\lambda_\sharp}{\lambda_{\min}(\boldsymbol{\Sigma})c_*^2},$$

which further leads to

$$\|u\|_2 \leq \frac{C_3}{c_*^2}\frac{\sqrt{s^\circ}\lambda_\sharp}{\lambda_{\min}(\boldsymbol{\Sigma})}. \tag{A.21}$$

Meanwhile, from Equation (A.20), one also has

$$\rho\lambda\|\widehat{\boldsymbol{\gamma}}^{\circ} - \boldsymbol{\gamma}^{\circ *}\|_1 \leq 2\lambda_\sharp(2+\eta+\rho)\sqrt{s^\circ}\|u\|_2.$$

Inserting Equation (A.21) into above inequality, we obtain that

$$\|\widehat{\boldsymbol{\gamma}}^{\circ} - \boldsymbol{\gamma}^{\circ *}\|_1 \leq \frac{C_4}{c_*^2}\frac{s^\circ\lambda_\sharp}{\rho\lambda_{\min}(\boldsymbol{\Sigma})}. \tag{A.22}$$

Hence, combining Equation (A.18) and Equation (A.21), the ℓ_2-error bound of $\widehat{\boldsymbol{\gamma}}$ is

$$\|\widehat{\boldsymbol{\gamma}}^{\circ} - \boldsymbol{\gamma}^{\circ *}\|_2 \leq \frac{C'}{c_*^2}\sqrt{s^{\circ}}\lambda_{\sharp} \max\left\{\frac{\log(1/\delta_0)}{s^{\circ}\log(1/\delta(\lambda_{\sharp}))}, \frac{1}{\lambda_{\min}(\boldsymbol{\Sigma})}\right\}.$$

Moreover, combining Equation (A.19) and Equation (A.22), the ℓ_1-error bound of $\widehat{\boldsymbol{\gamma}}$ is

$$\|\widehat{\boldsymbol{\gamma}}^{\circ} - \boldsymbol{\gamma}^{\circ *}\|_1 \leq \frac{C'}{c_*^2}s^{\circ}\lambda_{\sharp} \max\left\{\frac{\log(1/\delta_0)}{s^{\circ}\log(1/\delta(\lambda_{\sharp}))}, \frac{1}{\lambda_{\min}(\boldsymbol{\Sigma})}\right\}.$$

Simplifying the above results with the $\lambda_{\sharp}$ given in Equation (2.11), we obtain that, with probability displayed in Theorem 2.3.1,

$$\|\widehat{\boldsymbol{\gamma}}^{\circ} - \boldsymbol{\gamma}^{\circ *}\|_q \leq \frac{C}{c_*^2}(s^{\circ})^{1/q}\lambda_{\sharp} \max\left\{\frac{\log(1/\delta_0)}{s^{\circ}\log(1/\delta(\lambda_{\sharp}))}, \frac{1}{\lambda_{\min}(\boldsymbol{\Sigma})}\right\} \asymp R_K B_n,$$

for $q = 1, 2$, where $R_K = \sigma/(c_\lambda c_*^2)$ and

$$B_n = \frac{(s^{\circ})^{1/q}}{\sqrt{n}}\sqrt{\frac{2}{s_0^{\circ}}\log\left(\frac{4ed}{s_g^{\circ}}\right) + \log\left(\frac{2eK^2}{s_0^{\circ}}\right)}.$$

A.1.4 Contradiction Region

Assume

$$2\max_{i\in[n]}\sum_{l=1}^{K}\left|(\mathbf{x}^{(i)})^{\mathrm{T}}(\widehat{\boldsymbol{\gamma}}_l^{\circ} - \boldsymbol{\gamma}_l^{\circ *})\right| \leq C, \tag{A.23}$$

for some constant $C > 0$. Similar to Equation (A.16), from Equation (A.8), one has

$$\begin{aligned}
&\rho\lambda\|\widehat{\boldsymbol{\gamma}}^{\circ} - \boldsymbol{\gamma}^{\circ *}\|_1 + c_*^2 F_{\max}\sum_{k=1}^{K}\|\mathbf{X}(\widehat{\boldsymbol{\gamma}}_k^{\circ} - \boldsymbol{\gamma}_k^{\circ *})\|_n^2 \\
&\quad\leq \max\{F(u), G(u)\} + 2\lambda_{\sharp}\left[(2+\rho)\sqrt{s^{\circ}}\|\tilde{u}\|_2 - \left((1-\rho)\sum_{j=s+1}^{D} v_{j^{\sharp}} + \sum_{j=s_g+1}^{d-1}\sqrt{s_0}\|V_{j^{\sharp}}\|_2\right)\right],
\end{aligned}$$

for any given constant $\rho \in [0, 1/2)$, and $F(\cdot)$ and $G(\cdot)$ introduced in Equation (A.14). Recalling that

$$F_{\max} = \frac{1 - \exp\left(-2 \max_{i\in[n]} \sum_{l=1}^{K} \left|(\mathbf{x}^{(i)})^{\mathrm{T}}(\widehat{\boldsymbol{\gamma}}_l^{\circ} - \boldsymbol{\gamma}_l^{\circ *})\right|\right)}{2 \max_{i\in[n]} \sum_{l=1}^{K} |(\mathbf{x}^{(i)})^{\mathrm{T}}(\widehat{\boldsymbol{\gamma}}_l^{\circ} - \boldsymbol{\gamma}_l^{\circ *})|},$$

and with the assumption Equation (A.23), for each $k \in [K]$, we have

$$F_{\max} \geq \frac{1 - \exp(-C)}{2 \max_{i\in[n]} \sum_{l=1}^{K} |(\mathbf{x}^{(i)})^{\mathrm{T}}(\widehat{\boldsymbol{\gamma}}_l^{\circ} - \boldsymbol{\gamma}_l^{\circ *})|},$$

which implies

$$\begin{aligned}
&\rho\lambda\|\widehat{\boldsymbol{\gamma}}^{\circ} - \boldsymbol{\gamma}^{\circ *}\|_1 + c_*^2 F_{\max} \sum_{k=1}^{K} \|\mathbf{X}(\widehat{\boldsymbol{\gamma}}_k^{\circ} - \boldsymbol{\gamma}_k^{\circ *})\|_n^2 \\
&\quad \leq \max\{F(u), G(u)\} + 2\lambda_\sharp \left[(2+\rho)\sqrt{s^\circ}\|\tilde{u}\|_2 - \left((1-\rho)\sum_{j=s+1}^{D} v_{j^\sharp} + \sum_{j=s_g+1}^{d-1} \sqrt{s_0}\|V_{j^\sharp}\|_2\right)\right].
\end{aligned} \tag{A.24}$$

(Case $F(u) \leq G(u)$). In this case, inserting Equation (A.17) into Equation (A.24) gives

$$\left(\sum_{k=1}^{K} \|\mathbf{X}u_k\|_n^2\right)^{1/2} \leq \frac{2}{c_*^2}(2+\rho+\eta)\lambda_\sharp\sqrt{s^\circ}\sqrt{\frac{\log(1/\delta_0)}{s^\circ \log(1/\delta(\lambda_\sharp))}}F_{\max}^{-1},$$

Meanwhile, from Equation (A.24), one also has

$$\begin{aligned}
\rho\lambda\|\widehat{\boldsymbol{\gamma}}^{\circ} - \boldsymbol{\gamma}^{\circ *}\|_1 &\leq 2\lambda_\sharp(2+\rho+\eta)\sqrt{s^\circ}\sqrt{\frac{\log(1/\delta_0)}{s^\circ \log(1/\delta(\lambda_\sharp))}}\left(\sum_{k=1}^{K} \|\mathbf{X}u_k\|_n^2\right)^{1/2} \\
&\leq \frac{(2\lambda_\sharp)^2}{c_*^2}(2+\rho+\eta)^2 s^\circ \frac{\log(1/\delta_0)}{s^\circ \log(1/\delta(\lambda_\sharp))}F_{\max}^{-1},
\end{aligned}$$

