Mixture models to fit heavy-tailed, heterogeneous or sparse data

Zhen Miao

A dissertation

submitted in partial fulfillment of the
requirements for the degree of

Doctor of Philosophy

University of Washington

2023

Reading Committee:

Adrian Dobra, Chair

Yen-Chi Chen, Co-Chair

Wei Sun

Program Authorized to Offer Degree:
Statistics

University of Washington

Abstract

Mixture models to fit heavy-tailed, heterogeneous or sparse data

Zhen Miao

Chair of the Supervisory Committee:

Adrian Dobra
Department of Statistics

With the advent of modern technologies, many scientific fields collect and analyze increasingly large datasets. Unfortunately, the complexity and heterogeneity of these datasets cannot be properly captured through classical statistical models. In this thesis, we develop new classes of mixture models that alleviate these issues. Their key feature is the assumption that the overall population consists of several subpopulations and each of these subpopulations can be represented through simpler statistical models. Our new mixture models are defined through three classes of distributions for different data types, as follows. The first type of mixture model is called the bi-s^*-concave distribution for continuous data. We propose this distribution as a generalization of two popular distributions, i.e., the s-concave distribution and the bi-log-concave distribution, in the field of estimation under shape constraints to include multimodal and heavy-tail densities. Although its definition is not directly related to mixture models, this class include several important mixture distributions (e.g., mixture of Student-t distributions, mixture of Gaussian distributions) under some conditions. The second type of mixture model is the nonparametric Poisson mixture distribution for count data, which generalizes Poisson distribution by assuming its parameter following a totally unknown mixing distribution. We provide a minimax-optimal convergence rate for the nonparametric maximum likelihood estimation for the mixing distribution and apply it on a single-cell RNA-sequencing data. The third type of mixture model is the Ising mixture distribution for inferring associations between binary variables. This method combines the strengths of classic methods, such as Ising models and multivariate Bernoulli mixture models. We examine the conditions required for the identifiability of the Ising mixture model, and develop a Bayesian framework for implementation. Through simulations and two real data applications, we demonstrate the effectiveness of our proposed method.

Acknowledgement

I would like to express my deepest gratitude to my supervisors, Prof. Adrian Dobra and Prof. Yen-Chi Chen, for providing me with invaluable guidance, unwavering support, and endless encouragement throughout my research. Their expertise and insightful feedback have been instrumental in shaping my work and helping me to achieve my goals. I am especially grateful to Prof. Adrian Dobra, who not only supported me in research but also helped me figure out my career path. Thanks to his reference, I was able to secure a summer internship at Microsoft, which ultimately led to a full-time job offer and the start of my career in the industry. I am immensely grateful for his support and recommendation, as it opened the door to invaluable professional opportunities and provided a strong foundation for my career growth. His confidence in my abilities has been a driving force behind my success, and I am fortunate to have had such a supportive and influential mentor.

I would also like to thank my committee members, Prof. Adrian Dobra, Prof. Yen-Chi Chen, Prof. Jon A. Wellner, and Prof. Wei Sun, and Prof. Cristen Harris for their invaluable feedback, suggestions, and critiques that helped me to refine my research and provided me with a broader perspective on my work.

I am grateful to the Department of Statistics at the University of Washington for providing me with the resources and facilities necessary to carry out my research. In particular, I would like to thank Prof. Michael Perlman for his excellent lectures in courses STAT 512, 513, and 542, which helped me build a solid background in mathematical statistics. I also appreciate Prof. Jon A Wellner for his great lectures in STAT 521, 522, and 523, which provided me with solid materials on advanced probability. In addition, I am deeply grateful to Professor Jon A. Wellner for offering me the opportunity to engage in research during my master's studies. This experience resulted in my first publication, and I am indebted to him for his support and guidance throughout the process. Not only did I gain valuable knowledge and skills from his mentorship, but I also learned the importance of a strong work ethic and dedication to research. Professor Wellner's influence has had a lasting impact on my academic and professional pursuits, and I am grateful to have had him as a mentor.

I want to thank the staff, including Mee-Ling Hon, Tracy Pham, and Ellen Reynolds, for their assistance and support during my studies. Additionally, I would like to express my appreciation to the staff at the International Student Service at the University of Washington for their efficiency and professionalism in handling all things related to my studies, especially helping me apply for CPT for my summer internship.