which leads to

$$\begin{aligned}\|\widehat{\boldsymbol{\gamma}}^{\circ} - \boldsymbol{\gamma}^{\circ *}\|_1 &\leq \frac{C_1}{c_*^2} \frac{\log(1/\delta_0)}{s^{\circ} \log(1/\delta(\lambda_\sharp))} s^{\circ} \lambda_\sharp F_{\max}^{-1} \\ &\leq \frac{C_2}{c_*^2} s^{\circ} \lambda_\sharp F_{\max}^{-1} \\ &\leq \frac{C_3}{c_*^2} s^{\circ} \lambda_\sharp \max_{i,l} |\mathbf{x}_l^{(i)}| \sum_{k=1}^{K} \|\widehat{\boldsymbol{\gamma}}_k^{\circ} - \boldsymbol{\gamma}_k^{\circ *}\|_1 \\ &= \frac{C_3}{c_*^2} s^{\circ} \lambda_\sharp \max_{i,l} |\mathbf{x}_l^{(i)}| \|\widehat{\boldsymbol{\gamma}}^{\circ} - \boldsymbol{\gamma}^{\circ *}\|_1,\end{aligned}$$

where the second inequality is due to the assumption of $\frac{\log(1/\delta_0)}{s^{\circ} \log(1/\delta(\lambda_\sharp))} = O(1)$. The above inequality suggests that, if Equation (A.23) holds and $F(u) \leq G(u)$, then

$$\frac{s^{\circ}}{c_*^2} \lambda_\sharp \max_{i,l} |\mathbf{x}_l^{(i)}| \geq C_0,$$

for some absolute constant $C_0 > 0$. However, since $s^{\circ} \ll c_*^2 \sqrt{n/\log(dK)}$ by assumption and $|\mathbf{x}_l^{(i)}| \leq 1$, the left-hand side of the above inequality is of the order $o(1)$, resulting a contradiction.

(Case $F(u) > G(u)$). Similar to Equation (A.20), when $F(u) > G(u)$, we have

$$\begin{aligned}&\rho\lambda \|\widehat{\boldsymbol{\gamma}}^{\circ} - \boldsymbol{\gamma}^{\circ *}\|_1 + c_*^2 F_{\max} \sum_{k=1}^{K} \|\mathbf{X}(\widehat{\boldsymbol{\gamma}}_k^{\circ} - \boldsymbol{\gamma}_k^{\circ *})\|_n^2 \\ &\quad \leq \lambda_\sharp \left[2(2+\rho+\eta)\sqrt{s^{\circ}} \|\tilde{u}\|_2 - (2-\eta-2\rho) \left(\sum_{j=s+1}^{D} v_{j^\sharp} \right) - (2-\eta) \left(\sum_{j=s_g+1}^{d-1} \sqrt{s_0} \|V_{j^\sharp}\|_2 \right) \right],\end{aligned}$$

which also implies that $\tilde{u}$ belongs to a cone $\mathcal{C}_{SRE}(s^{\circ}, c')$, as defined right before Lemma A.2.3, for an absolute constant $c' > 0$. Hence, from Lemma A.2.3, with probability at least $1 - c((d-1)K + 1)^{-1}$, one has

$$\|\widehat{\boldsymbol{\gamma}}^{\circ} - \boldsymbol{\gamma}^{\circ *}\|_1 \leq (1+c')\sqrt{s^{\circ}} \|\widehat{\boldsymbol{\gamma}}^{\circ} - \boldsymbol{\gamma}^{\circ *}\|_2,$$

and

$$c_*^2 F_{\max} \lambda_{\min}(\boldsymbol{\Sigma}) \|\widehat{\boldsymbol{\gamma}}^{\circ} - \boldsymbol{\gamma}^{\circ *}\|_2^2 \leq 2\lambda_\sharp (2+\rho+\eta)\sqrt{s^{\circ}} \|\widehat{\boldsymbol{\gamma}}^{\circ} - \boldsymbol{\gamma}^{\circ *}\|_2.$$

Combining these two inequalities, we obtain that

$$\begin{aligned}\|\widehat{\boldsymbol{\gamma}}^{\circ} - \boldsymbol{\gamma}^{\circ *}\|_2 &\leq \frac{C_1}{c_*^2}\sqrt{s^{\circ}}\lambda_{\sharp}F_{\max}^{-1} \\ &\leq \frac{C_2}{c_*^2}\sqrt{s^{\circ}}\lambda_{\sharp}\max_{i,l}|\mathbf{x}_l^{(i)}|\|\widehat{\boldsymbol{\gamma}}^{\circ} - \boldsymbol{\gamma}^{\circ *}\|_1 \\ &\leq \frac{C_3}{c_*^2}s^{\circ}\lambda_{\sharp}\max_{i,l}|\mathbf{x}_l^{(i)}|\|\widehat{\boldsymbol{\gamma}}^{\circ} - \boldsymbol{\gamma}^{\circ *}\|_2.\end{aligned}$$

This also results in

$$\frac{s^{\circ}}{c_*^2}\lambda_{\sharp}\max_{i,l}|\mathbf{x}_l^{(i)}| \geq C_0,$$

for some absolute constant $C_0 > 0$, leading to a contradiction as the previous case.

A.2 Auxiliary Results

The following lemma is used to analyze B_1 defined in Equation (A.11), for a generic vector $\tilde{u} \in \mathbb{R}^{\tilde{D}}$ with $\tilde{D} = dK^2$ and the corresponding the group matrix $U \in \mathbb{R}^{K^2 \times d}$, we introduce the function $N(\tilde{u})$ as follows:

$$N(\tilde{u}) = \frac{1}{\sqrt{n}}\left(\sum_{j=1}^{\tilde{D}}\tilde{u}_{j\sharp}\tilde{\lambda}_j + \sum_{j=1}^{d}\|U_{j\sharp}\|_2\sqrt{s_0^{\circ}}\lambda_j\right), \tag{A.25}$$

where, $0 \leq s_0^{\circ} \leq K^2$,

$$\lambda_j = \sigma\sqrt{\log\frac{2eK^2}{s_0^{\circ}} + \frac{2}{s_0^{\circ}}\log\frac{4d}{j}}, \quad \forall j \in [d],$$

and

$$\tilde{\lambda}_j = \begin{cases}\lambda_{\lfloor j/s_0^{\circ}\rfloor} & j \leq s_0^{\circ}d, \\ \lambda_d & j > s_0^{\circ}d,\end{cases}$$

for $\sigma^2 = \max_{i,k}\text{Var}(y_k^{(i)}|\mathbf{x}^{(i)})$ and c_* given in Assumption 2. By this definition, $N(u)$ is positive homogeneous, that is, $N(a\tilde{u}) = aN(\tilde{u})$ for all $a \geq 0, \tilde{u} \in \mathbb{R}^{\tilde{D}}$ and $N(\tilde{u}) > 0$ for $u \neq 0$. Moreover,

let $\tilde{\mathbf{X}} = [\mathbf{0}_{n\times(K-1)}, \mathbf{X}] \in \mathbb{R}^{n\times dK}$ with $\mathbf{X}$ specified in Equation (A.5). With these notations, we propose the following lemma.

Lemma A.2.1. *Consider $N(\tilde{u})$ in Equation* (A.25) *and let $\delta_0 \in (0,1)$. Then, under Assumption 2, for $\{\epsilon_k, k = 1,\ldots,K\}$ and $\tilde{\mathbf{X}}$ given above, with probability at least $1-\delta_0$ we have, for all $u = (u_1^{\mathrm{T}},\ldots,u_K^{\mathrm{T}})^{\mathrm{T}}$ with each $\tilde{u}_k \in \mathbb{R}^{dK}$,*

$$\sum_{k=1}^{K} \frac{1}{n}\epsilon_k^{\mathrm{T}}\tilde{\mathbf{X}}\tilde{u}_k \le 40\max\left\{ N(\tilde{u}), \left(\sum_{k=1}^{K}\|\tilde{\mathbf{X}}\tilde{u}_k\|_n^2\right)^{1/2} \frac{\sigma\left(\sqrt{\pi/2}+\sqrt{2\log(1/\delta_0)}\right)}{\sqrt{n}} \right\},$$

for some absolute constant $c_1 > 0$. Furthermore, assuming $\delta_0 = o(1)$ as $n \to \infty$, then, for sufficiently large n, we have that

$$\sum_{k=1}^{K} \frac{1}{n}\epsilon_k^{\mathrm{T}}\tilde{\mathbf{X}}\tilde{u}_k \le 40\max\left\{ N(\tilde{u}), \left(\sum_{k=1}^{K}\|\tilde{\mathbf{X}}\tilde{u}_k\|_n^2\right)^{1/2} \frac{2\sigma\sqrt{\log(1/\delta_0)}}{\sqrt{n}} \right\},$$

holds with probability tending to one.