Life as a PhD student can be challenging and demanding, and I am incredibly grateful for the support and guidance provided by Professors Yen-Chi Chen, Adrian Dobra, Wei Sun, and Abel Rodriguez. During a difficult period of deep depression, these professors were willing to lend an ear and offer valuable advice

that helped me navigate my way forward. I cannot overstate how much their compassion and understanding meant to me during this time, and I am grateful to have had such caring mentors in my corner.

I would like to express my appreciation to my collaborators, including Prof. Jonathan Bricker (Bricker et al., 2022, 2023), Prof. Fang Han (Han et al., 2021), Prof. Wei Sun (Zhang et al., 2022), and others. Working with all of you has been an exceptional experience. I have gained valuable insights, especially in interpreting our statistical analysis results from a scientific standpoint. Your perspectives have inspired me to embrace a more practical approach rather than solely emphasizing statistical aspects.

I owe a debt of gratitude to my family, friends and my love, Wenbo, for their unwavering support, encouragement, and love. Their belief in me has been a constant source of inspiration and motivation throughout my academic journey.

Finally, I would like to express my heartfelt appreciation to all the participants who generously gave their time and shared their experiences with me. Without their contributions, this research would not have been possible.

Thank you all for your contributions, encouragement, and support.

Contents

List of Figures

1 Introduction

1.1 Heavy-tailed, heterogeneous or sparse data

As the cost of collecting and storing data continues to decrease, individuals increasingly collect and analyze larger data sets in various fields. However, these larger data sets often come with a level of complexity that was previously unappreciated, making classical statistical models inadequate for capturing the distribution of observations. This challenge has drawn attention from researchers in both academia and industry, as highlighted in a comprehensive review by Hilbert (2016). Specific difficulties include heavy-tailedness, heterogeneity, and sparsity.

Data collected from complex multi-component systems, such as economics, ecological systems, sociology, finance, and business, often exhibit heavy-tailedness. This is evident in examples such as service time and input in queuing models, flood levels of rivers, major insurance claims, extreme levels of ozone concentration, high wind-speed values, wave heights during a storm, and low and high temperatures. Sparse observations at the tail domain of the distribution, commonly referred to as outliers, can cause difficulties when analyzing such data. In classical textbooks, outliers are typically regarded as anomalies resulting from mistakes in the sample. However, in many cases, outliers are a crucial part of the data. For example, the size of files transported by a network during the transfer of some firm's home page may vary from kilobytes to megabytes (Crovella et al., 1998). A network administrator who controls the operation of the network must take into account the existence of such files to avoid network overload. In a histogram, large values will be viewed as apparent outliers. Example 1.1 below illustrates heavy-tailed data with outliers. In the field of estimation under shape constraints, there has been an increasing focus on the class of distributions with log-concave densities (Walther, 2009; Samworth and Sen, 2018). Additionally, Dümbgen et al. (2017) introduced the class of bi-log-concave distribution functions to include distributions with multimodal densities. However, an issue with the assumption of both log-concavity and bi-log-concavity is that the corresponding density functions inherit the requirement of exponentially decaying tails from the class of log-concave densities. This excludes distribution functions with tails that decay more slowly than exponentially, which are necessary to model data with heavy tails or outliers. Consequently, new constraints need to be proposed.

Example 1.1 (The salary data from Wooldridge (2000))**.** The dataset consists of the annual salaries of 177 randomly selected CEOs from companies in the U.S. in 1990, rounded to the nearest thousand USD. The histogram of this dataset is displayed in Figure 1, which suggests the presence of one or two outliers. Additionally, the histogram is highly skewed with a heavy right tail and appears to have multiple modes. These characteristics indicate that classical shape-constraints, such as unimodality or light-tailedness, may

not be suitable for modeling this salary data. Instead, the bi-s^*-concave classes proposed in Chapter 2 are more appropriate for modeling this dataset.