Proof. With $\epsilon = (\epsilon_1^{\mathrm{T}},\ldots,\epsilon_K^{\mathrm{T}})^{\mathrm{T}} \in \mathbb{R}^{nK}$ and the following diagonal-block matrix

$$\tilde{\mathbb{X}} = \begin{pmatrix} \tilde{\mathbf{X}} & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \tilde{\mathbf{X}} \end{pmatrix} \in \mathbb{R}^{nK\times\tilde{D}},$$

we observe that $\sum_{k=1}^{K} n^{-1}\epsilon_k^{\mathrm{T}}\tilde{\mathbf{X}}\tilde{u}_k = n^{-1}\epsilon^{\mathrm{T}}\tilde{\mathbb{X}}\tilde{u}$. According to the group partition $(1),\ldots,(d)$, we can split the columns $\{1,\ldots,\tilde{D}\}$ of $\tilde{\mathbb{X}}$ into d disjoint groups $\cup_{j=1}^{d} G_{(j)}$, each containing K^2 elements. Set $\Phi = n^{-1/2}\epsilon^{\mathrm{T}}\tilde{\mathbb{X}}$, and for each $j \in [d]$ and any $S \subset G_{(j)}$ with $|S| = s_0^{\circ}$, denote $\Phi_S = n^{-1/2}\epsilon^{\mathrm{T}}\tilde{\mathbb{X}}_S$ where $\tilde{\mathbb{X}}_S$ is the submatrix of $\tilde{\mathbb{X}}$ with columns indexed by the set S.

By calculation,

$$\mathbf{\Sigma}_\epsilon := \mathrm{cov}(\epsilon) = \mathbb{E}_{\mathbf{X}} \begin{pmatrix} p_1(1-p_1)I_n & -p_1p_2I_n & \dots & -p_1p_KI_n \\ -p_2p_1I_n & p_2(1-p_2)I_n & \dots & -p_2p_KI_n \\ \vdots & \vdots & \ddots & \vdots \\ -p_Kp_1I_n & \dots & \dots & p_K(1-p_K)I_n \end{pmatrix} \in \mathbb{R}^{nK\times nK}, \quad \text{(A.26)}$$

where we denote $p_k = p_k^{(1)}(\boldsymbol{\gamma}^{\circ *})$ for short, and the expectation $\mathbb{E}_{\mathbf{X}}$ is taken with respect to $\mathbf{X}$. Observe that $\mathbf{\Sigma}_\epsilon = \mathbb{E}_{\mathbf{X}}[P \otimes I_n]$, where $\otimes$ is the Kronecker product, and

$$P = \begin{pmatrix} p_1(1-p_1) & -p_1p_2 & \dots & -p_1p_K \\ -p_2p_1 & p_2(1-p_2) & \dots & -p_2p_K \\ \vdots & \vdots & \ddots & \vdots \\ -p_Kp_1 & \dots & \dots & p_K(1-p_K) \end{pmatrix}.$$

From Lemma A.2.5, $\lambda_{\min}(P) \geq c_*^2 > 0$ by Assumption 2. Consequently, $\lambda_{\min}(\mathbf{\Sigma}_\epsilon) \geq c_*^2 > 0$.

Let

$$T = \left\{ \tilde{u} = (\tilde{u}_1^{\mathrm{T}}, \dots, \tilde{u}_K^{\mathrm{T}})^{\mathrm{T}} \in \mathbb{R}^{\tilde{D}} : \max\left(N(\tilde{u}), \frac{1}{L}\sqrt{\sum_{k=1}^K \|\tilde{\mathbf{X}}\tilde{u}_k\|_n^2} \right) \leq 1 \right\},$$

with $L = \sqrt{n}/[\sigma(\sqrt{\pi/2} + \sqrt{2\log(1/\delta_0)})]$. Using Lemma A.2.2 and Proposition 1, we obtain that, with probability at least $1-\delta_0$,

$$\begin{aligned} \sup_{u\in T} \sum_{k=1}^K \frac{1}{n} \epsilon_k^{\mathrm{T}} \tilde{\mathbf{X}} \tilde{u}_k &\leq 8\sigma \mathrm{Med}\left[\sup_{\mathbf{u}\in T} \sum_{k=1}^K \frac{1}{n} \mathbf{z}_k^{\mathrm{T}} \tilde{\mathbf{X}} \tilde{u}_k \right] + \frac{8L\sigma}{\sqrt{n}}(\sqrt{\pi/2} + \sqrt{2\log(1/\delta_0)}) \\ &\leq 32 + \frac{8L\sigma}{\sqrt{n}}(\sqrt{\pi/2} + \sqrt{2\log(1/\delta_0)}) = 40, \end{aligned}$$

where $\mathbf{z} = (\mathbf{z}_1^{\mathrm{T}}, \dots, \mathbf{z}_K^{\mathrm{T}})^{\mathrm{T}} \sim N_{nK}(0, I_{nK})$ is a Gaussian random vector with $\mathbf{z}_k = (z_k^{(1)}, \dots, z_k^{(n)}) \in \mathbb{R}^n$ for $k \in [K]$. □

Lemma A.2.2. *Consider $N(\tilde{u})$ in Equation* (A.25) *and $\mathbf{\Sigma}_\epsilon$ in Equation* (A.26)*, and let $\delta_0 \in (0,1)$. Then, under Assumption 2, with probability at least $1-\delta_0$ we have,*

$$\sup_{\tilde{u}=(\tilde{u}_1^{\mathrm{T}},\ldots,\tilde{u}_K^{\mathrm{T}})^{\mathrm{T}}\in\mathbb{R}^{\tilde{D}}:N(\tilde{u})\leq 1}\left|\sum_{k=1}^{K}\frac{1}{n}\xi_k^{\mathrm{T}}\tilde{\mathbf{X}}\tilde{u}_k\right|\leq 4,$$

where $\xi=(\xi_1^{\mathrm{T}},\ldots,\xi_K^{\mathrm{T}})^{\mathrm{T}}\sim N_{nK}(0,\sigma^2 I_{nK})$, and $N_{nK}(0,I_{nK})$ is the nK-dimensional standard Gaussian distribution.

Proof. Recall the notation $\tilde{\mathbb{X}}$ and the index set S with $|S|=s_0^\circ$ introduced in the proof of Lemma A.2.1. Let $\Phi_S=n^{-1/2}\tilde{\xi}^{\mathrm{T}}\tilde{\mathbb{X}}_S$, where $\tilde{\xi}=\sigma\xi\sim N_{nK}(0,\sigma I_{nK})$ and

$$\|\tilde{\mathbb{X}}_S\|_2\leq\sqrt{\|\tilde{\mathbb{X}}_S\|_1\|\tilde{\mathbb{X}}_S\|_\infty}\leq\sqrt{n},$$

where the second inequality holds because the absolute sum of each row in $\tilde{\mathbb{X}}_S$ is at most 1 due to the construction of S. Consequently, we can re-establish Theorem 1 in Li et al. (2023) by the argument therein, which leads to the desired result. □

Proposition 1. *Let $U\subset\{\mathbf{u}\in\mathbb{R}^{nK}:\|\mathbf{u}\|_2\leq 1\}$ be a subset of the unit ball. Given ϵ and $\sigma>0$ in Lemma A.2.1, for any $x>0$, with probability with at least $1-\exp(-x)$ we have*

$$\sup_{\mathbf{u}\in U}\epsilon^{\mathrm{T}}\mathbf{u}\leq 8\sigma\mathbb{E}\left[\sup_{\mathbf{u}\in U}\mathbf{z}^{\mathrm{T}}\mathbf{u}\right]+8\sigma\sqrt{2x}\leq 8\sigma\mathit{Med}\left[\sup_{\mathbf{u}\in U}\mathbf{z}^{\mathrm{T}}\mathbf{u}\right]+8\sigma(\sqrt{\pi/2}+\sqrt{2x}),$$

where $\mathbf{z}=(\mathbf{z}_1^{\mathrm{T}},\ldots,\mathbf{z}_K^{\mathrm{T}})^{\mathrm{T}}\sim N_{nK}(0,I_{nK})$ is a Gaussian random vector with $\mathbf{z}_k=(z_k^{(1)},\ldots,z_k^{(n)})\in\mathbb{R}^n$ for $k\in[K]$, and Med$(\cdot)$ is the median function.

Proof. Let $(\varepsilon^{(1)}, \ldots, \varepsilon^{(n)})$ be a vector of i.i.d. Rademacher variables independent of ξ. Then, for any $t > 0$, we have

$$\begin{aligned}
\mathbb{E}\left[\exp\left(t \sup_{\mathbf{u}\in U} \epsilon^{\mathrm{T}}\mathbf{u}\right)\right] &= \mathbb{E}\left[\exp\left(t \sup_{\mathbf{u}\in U} \sum_{k=1}^{K}\sum_{i=1}^{n} \epsilon_k^{(i)} u_k^{(i)}\right)\right] \\
&= \mathbb{E}\left[\exp\left(t \sup_{\mathbf{u}\in U} \sum_{i=1}^{n}\left[\sum_{k=1}^{K} \epsilon_k^{(i)} u_k^{(i)}\right]\right)\right] \\
&\leq \mathbb{E}\left[\exp\left(t \sup_{\mathbf{u}\in U} \sum_{i=1}^{n} 2\varepsilon^{(i)}\left[\sum_{k=1}^{K} \epsilon_k^{(i)} u_k^{(i)}\right]\right)\right] \\
&\leq \mathbb{E}\left[\exp\left(8t\sigma \sup_{\mathbf{u}\in U} \mathbf{z}^{\mathrm{T}}\mathbf{u}\right)\right], \qquad \text{(A.27)}
\end{aligned}$$

where the first inequality is due to the symmetrization inequality (Theorem 2.1 in Koltchinskii, 2011), and the second inequality is from a contraction inequality (Lemma 4.6 in Ledoux and Talagrand, 1991) and the fact that $P\left(|\varepsilon^{(i)}\epsilon_k^{(i)}| > t\right) \leq 4P\left(\sigma|z_k^{(i)}| > t\right)$ for $i \in [n]$ and $k \in [K]$ due to Lemma H.1 in Bellec et al. (2018). Let $f(\mathbf{x}) = \sup_{\mathbf{u}\in U} \mathbf{x}^{\mathrm{T}}\mathbf{u}$, then f is 1-Lipshitz. Thus, by Theorem 5.5 in Boucheron et al. (2013),

$$\begin{aligned}
\mathbb{E}\left[\exp\left(8\sigma t \sup_{\mathbf{u}\in U} \mathbf{z}^{\mathrm{T}}\mathbf{u}\right)\right] &= \mathbb{E}\left[\exp\left(8\sigma t f(\mathbf{z})\right)\right] \\
&\leq \exp\left(8\sigma t \mathbb{E}[f(\mathbf{z})] + 32\sigma^2 t^2\right) \\
&= \exp\left(8\sigma t \mathbb{E}\left[\sup_{\mathbf{u}\in U} \mathbf{z}^{\mathrm{T}}\mathbf{u}\right] + 32\sigma^2 t^2\right). \qquad \text{(A.28)}
\end{aligned}$$

Furthermore, by the discussion after equation (1.6) in Ledoux and Talagrand (1991),

$$\left|\text{Med}\left[\sup_{\mathbf{u}\in U} \mathbf{z}^{\mathrm{T}}\mathbf{u}\right] - \mathbb{E}\left[\sup_{\mathbf{u}\in U} \mathbf{z}^{\mathrm{T}}\mathbf{u}\right]\right| = |\text{Med}[f(\mathbf{z})] - \mathbb{E}[f(\mathbf{z})]| \leq \sqrt{\pi/2}. \qquad \text{(A.29)}$$

Inserting Equation (A.28) into Equation (A.27), a Chernoff argument with Equation (A.29) completes the proof. □

Introduce the cone

$$\mathcal{C}_{SRE}(s,c) = \{\mathbf{v} = (\mathbf{v}_1^{\mathrm{T}}, \ldots, \mathbf{v}_K^{\mathrm{T}})^{\mathrm{T}} \in \mathbb{R}^D : \|\mathbf{v}\|_1 \leq (1+c)\sqrt{s}\|\mathbf{v}\|_2\}$$

for some absolute constant $c > 0$.

Lemma A.2.3. *For a design matrix $\mathbf{X} \in \mathbb{R}^{n\times p}$ with rows $\mathbf{x}^{(i)}$ of bounded elements for $1 \leq i \leq n$, let $\mathbf{\Sigma} = \mathbb{E}[\mathbf{x}^{(1)}(\mathbf{x}^{(1)})^{\mathrm{T}}]$ and assume its smallest eigenvalue satisfies $\lambda_{\min}(\mathbf{\Sigma}) \geq c_\lambda$ for some constant c_λ. Assuming $s \ll \sqrt{n/\log p}$, with probability at least $1 - c_1 p^{-1}$ we have, for sufficiently large n and all $\mathbf{v} \in \mathcal{C}_{SRE}(s,c)\backslash\{0\}$,*

$$\frac{\sum_{k=1}^K \|\mathbf{X}\mathbf{v}_k\|_n^2}{\|\mathbf{v}\|_2^2} \geq \lambda_{\min}(\mathbf{\Sigma}).$$

Proof. By Lemma A.2.4, with probability at least $1 - c_1 p^{-1}$, for all $k \in [K]$,

$$\|\mathbf{X}\mathbf{v}_k\|_n^2 \geq \lambda_{\min}(\mathbf{\Sigma})\|\mathbf{v}_k\|_2^2 - C_1\sqrt{\frac{\log p}{n}}\|\mathbf{v}_k\|_1^2.$$

Therefore,

$$\begin{aligned}
\sum_{k=1}^K \|\mathbf{X}\mathbf{v}_k\|_n^2 &\geq \lambda_{\min}(\mathbf{\Sigma})\|\mathbf{v}\|_2^2 - C_1\sqrt{\frac{\log p}{n}}\sum_{k=1}^K \|\mathbf{v}_k\|_1^2 \\
&\geq \lambda_{\min}(\mathbf{\Sigma})\|\mathbf{v}\|_2^2 - C_1\sqrt{\frac{\log p}{n}}\left(\sum_{k=1}^K \|\mathbf{v}_k\|_1\right)^2 \\
&= \lambda_{\min}(\mathbf{\Sigma})\|\mathbf{v}\|_2^2 - C_1\sqrt{\frac{\log p}{n}}\|\mathbf{v}\|_1^2 \\
&\geq \lambda_{\min}(\mathbf{\Sigma})\|\mathbf{v}\|_2^2 - C_2\sqrt{\frac{\log p}{n}}s\|\mathbf{v}\|_2^2
\end{aligned}$$

where the first inequality holds since $\|\mathbf{v}\|_2^2 = \sum_{k=1}^K \|\mathbf{v}_k\|_2^2$, and the last inequality is due to the fact $\mathbf{v} \in \mathcal{C}_{SRE}(s,c) \setminus \{0\}$. In the above inequality, the second term on the right-hand side tend to zero as n goes to infinity due to the assumption that $s \ll \sqrt{n/\log p}$. □

Lemma A.2.4. *For a design matrix $\mathbf{X} \in \mathbb{R}^{n\times p}$, assume its rows $\mathbf{x}^{(i)} \in \mathbb{R}^p$, $i = 1, \ldots, n$, have bounded elements. That is, $\max_{i,j} |\mathbf{x}_j^{(i)}| \leq c$ for some absolute constant $c > 0$. Denote $\mathbf{\Sigma} =$*

$\mathbb{E}[\mathbf{x}^{(1)}(\mathbf{x}^{(1)})^{\mathrm{T}}]$ *and suppose that its smallest eigenvalue satisfies* $\lambda_{\min}(\mathbf{\Sigma}) \geq c_\lambda$ *for some constant* c_λ*. Then, with probability at least* $1 - c_1 p^{-1}$*, we have*

$$\|\mathbf{X}\mathbf{v}\|_n^2 \geq \lambda_{\min}(\mathbf{\Sigma})\|\mathbf{v}\|_2^2 - C_1\sqrt{\frac{\log p}{n}}\|\mathbf{v}\|_1^2$$

for all $\mathbf{v} \in \mathbb{R}^p$*.*

Proof. Let $\tilde{\lambda} = \|\widehat{\mathbf{\Sigma}} - \mathbf{\Sigma}\|_\infty$, where $\mathbf{X} = (\mathbf{x}^{(1)}, \ldots, \mathbf{x}^{(n)})^{\mathrm{T}} \in \mathbb{R}^{n\times p}$ is the design matrix, $\widehat{\mathbf{\Sigma}} = n^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{X}$ is the sample covariance matrix and $\|A\|_\infty$ for a matrix A is the element-wise maximum norm of A. Then for any $\mathbf{v} \in \mathbb{R}^p$,

$$\left|\mathbf{v}^{\mathrm{T}}\widehat{\mathbf{\Sigma}}\mathbf{v} - \mathbf{v}^{\mathrm{T}}\mathbf{\Sigma}\mathbf{v}\right| = \left|\mathbf{v}^{\mathrm{T}}(\widehat{\mathbf{\Sigma}} - \mathbf{\Sigma})\mathbf{v}\right| \leq \tilde{\lambda}\|\mathbf{v}\|_1^2.$$