Heterogeneity, which refers to a mixture of components, is a common phenomenon in multi-component systems, as described above. Examples of such systems include fishery data (Titterington et al., 1985), single-cell RNA-sequencing (scRNA-seq) data (Haque et al., 2017), clinical data (Kriston, 2013), and highway accident data (Mannering et al., 2016); see also Li and Reynolds (1995) for more examples. Increasing levels of heterogeneity can make classical models fail to fit the distribution of the data. For instance, Poisson models and their variants (such as the over-dispersed Poisson model (Robinson et al., 2010), Poisson-Gamma model (Love et al., 2014; Huang et al., 2018), Poisson-Beta model (Vu et al., 2016), Poisson-log normal model (Silva et al., 2019), and zero-inflated mixture Poisson linear models (Liu et al., 2019)) are widely used to fit bulk RNA-sequencing data, which measures gene expression of an entire sample without differentiating among cells within the sample. However, the development of scRNA-seq, which allows the detection and quantification of gene expression in individual cells with high resolution and on a transcriptomic scale, has revealed previously unappreciated levels of heterogeneity and complexity in scRNA-seq data. This has made classical statistical methods fail to capture the distribution of scRNA-seq data, necessitating the development of new models, as demonstrated in Sarkar and Stephens (2021) through the use of mixture models. The histogram of heterogeneous data often exhibits various modes, which can also indicate the presence of heterogeneity. We illustrate this using an example of scRNA-seq data in Example 1.2.

Example 1.2 (The scRNA-seq data from Velmeshev et al. (2019))**.** The data consists of gene expression measurements for 18,041 genes in 23 subjects, and an example of the expression of gene KIF1B is shown in the barplot in Figure 2. The barplot suggests that the distribution of gene expression for KIF1B is likely to be multimodal, which violates the assumptions of most parametric Poisson mixture models such as the over-dispersed Poisson model and Poisson-Gamma model mentioned earlier.

Sparsity is a common phenomenon observed in large contingency tables for multivariate binary random variables. For instance, functional disability data (Erosheva et al., 2007), the Rochdale data in (Whittaker, 1990), and the NLTCS data (see Example 1.3) are some specific examples of such tables. The Ising model is one of the most widely used statistical methods for contingency table analysis in various scientific fields such as biological sciences, natural language processing, and data mining (Bishop et al., 1975; Christensen, 1997). However, despite its widespread use and great interpretability, its ability to fit data well can be limited, especially applied to sparse contingency tables. On the other hand, several classes of mixture models have been proposed as alternatives. For example, finite mixtures of Bernoulli products, known as

multivariate Bernoulli mixture models, have shown promising results in both applications (Juan and Vidal, 2002, 2004) and theory (Carreira-Perpinán and Renals, 2000; Allman et al., 2009). Other examples include parallel factor analysis (PARAFAC) (Bro, 1997), simplex factor models (Bhattacharya and Dunson, 2012), sparse PARAFAC (Zhou et al., 2015), Tucker decomposition (De Lathauwer et al., 2000), collapsed Tucker decomposition (Johndrow et al., 2017). These models are effective at fitting sparse contingency tables, however, they are not easily interpretable, especially when it comes to inferring multivariate associations between variables. Therefore, we will need a method which not only has great power to fit the data well, but also has strong interpretability.

Example 1.3 (National Long Term Care Survey (NLTCS) data)**.** The data utilized is extracted from the National Long Term Care Survey (NLTCS) data set created by the Center of Demographic Studies at Duke University. It includes eight binary variables that measure functional disability in daily living activities such as (1) eating, (2) getting around inside, (3) dressing, (4) cooking, (5) grocery shopping, (6) getting about outside, (7) traveling, and (8) managing money. Each measure classifies subjects as healthy (level 1) or disabled (level 2). The data is cross-classified for elderly individuals aged 65 and above, pooled across four survey waves from 1982, 1984, 1989, and 1994. With a sample size of 21574, the resulting 2^8 contingency table is sparse, Table 1, with 17.1% cells having 0 counts, 46.9% cells having small counts no larger than 5, and richest 1.6% cells bagged 40.5% observations.