Since elements in $\mathbf{X}$ are bounded, according to Problem 14.3 in Bühlmann and van de Geer (2011), we have

$$P\left(\|\widehat{\mathbf{\Sigma}} - \mathbf{\Sigma}\|_\infty \geq C_1\sqrt{\frac{\log p}{n}}\right) \leq c_1 p^{-1},$$

for some absolute constants $C_1, c_1 > 0$. Therefore, with probability at least $1 - c_1 p^{-1}$, one has

$$\|\mathbf{X}\mathbf{v}\|_n^2 = \mathbf{v}^{\mathrm{T}}\widehat{\mathbf{\Sigma}}\mathbf{v} \geq \mathbf{v}^{\mathrm{T}}\mathbf{\Sigma}\mathbf{v} - C_1\sqrt{\frac{\log p}{n}}\|\mathbf{v}\|_1^2 \geq \lambda_{\min}(\mathbf{\Sigma})\|\mathbf{v}\|_2^2 - C_1\sqrt{\frac{\log p}{n}}\|\mathbf{v}\|_1^2$$

□

Lemma A.2.5. *Consider a probability sequence* $(p_0, p_1, \ldots, p_K)$ *satisfying that* $\sum_{k=0}^K p_k = 1$ *and* $\min_{k\in[K]_0} p_k \geq c_*$ *for some constant* $c_* > 0$*. Define the matrix*

$$Q = \begin{pmatrix} p_1(1-p_1) & -p_1p_2 & \ldots & -p_1p_K \\ -p_2p_1 & p_2(1-p_2) & \ldots & -p_2p_K \\ \vdots & \vdots & \ddots & \vdots \\ -p_Kp_1 & \ldots & \ldots & p_K(1-p_K) \end{pmatrix}.$$

Then, the smallest eigenvalue of Q, $\lambda_{\min}(Q)$, satisfies

$$\lambda_{\min}(Q) \geq \min_{1 \leq k \leq K} p_k p_0 \geq c_*^2.$$

Proof. By the definition of $(p_0, p_1, \ldots, p_K)$, for each $k \in [K]$, the sum of the absolute values of the non-diagonal entries in the kth row is $r_k = \sum_{j \neq k} p_k p_j = p_k(1 - p_k - p_0)$. Hence, by the Gershgorin circle theorem, there exists some $k_0 \in [K]$, such that the smallest eigenvalue of Q satisfies

$$|\lambda_{\min}(Q) - p_{k_0}(1 - p_{k_0})| \leq r_{k_0},$$

which implies

$$\lambda_{\min}(Q) \geq p_{k_0}(1 - p_{k_0}) - r_{k_0} = p_{k_0} p_0 \geq \min_{1 \leq k \leq K} p_k p_0 \geq c_*^2,$$

as desired. □

Remark A.2.6. *Notably, the lower bound derived in Lemma A.2.5 is tight. Observe that $Q = P_1 - P_2$, where $P_1 = \mathit{diag}(v)$ is a diagonal matrix with diagonal elements v, and $P_2 = vv^{\mathrm{T}}$, with $v = (p_1, \ldots, p_K) \in \mathbb{R}^K$. Clearly, the eigenvalues of P_1 are the elements in v, while the eigenvalues of P_1 are $\mathit{trace}(P_2) = \sum_{k=1}^K p_k^2$ and 0. Consider the case where $p_0 = p_1 = \cdots = p_K = 1/(K+1)$. That is, $c_* = 1/(K+1)$. Then, the largest eigenvalue of Q is $1/(K+1)$, while the smallest eigenvalue of Q is $1/(K+1)^2$. Let $\lambda_1(Q) \geq \lambda_2(Q) \geq \cdots \geq \lambda_K(Q)$ be the eigenvalues of Q in descending order. By Weyl's inequality, we have*

$$1/(K+1) = \lambda_2(P_1) - \lambda_2(P_2) \leq \lambda_1(Q) \leq \lambda_1(P_1) - \lambda_K(P_2) = 1/(K+1),$$

and

$$\lambda_1(P_1) - \lambda_1(P_2) \leq \lambda_K(Q) \leq \lambda_1(P_1) - \lambda_1(P_2),$$

where $\lambda_1(P_1) - \lambda_1(P_2) = p_1 - \sum_{k=1}^K p_k^2 = 1/(K+1) - K/(K+1)^2 = 1/(K+1)^2$.

Lemma A.2.7. *Given a vector $x = (x_1, \ldots, x_K)^{\mathrm{T}} \in \mathbb{R}^K$, for each $1 \leq k \leq K$, define the function $g_k(x) = \frac{\exp(x_k)}{[1+\sum_{k=1}^K \exp(x_k)]^2}$. Then, we have that*

$$\exp\left(-2\sum_{l=1}^{K}|x_k - y_k|\right) \leq \frac{g_k(x)}{g_k(y)} \leq \exp\left(2\sum_{l=1}^{K}|x_k - y_k|\right),$$

holds for each $1 \leq k \leq K$ and any vector $y = (y_1, \ldots, y_K)^{\mathrm{T}}$.

Proof. Note that

$$\begin{aligned}
\frac{g_k(x)}{g_k(y)} &= \exp(x_k - y_k)\frac{[1+\sum_{k=1}^K \exp(y_k)]^2}{[1+\sum_{k=1}^K \exp(x_k)]^2} \\
&\leq \exp(x_k - y_k)\exp\left(2\sum_{l=1}^{K}(y_k - x_k)_+\right) \\
&\leq \exp\left(2\sum_{l=1}^{K}|x_k - y_k|\right),
\end{aligned}$$

where we define $(a)_+ = \max\{a, 0\}$ for any number a. Similarly, we also have

$$\frac{g_k(y)}{g_k(x)} \leq \exp\left(2\sum_{l=1}^{K}|x_k - y_k|\right).$$

Combining these two results completes the proof. □

A.3 Extra Results from Numerical Studies

A.3.1 Data generation

For simplicity, we assume that each site of $\mathbf{z}$ can take on K distinct values instead of $K+1$. The space of $\mathbf{z}$ from the Potts model Equation (2.1) consists of K^d elements, making direct computation infeasible even for moderate K and d. To clarify the generation process, we adopt a categorical representation for protein sequences $\mathbf{z}$: for a sequence with d sites, $\mathbf{z} = (z_1, ..., z_d)^{\mathrm{T}} \in \mathbb{R}^d$, where $z_j \in [K]$ for $j \in [d]$. Denote the sequence that excludes the j-th site as $\mathbf{z}_{-j} = (z_1, ..., z_{j-1}, z_{j+1}, ..., z_d)^{\mathrm{T}} \in \mathbb{R}^{d-1}$.

The widely-used Swendsen–Wang algorithm (Edwards and Sokal, 1988; Izenman, 2021) applies only to the Potts model in the specific form $P(\mathbf{z}) \propto \exp\left\{\sum_{i,j} J_{jr}[1 - \mathbb{1}(z_j = z_r)]\right\}$, where the direct couplings between two sites J_{jr} are scalars. Specifically, when $z_j \neq z_r$, the pair contributes J_{jr} to the energy function in Equation (2.2), whereas $z_j = z_r$ contributes nothing. In contrast, our study focuses on a more general case where the direct couplings between sites are characterized by $\boldsymbol{\gamma}_{jr} = (\gamma_{jr,11}, ..., \gamma_{jr,kl}, ..., \gamma_{jr,KK}) \in \mathbb{R}^{K^2}$. We adopt a Gibbs sampler as an alternative for generating samples from the Potts model (Casella and George, 1992; Gelfand, 2000; Li et al., 2019). The Gibbs sampler sequentially samples values for each site z_j based on the conditional probability given the current values of other sites $\mathbf{z}_{-j}$:

$$P(z_j|\mathbf{z}_{-j}) = \frac{\prod_{k=1}^{K} \exp\{\theta_{jk} + \sum_{r\neq j} \sum_{l=1}^{K} \gamma_{jr,kl} \mathbb{1}(z_r = l)\}^{\mathbb{1}(z_j=k)}}{\sum_{k=1}^{K} \exp\{\theta_{jk} + \sum_{r\neq j} \sum_{l=1}^{K} \gamma_{jr,kl} \mathbb{1}(z_r = l)\}}. \tag{A.30}$$

The data generation process is outlined in Algorithm 3. To generate a data set, we ran Algorithm 3 for 10^5 iterations, initializing all sites with values randomly generated from a discreet uniform distribution (Li et al., 2017, 2019). To mitigate the effects of random initialization and thoroughly explore the state space of $\mathbf{z}$, the first 40% of iterations were discarded as the burn-in process (Carlo, 2004).

Algorithm 3: Data generation from the Potts model

Input: Model coefficients $\boldsymbol{\Gamma}_{d(d-1)\times K^2}$ and $\boldsymbol{\theta}$, number of sites d and sample size n, initial $\mathbf{z}^0$ and total iterations R

for $r \in [R]$, $j \in [d]$ **do**

  Sample z_j^r from the multinomial distribution $P(z_j^r | z_{1:(j-1)}^r, z_{(j+1):d}^{r-1})$ in Equation (A.30)

end

Collect the pool $\mathcal{Z} := \{\mathbf{z}^r \mid 0.4R < r \leq R\}$;

Randomly draw n samples from $\mathcal{Z}$, denoted as $\mathcal{Z}_n$.

Output: $\mathcal{Z}_n$

To demonstrate the capability of Algorithm 3 to simulate data that closely resemble real data, we used the estimated coefficients from Equation (2.8) based on real data from Tyrosine-protein kinase Fyn (FYN, Di Nardo et al. (2003)) and Yes-associated protein (YAP, Araya et al.

(2012)) along with Algorithm 3 to generate data, and then computed the similarity matrix between real and simulated data using the LIN1 measure for categorical data Lin (1998); Boriah et al. (2008). This matrix was further projected onto a two-dimensional space using Uniform Manifold Approximation and Projection (UMAP, Becht et al. (2019)). Similar idea to validate the simulated data by comparing it with the real data has also been employed in synthetic single-cell RNASeq experiments (Song et al., 2024; Xia et al., 2024). As shown in Figure A.1, the simulated data occupies a similar two-dimensional space as the real data, demonstrating the Algorithm 3's capability to emulate real data effectively.

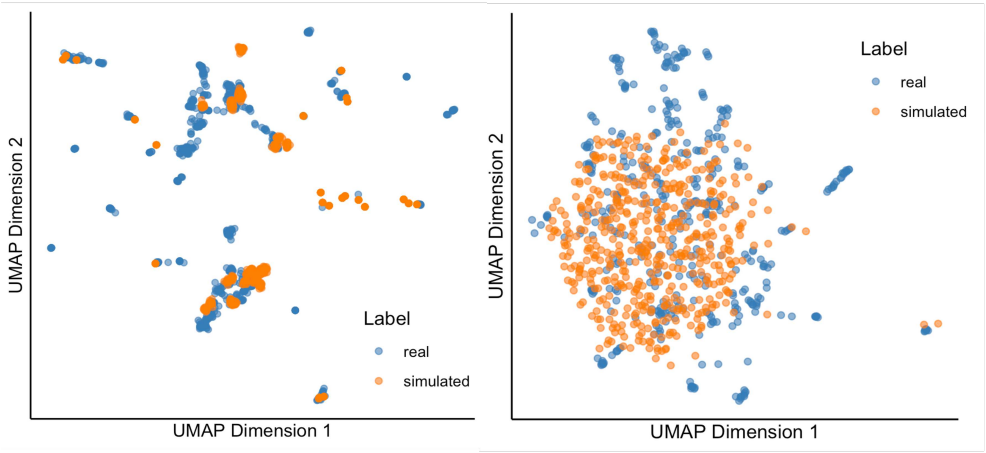


Figure A.1: UMAP of real and simulated data with similarity matrix calculated from LIN1 measure. Left: YAP; Right: FYN.

A.3.2 Extra simulations

Table A.1 presents the results for setting (M2) with $n = 1000, 2000, 4000$, $d = 25, 50$, and $K = 20$. Additionally, we consider simulations with $d = 150$ sites and large sample size of $n = 4000, 6000, 8000$ under settings of (M1) and (M2), resembling the size of real data. For setting (M2) with $d = 150$, we set $\tau = 2$. The results are shown in Table A.2, which show that our method outperforms others in both estimation and variable selection, even when the group weights function is misspecified. Furthermore, the sparse group Lasso without structural

information performs better than Lasso and ridge, highlighting the benefits of incorporating both group-wise and element-wise sparsity.

Table A.1: Results in (M2) with $K = 20$

d	$\sum_{j<r} A_{jr}$	Methods	n	MSE	TPR	FDR	TPR_g	FDR_g
25	26	Our method with (N1)	1000	40.395	0.774	0.062	0.942	0.228
		Our method with (N2)		43.658	0.743	0.078	0.942	0.234
		Lasso		59.028	0.655	0.328	0.846	0.353
		Sparse Group Lasso		49.444	0.720	0.137	0.923	0.284
		Ridge		95.815	–	–	–	–
		Our method with (N1)	2000	31.581	0.831	0.054	0.962	0.206
		Our method with (N2)		32.967	0.807	0.067	0.962	0.212
		Lasso		45.172	0.732	0.295	0.904	0.309
		Sparse Group Lasso		35.084	0.783	0.110	0.942	0.246
		Ridge		69.860	–	–	–	–
		Our method with (N1)	4000	25.039	0.862	0.047	1.000	0.161
		Our method with (N2)		28.743	0.826	0.059	1.000	0.175
		Lasso		37.588	0.765	0.266	0.942	0.290
		Sparse Group Lasso		31.117	0.801	0.088	1.000	0.224
		Ridge		55.145	–	–	–	–
50	89	Our method with (N1)	1000	92.414	0.756	0.069	0.907	0.225
		Our method with (N2)		94.252	0.720	0.074	0.903	0.232
		Lasso		113.466	0.664	0.348	0.854	0.340
		Sparse Group Lasso		102.733	0.701	0.092	0.892	0.264
		Ridge		143.535	–	–	–	–
		Our method with (N1)	2000	68.252	0.815	0.061	0.959	0.206
		Our method with (N2)		70.564	0.787	0.070	0.952	0.217
		Lasso		83.056	0.716	0.315	0.898	0.320
		Sparse Group Lasso		77.480	0.752	0.089	0.925	0.241
		Ridge		99.579	–	–	–	–
		Our method with (N1)	4000	54.794	0.870	0.057	1.000	0.188
		Our method with (N2)		58.138	0.846	0.064	1.000	0.201
		Lasso		69.266	0.736	0.281	0.934	0.287
		Sparse Group Lasso		61.852	0.827	0.082	1.000	0.233
		Ridge		85.917	–	–	–	–

Table A.2: Results with $K = 20, d = 150$ under different model settings

Model	$\sum_{j<r} A_{jr}$	Methods	n	MSE	TPR	FDR	TPR_g	FDR_g
(M1)	353	Our method with (N1)	4000	317.497	0.788	0.096	0.935	0.211
		Our method with (N2)		352.646	0.751	0.131	0.934	0.220
		Lasso		421.714	0.694	0.316	0.852	0.319
		Sparse Group Lasso		384.022	0.726	0.158	0.907	0.247
		Ridge		488.528	–	–	–	–
		Our method with (N1)	6000	273.158	0.823	0.087	0.952	0.194
		Our method with (N2)		297.134	0.791	0.115	0.948	0.208
		Lasso		359.365	0.725	0.279	0.884	0.290
		Sparse Group Lasso		325.645	0.752	0.124	0.922	0.233
		Ridge		412.459	–	–	–	–
		Our method with (N1)	8000	210.518	0.881	0.062	0.992	0.169
		Our method with (N2)		233.982	0.857	0.084	0.984	0.184
		Lasso		295.008	0.802	0.236	0.935	0.276
		Sparse Group Lasso		248.156	0.825	0.095	0.964	0.213
		Ridge		377.510	–	–	–	–
(M2)	355	Our method with (N1)	4000	351.374	0.775	0.105	0.915	0.253
		Our method with (N2)		369.702	0.757	0.143	0.909	0.272
		Lasso		406.606	0.686	0.330	0.855	0.318
		Sparse Group Lasso		374.742	0.729	0.176	0.901	0.290
		Ridge		502.565	–	–	–	–
		Our method with (N1)	6000	309.590	0.817	0.084	0.945	0.225
		Our method with (N2)		322.535	0.801	0.109	0.941	0.243
		Lasso		360.367	0.711	0.296	0.876	0.305
		Sparse Group Lasso		339.063	0.785	0.133	0.934	0.278
		Ridge		440.593	–	–	–	–
		Our method with (N1)	8000	244.307	0.869	0.076	0.985	0.203
		Our method with (N2)		261.112	0.844	0.094	0.972	0.222
		Lasso		314.978	0.772	0.267	0.931	0.280
		Sparse Group Lasso		278.886	0.825	0.118	0.968	0.264
		Ridge		416.751	–	–	–	–