4419	97	67	472	2063	55	335	44	313	18	33	76	1	5	2	6
119	115	1	16	0	4	1189	17	112	6	130	64	529	52	453	56
2	22	13	116	10	67	47	0	2	0	1	92	0	4	0	4
1	12	5	19	1	0	0	3	1	0	354	2	27	4	16	5
55	3	24	1	0	0	1	7	1	60	667	29	601	14	1	16
3	55	8	85	7	65	69	400	24	5	62	2	10	164	0	8
2	6	3	15	3	5	0	0	0	0	0	0	3	32	4	41
1	0	1	0	1	0	4	1	3	3	9	0	0	0	0	0
14	226	11	140	5	0	2	2	10	0	7	3	3	11	31	3
0	4	0	2	125	8	134	81	654	34	0	34	1	5	25	215
8	80	30	5	105	19	50	1	1	0	2	3	0	3	0	3
13	9	4	1	0	0	4	7	1	6	6	54	3	0	1	0
0	1	6	0	6	1	42	3	28	48	207	12	0	5	0	2
4	34	1	13	6	38	549	19	180	21	196	27	2	14	72	88
8	3	0	0	2	8	11	3	15	9	5	19	3	26	0	28
29	158	10	89	5	66	764	66	86	8	175	7	151	131	516	1056

Table 1: The NLTCS data. The cells counts appear row by row in lexicographical order with Variable 8 varying fastest and Variable 1 varying slowest.

To analyze these heavy-tailed, heterogeneous or sparse data illustrated above, we want to generalize classical models to make them more powerful and this motivates the idea of mixture models.

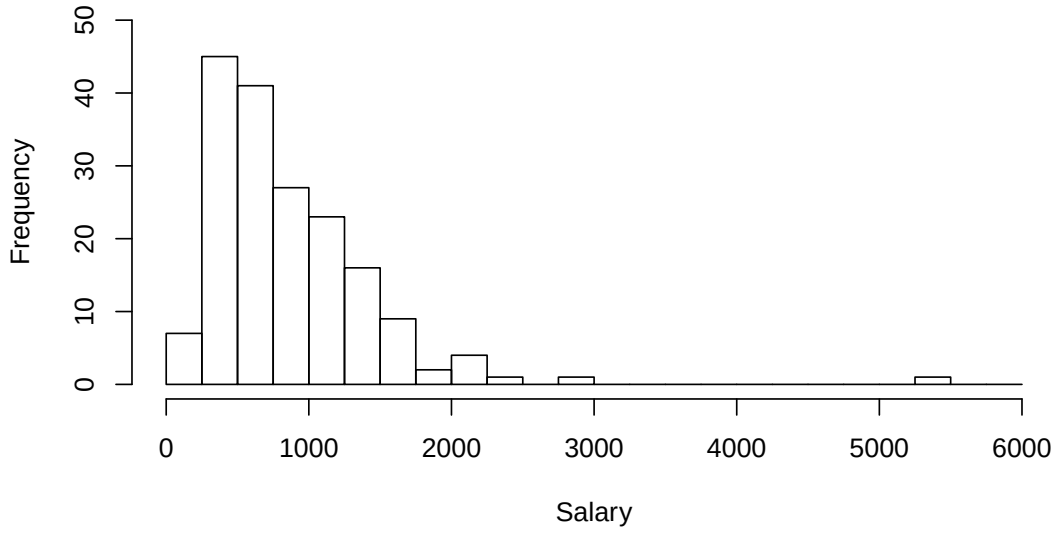

Figure 1: Histogram of the annual salaries of CEOs for $n = 177$ randomly chosen companies, rounded to multiples of 1000 USD.