Appendix B

Supplemental materials for Chapter 3

B.1 Equivalence of Poisson non-negative model and pLSI

Let $\mathbf{Z} = (Z_{ij})_{(i,j)} \in \mathbb{R}_+^{n\times V}$ denote the corpus matrix. Both our model without zero-inflation and pLSI can be considered as different fitting models for $\mathbf{Z}$. Given $K \geq 2$, the topic-vocabulary matrix $\mathbf{A} = (A_{jk})_{(j,k)} \in \mathbb{R}_+^{K\times V}$ and the document-topic matrix $\mathbf{W} = (W_{ik})_{(i,k)} \in \mathbb{R}_+^{n\times K}$, our model is

$$Z_{ij}|\mathbf{A}, \mathbf{W} \sim \text{Poisson}(\lambda_{ij}) \tag{B.1}$$

$$\lambda_{ij} = \sum_{k=1} W_{ik}A_{kj} \tag{B.2}$$

We impose non-negativity constraints on both $\mathbf{A}$ and $\mathbf{W}$, and assume that the rows of $\mathbf{A}$ lie on a simplex. Relatively, the pLSI model given K, $\mathbf{A}^*$, $\mathbf{W}^*$ and document length N_i for each i is

$$Z_{ij}|\mathbf{A}^*, \mathbf{W}^* \sim \text{Multinomial}(N_i; \pi_{i1}, ..., \pi_{ij}) \tag{B.3}$$

$$\pi_{ij} = \sum_{k=1} W^*_{ik}A^*_{kj} \tag{B.4}$$

where both $\mathbf{A}^*$ and $\mathbf{W}^*$ have the simplex constrains on rows.

Denote $\mathbf{A}_{\cdot j}$ as the jth column of $\mathbf{A}$ and $\mathbf{W}_{i\cdot}$ as the ith row of $\mathbf{W}$. We state the equivalence of $\mathbf{A}$ and $\mathbf{A}^*$ in the models in the following lemma.

Lemma B.1.1. *Denote the Poisson non-negative model density in* (B.1) *as* $\mathbf{P}_{PN}(\mathbf{Z}_{i\cdot}|\mathbf{W}_{i\cdot}, \mathbf{A}_{\cdot j})$ *and the multinomial model density in* (B.3) *as* $\mathbf{P}_{pLSI}(\mathbf{Z}_{i\cdot}|\mathbf{W}^*_{i\cdot}, \mathbf{A}^*_{\cdot j}, N_i)$. *For each document* i, *since the document length* N_i *in the pLSI model carries no information about* $\mathbf{\Pi}_{i\cdot}$ *where* $\mathbf{\Pi} = (\pi_{ij})_{i,j}$, *we*

can assume it follows a Poisson distribution to link the two models. We have

$$N_i := \sum_{j=1}^{V} Z_{ij} \sim \textit{Poisson}\left(\sum_{j=1}^{V} \lambda_{ij}\right)$$

Define the mapping $\psi : (\mathbf{A}, \mathbf{W}) \longmapsto (\mathbf{A}^*, \mathbf{W}^*, \boldsymbol{\omega} = (\omega_1, \ldots, \omega_n)^{\mathrm{T}})$ by the procedures:

$$\begin{aligned} A^*_{kj} &\leftarrow A_{kj} \;\; k = 1, ..., K; j = 1, ..., V \\ \omega_i &\leftarrow \sum_{k=1}^{K} W_{ik} \;\; i = 1, ..., n \\ W^*_{ik} &\leftarrow W_{ik}/\omega_i \;\; k = 1, ..., K; i = 1, ..., n \end{aligned}$$

Obviously, this is a one-to-one mapping with the inverse ψ^{-1} of $A_{kj} \leftarrow A^*_{kj}$ and $W_{ik} \leftarrow W^*_{ik}\omega_i$.

Notice that

$$\begin{aligned} \sum_{j=1}^{V} \lambda_{ij} = \sum_{j=1}^{V}\sum_{k=1}^{K} W_{ik}A_{kj} &= \sum_{k=1}^{K}\left(W_{ik}\sum_{j=1}^{V} A_{kj}\right) \\ &\xlongequal{\sum_{j=1}^{V} A_{kj}=1} \sum_{k=1}^{K} W_{ik} = \omega_i \sum_{k=1}^{K} W^*_{ik} \\ &\xlongequal{\sum_{k=1}^{K} W^*_{ik}=1} \omega_i \end{aligned}$$

Then the following equation of likelihoods holds

$$\mathbf{P}_{\text{PN}}(\mathbf{Z}_{i\cdot}|\mathbf{W}^*_{i\cdot}, \mathbf{A}^*, \omega_i) = \mathbf{P}_{\text{pLSI}}(\mathbf{Z}_{i\cdot}|\mathbf{W}^*_{i\cdot}, \mathbf{A}^*, N_i)\mathbf{P}(N_i|\mathbf{W}^*_{i\cdot}, \mathbf{A}^*, \omega_i)$$

Notice that

$$\begin{aligned} RHS &= \mathbf{P}_{\text{pLSI}}(\mathbf{Z}_{i\cdot}|\mathbf{W}^*_{i\cdot}, \mathbf{A}^*, N_i, \omega_i)\mathbf{P}(N_i|\mathbf{W}^*_{i\cdot}, \mathbf{A}^*, \omega_i) \\ &= \mathbf{P}(\mathbf{Z}_{i\cdot}, N_i|\mathbf{W}^*_{i\cdot}, \mathbf{A}^*, \omega_i) \\ &\xlongequal{N_i=\sum_{j=1}^{V} Z_{ij}} \mathbf{P}(\mathbf{Z}_{i\cdot}|\mathbf{W}^*_{i\cdot}, \mathbf{A}^*, \omega_i) \end{aligned}$$

Proof:

$$
\begin{aligned}
RHS &= \left(N_i! \prod_{j=1}^{V} \frac{\pi_{ij}^{Z_{ij}}}{Z_{ij}!}\right) \left(\frac{\left(\sum_{j=1}^{V} \lambda_{ij}\right)^{N_i} e^{-\sum_{j=1}^{V} \lambda_{ij}}}{N_i!}\right) \\
&= \left(\prod_{j=1}^{V} \frac{e^{-\lambda_{ij}} \pi_{ij}^{Z_{ij}}}{Z_{ij}!}\right) \left(\sum_{j=1}^{V} \lambda_{ij}\right)^{N_i} = \left(\prod_{j=1}^{V} \frac{e^{-\lambda_{ij}} \pi_{ij}^{Z_{ij}}}{Z_{ij}!}\right) \left(\sum_{j=1}^{V} \sum_{k=1}^{K} W_{ik} A_{kj}\right)^{N_i} \\
&= \left(\prod_{j=1}^{V} \frac{e^{-\lambda_{ij}} \pi_{ij}^{Z_{ij}}}{Z_{ij}!}\right) \left(\sum_{k=1}^{K} W_{ik} \left(\sum_{j=1}^{V} A_{kj}\right)\right)^{N_i} \\
&= \left(\prod_{j=1}^{V} \frac{e^{-\lambda_{ij}} \pi_{ij}^{Z_{ij}}}{Z_{ij}!}\right) \left(\sum_{k=1}^{K} W_{ik}\right)^{N_i} = \left(\prod_{j=1}^{V} \frac{e^{-\lambda_{ij}} \pi_{ij}^{Z_{ij}}}{Z_{ij}!}\right) \left(\sum_{k=1}^{K} W_{ik}\right)^{\sum_{j=1}^{V} Z_{ij}} \\
&= \prod_{j=1}^{V} \frac{e^{-\lambda_{ij}}}{Z_{ij}!} \left(\pi_{ij} \sum_{k=1}^{K} W_{ik}\right)^{Z_{ij}} = \prod_{j=1}^{V} \frac{e^{-\lambda_{ij}}}{Z_{ij}!} \left(\sum_{k^*=1}^{K} \left(\sum_{k=1}^{K} W_{ik}\right) W_{ik^*}^* A_{k^*j}^*\right)^{Z_{ij}} \\
&\xlongequal{\text{Use the mapping } \psi} \prod_{j=1}^{V} \frac{e^{-\lambda_{ij}}}{Z_{ij}!} \left(\sum_{k=1}^{K} W_{ik} A_{kj}\right)^{Z_{ij}} \\
LHS &= \prod_{j=1}^{V} \frac{e^{-\lambda_{ij}}}{Z_{ij}!} \lambda_{ij}^{Z_{ij}} = \prod_{j=1}^{V} \frac{e^{-\lambda_{ij}}}{Z_{ij}!} \left(\sum_{k=1}^{K} W_{ik} A_{kj}\right)^{Z_{ij}}
\end{aligned}
$$