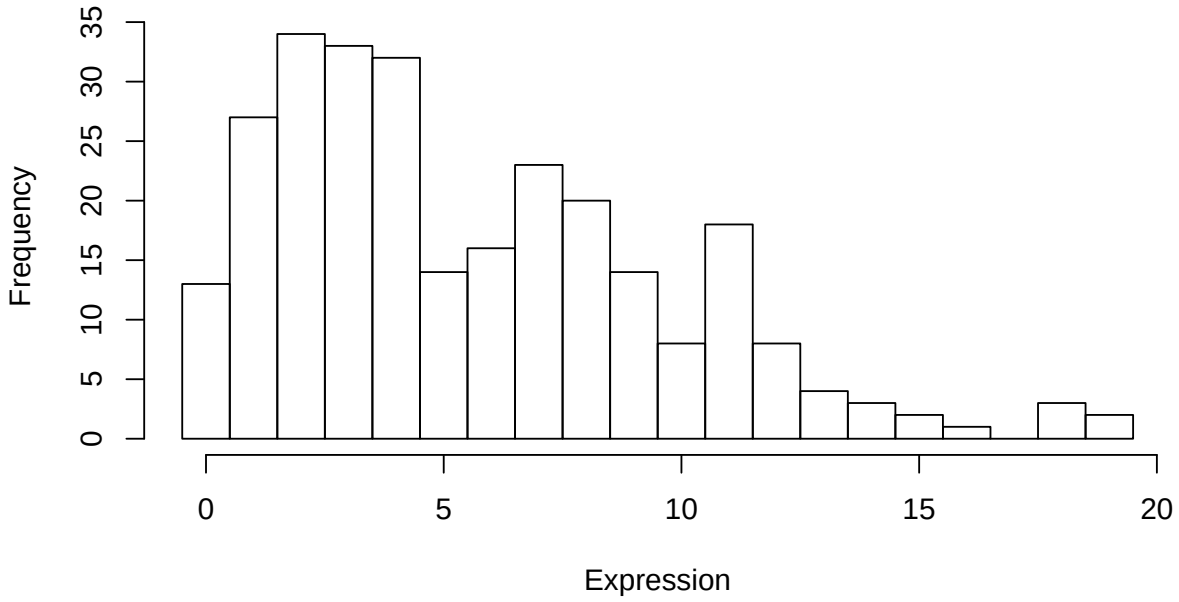

Figure 2: Barplot of expression of gene KIF1B.

1.2 Mixture models

Mixture models have become a popular tool across various fields, such as biometrics (Hosseinzadeh and Krishnan, 2008), genetics (Snipen et al., 2009), medicine (De Angelis et al., 1999), marketing (Gupta and Chintagunta, 1994), and clustering (McLachlan and Basford, 1988). They are used to capture specific properties of real data, such as heavy-tailedness (Swami, 2000), heterogeneity (Lubke and Muthén, 2005), and sparsity (Yu et al., 2011). The basic idea of a mixture model is to assume that outliers, multiple modes, or excess zeros come from several different subpopulations. Mixture models are mainly classified as finite or nonparametric.

1.2.1 Finite mixture models

Finite mixture models are widely used in statistical modeling involving discrete latent variables, such as clustering or latent class models. They assume that the observed data comes from a finite number of subpopulations or components within the population. The concept of finite mixture models dates back to Feller (1943), and includes popular examples such as normal mixtures (Pearson, 1894) and Poisson mixtures (Feller, 1943) that are special cases of mixtures from the general exponential family (Barndorff-Nielsen, 1978; Shaked, 1980). Estimation of the number of subpopulations and diagnosing their similarity is studied in Ray and Lindsay (2005); Leisch (2004), generalizing previous results from Robertson and Fryer (1969); Behboodian (1970). Identifiability in finite mixture models is studied by Teicher (1961) for mixtures of Poisson distributions, Teicher (1963) for mixtures of Gamma (or of normal) distributions, Yakowitz and Spragins (1968) for mixtures of products of exponential distributions. See also Chandra (1977); Redner and Walker (1984); Crawford (1994). The maximum likelihood estimation (MLE) for finite mixture models is one of the most popular estimation methods, and its consistency has been studied in Cheng and Liu (2001); Atienza et al. (2007); Tan et al. (2007), while Chen (1995); Kim (2014); Brunel (2019) establish the convergence rate. For further properties and inferences related to finite mixture models, Frühwirth-Schnatter (2006) provides a comprehensive review.

1.2.2 Nonparametric mixture models

Replacing discrete latent variables with continuous latent variables with a nonparametric unknown distribution leads to the definition of nonparametric mixture models. Results on nonparametric mixture models mainly focus on the nonparametric maximum likelihood estimation (NPMLE). Earlier results on the existence, discreteness (of the NPMLE support), and computation include Simar (1976) for nonparametric Poisson mixture, Jewell (1982) for nonparametric exponential mixture, Laird (1978); Lindsay (1983a,b);

Lindsay and Roeder (1993) for general nonparametric mixture models; see also Lindsay (1995) for a survey. Consistency of NPMLEs were established in Kiefer and Wolfowitz (1956) and Pfanzagl (1988); see also (Chen, 2017) for a survey. Other estimation methods include kernel estimators (Zhang, 1995; Loh and Zhang, 1996), orthogonal polynomial mixing density estimators (Hengartner, 1997; Loh and Zhang, 1997), and projection estimators (Roueff and Rydén, 2005).

1.3 Structure of the thesis

In this thesis, we explore three classes of distributions for different data types under the idea of mixture models. Specifically, we introduce the bi-s^*-concave distribution for continuous data, the nonparametric Poisson mixture distribution for count data (e.g., scRNA-seq data), and the Ising mixture distribution for sparse contingency tables. Each of these classes will be described in detail in the following chapters.

1.3.1 The bi-s^*-concave distribution

The bi-s^*-concave distribution is a proposed generalization of well-studied density classes in estimation under shape constraints and mixture models. While estimation under shape constraints has been well-developed for the past 15-20 years, most constraints assume the unimodality of density functions (e.g., monotonicity (Grenander, 1956; van Eeden, 1956), log-concavity (Walther, 2009; Samworth and Sen, 2018), and s-concavity (Dharmadhikari and Joag-Dev, 1988; Han and Wellner, 2016)), and do not include mixture distributions with multimodal densities. To include multimodal density functions, Dümbgen et al. (2017) introduced the class of bi-log-concave distribution functions, which include normal mixtures under certain conditions. However, the assumption of bi-log-concavity has the difficulty that the corresponding density functions inherit the requirement of exponentially decaying tails of the class of log-concave densities. This restriction eliminates the ability to model data with heavy tails or outliers. Chapter 2 presents the work with Nilanjana Laha and Jon A. Wellner (Laha et al., 2021) that introduces new shape-constrained families of distribution functions called bi-s^*-concave distributions, which allow for tails that decay more slowly or more rapidly than exponentially. The bi-s^*-concave classes generalize the bi-log-concave class and include not only normal mixtures but also Student-t mixtures under some conditions.

1.3.2 Nonparametric Poisson mixture distribution

The nonparametric Poisson mixture distribution is a generalization of the Poisson distribution, where the parameter is assumed to follow from a completely unknown mixing distribution. It serves as a generalization of

classical Poisson models and their variations, such as over-dispersed Poisson, Poisson-Gamma, Poisson-Beta, and Poisson-log normal, allowing for the capture of heavy-tailedness, heterogeneity, and sparsity in scRNA-seq data. Chapter 3 presents the work with Weihao Kong, Ramya Korlakai Vinayak, Wei Sun, and Fang Han (Miao et al., 2022) that focuses on nonparametric Poisson mixture models adapted for single-cell RNA sequencing data. In terms of theoretical contributions, we establish the existence and consistency of the adapted nonparametric maximum likelihood estimators. Additionally, we investigate and establish the minimax convergence rate of the maximum likelihood estimators of nonparametric Poisson mixture models. In terms of computation, we provide three algorithms for obtaining the nonparametric maximum likelihood estimators. We conduct time complexity analyses for each algorithm and perform simulations and real-data applications to demonstrate their effectiveness.

1.3.3 Ising mixture distributions

In Chapter 4, we introduce Ising mixture models as a novel approach to fitting multivariate binary random variables, which is a joint work with Adrian Dobra and Yen-Chi Chen (Miao et al., 2023). The proposed model is a generalization of Ising models and multivariate Bernoulli mixture models, and it is not only effective in fitting sparse contingency tables but also provides interpretable results. We propose a set of sufficient and necessary assumptions for the identifiability of Ising mixture models. Our proof technique provides another means of establishing identifiability, which is valuable since traditional proof tools are unavailable due to the lack of the conditional independence assumption. To address the non-identifiability issue, we propose using the Bayesian framework. For each interaction effect, we use continuous spike-and-slab prior distributions that contain Bernoulli random variables, called association indicators, to indicate the existence of interaction effects between variables. The posterior means of these association indicators provide a solution to infer associations between binary variables, which is meaningful even in cases where the model is non-identifiable.

2 Bi-s^*-concave distributions

2.1 Introduction

Statistical methods based on shape constraints have been developing rapidly during the past 15 - 20 years. From the classical univariate methods based on monotonicity going back to the work of Grenander (1956) and van Eeden (1956) in the 1950's and 1960's, research has progressed to consideration of convexity type constraints in a variety of problems including estimation of density functions, regression functions, and other "nonparametric" functions such as hazard (rate) functions. See Samworth and Sen (2018) for a summary and overview of some of this recent activity.