B.2 Extra simulation with different random effects

In addition to evaluating random effects generated from $\mathrm{Unif}(0, 2)$ with mild variance and no skewness, we further examine settings with either no or more heterogeneous random effects. Since both our method and Topic-SCORE consistently outperform TTS, we exclude TTS from this comparison. Specifically, we consider distributions with greater variance and skewness, including the truncated normal $\mathrm{TN}(\mu = 0, \sigma^2 = 25)$, Gamma distributions with shape-scale parameterizations $\mathrm{Gamma}(3, \frac{1}{3})$ and $\mathrm{Gamma}(\frac{1}{2}, 2)$ and the log-normal distribution $\mathrm{LN}(\mu = -\frac{1}{6}, \sigma^2 = \frac{1}{3})$. We evaluate the relative Frobenius norm as the estimation error, which is

$$
L_2(\mathbf{A} - \widehat{\mathbf{A}}) = \min_{\mathbf{r} \in S([K])} \frac{\|\mathbf{A} - \widehat{\mathbf{A}}_{\mathbf{r}}\|_F}{\|\mathbf{A}\|_F}.
$$

Across all settings, as shown in Figure B.1, the estimation error decreases as n increases, confirming the consistency of both methods. In contrast, the error increases with V, reflecting the greater number of parameters in the matrix $\mathbf{A}$ that must be estimated. Estimation errors are also smaller when random effects are absent, which is expected given the reduction in variability due to the exclusion of latent noise. Beyond these general trends, our method consistently achieves lower estimation error than Topic-SCORE, with the performance gap becoming more pronounced in the presence of random effects exhibiting greater variance and skewness. Notably, under the heavy-tailed Gamma distribution with large variance, our method demonstrates substantial improvements and stable performance when V is small. In contrast, when the random effects are milder and more symmetric, the two methods perform similarly, especially as the sample size grows.

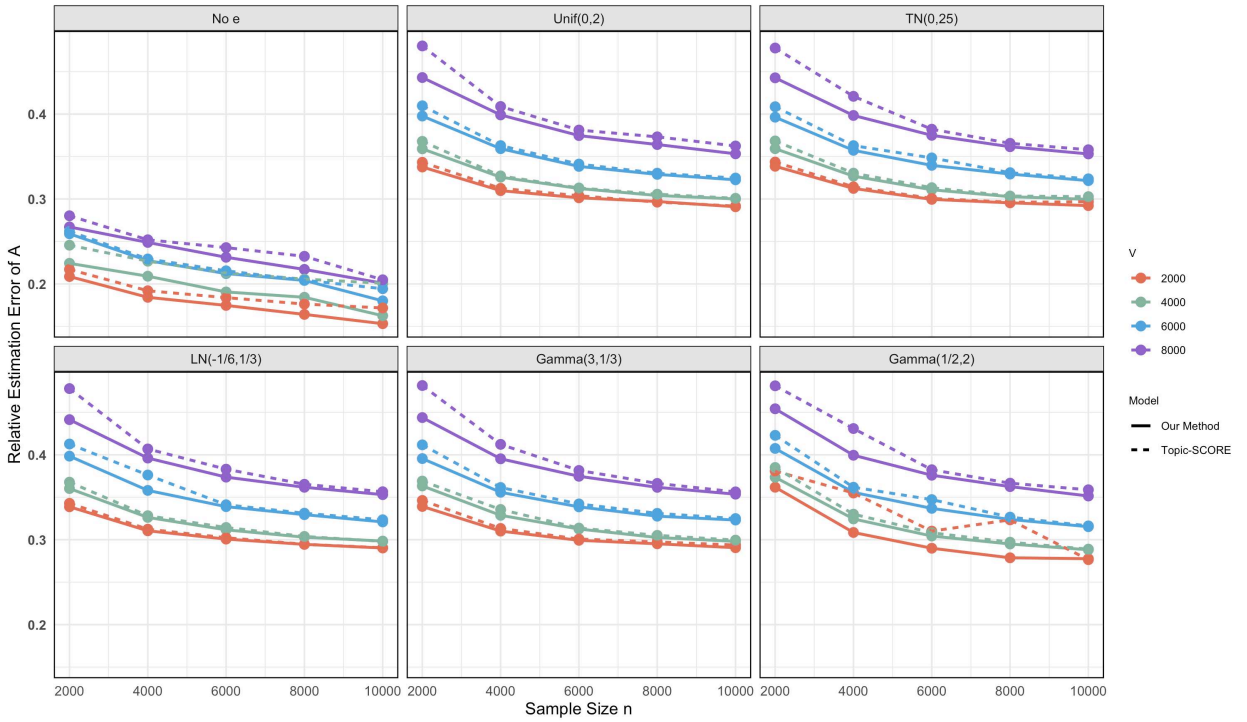


Figure B.1: $L_2(\mathbf{A} - \widehat{\mathbf{A}})$ with $K = 3$ under different random effects distributions, evaluated across varying (n, V), comparing our method with Topic-SCORE.

We further explore the random effect with heavier tails with $\mathbf{e}$ generated from Gamma$(1/2, 2)$ and Gamma$(1/4, 4)$ as shown in B.2. The results indicate that as the tail of the random effect

distribution becomes heavier, the estimation performance of our model deteriorates significantly for small n, with near failure even for low values of K. However, when n is sufficiently large, the estimation remains competitive despite the presence of heavy-tailed effect.

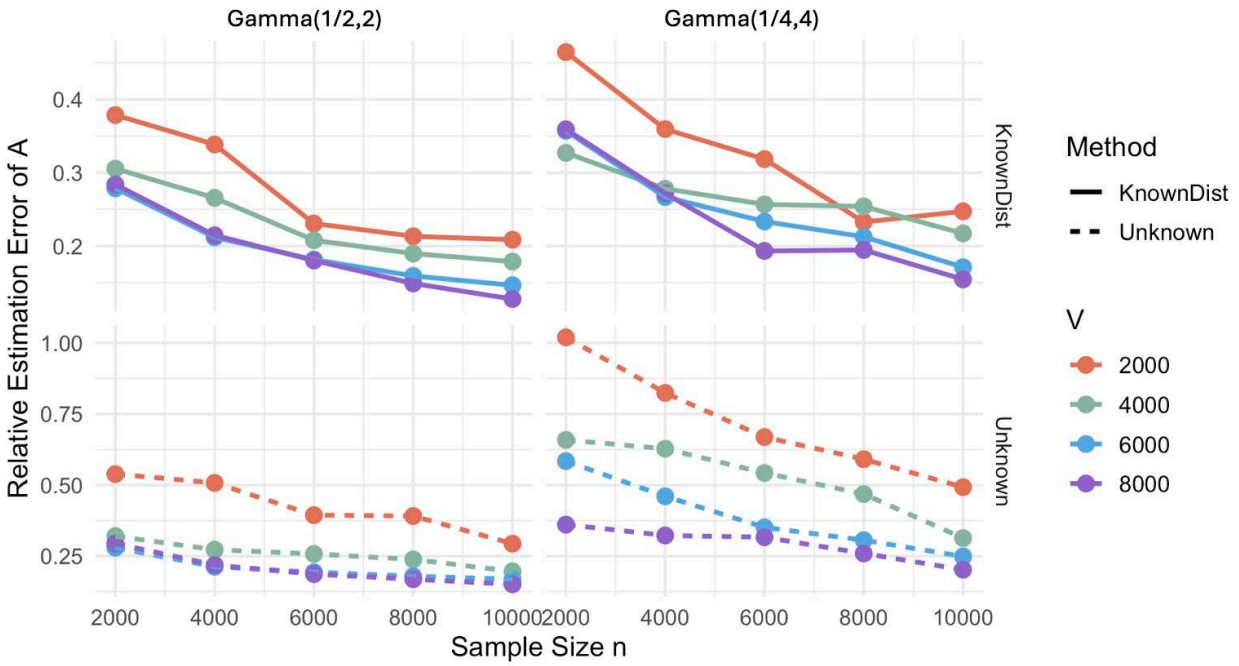


Figure B.2: $L_2(\mathbf{A} - \widehat{\mathbf{A}})$ with $K = 3$, where $\mathbf{e}$ sampled from Gamma$(1/2, 2)$ (left) and Gamma$(1/4, 4)$ (right) is known for the distribution.