One very appealing shape constraint is log-concavity: a (density) function $f : \mathbb{R}^d \to [0, \infty]$ is *log-concave* if $\log f$ is concave (with $\log 0 = -\infty$). See Samworth (2018) for a recent review of the properties of log-concave densities and their relevance for statistical applications. While much of the current literature has focused on point estimation, our main focus here will be on inference for one-dimensional distribution functions and especially on (honest, exact) confidence bands for distribution functions which take advantage of shape constraints.

To this end, Dümbgen et al. (2017) introduced the class of *bi-log-concave* distribution functions defined as follows: a distribution function F on $\mathbb{R}$ is bi-log-concave if both F and $1 - F$ are log-concave. They provided several different equivalent characterizations of this property, and noted (the previously known fact) that if f is a log-concave density, then the corresponding distribution function F and survival function $1 - F$ are both log-concave. But the converse is false: there are many bi-log-concave distribution functions F with density f which fail to be log-concave; see Chapter 2.2 below for an explicit example. Dümbgen et al. (2017) also showed how to construct confidence bands which exploit the bi-log-concave shape constraint and thereby obtain narrower bands, especially in the tails, with correct coverage when the bi-log-concave assumption holds.

However, a difficulty with the assumption of bi-log-concavity is that the corresponding density functions inherit the requirement of exponentially decaying tails of the class of log-concave densities, and this rules out distribution functions F with tails decaying more slowly than exponentially. Here we introduce new shape-constrained families of distribution functions F, which we call the *bi-s^*-concave distributions*, with tails possibly decaying more slowly (or more rapidly) than exponentially. As the name indicates, these families involve a parameter $s^* \in (-\infty, 1]$ which allows heavier than exponential tails when $s^* < 0$, lighter than exponential tails when $s^* > 0$, and which correspond to exactly the bi-log-concave class introduced by Dümbgen et al. (2017) when $s^* = 0$.

Here is an outline of the rest of the chapter. In subchapter 2.2 we give careful definitions of the new classes of *bi-s^*-concave distributions*. We also present several helpful examples and discuss some basic properties of the new classes and their connections to the classes of s-concave densities studied by Borell (1975), Brascamp and Lieb (1976), and Rinott (1976). (See also Dharmadhikari and Joag-Dev (1988), and Gardner (2002).) Chapter 2.3 contains the main theoretical results of the chapter. The connection between the bi-s^*-concave class and a key condition in the theory of quantile processes, the Csörgő - Révész condition, is discussed in Corollary 2.1. Finally, we give two tail bounds for distribution functions $F \in \mathcal{P}_{s^*}$, see Corollary 2.2.

In Chapter 2.4 we first introduce the new confidence bands for a distribution function $F \in \mathcal{P}_{s^*}$ assuming s^* is known. We also discuss some of their theoretical properties: the consistency of confidence bands is discussed in Theorem 2.2, and Theorem 2.3 provides a rate of convergence for linear functionals of bi-s^*-distribution functions contained in the bands. This extends theorem 5 of Dümbgen et al. (2017). We then briefly discuss the algorithms used to compute the new bands, and illustrate the new bands with real and artificial data. Chapter 2.5 gives a brief summary and statements of further problems. An especially important remaining problem concerns construction of confidence bands when s^* is unknown. The proofs for all the results in Chapters 2.2, 2.3, and 2.4 are given in Chapters 2.6.

We conclude this chapter with some notation which will be used throughout the rest of the chapter. The supremum norm of a function $h : \mathbb{R} \to \mathbb{R}$ is denoted by $\|h\|_\infty := \sup_{x\in\mathbb{R}} |h(x)|$, and for $K \subset \mathbb{R}$ we write $\|h\|_{K,\infty} := \sup_{x\in K} |h(x)|$. For a function $x \mapsto f(x)$,

$$f'_+(x) := \lim_{\lambda\downarrow 0} \frac{f(x+\lambda) - f(x)}{\lambda}, \qquad \text{and} \quad f'_-(x) := \lim_{\lambda\uparrow 0} \frac{f(x+\lambda) - f(x)}{\lambda},$$
$$f(x+) := \lim_{y\downarrow x} f(y), \qquad \text{and} \quad f(x-) := \lim_{y\uparrow x} f(y),$$

assuming that the indicated limits exist. In general, we use F and f to denote a distribution function and the corresponding density function with respect to Lebesgue measure, and we set $J(F) := \{x \in \mathbb{R} : 0 < F(x) < 1\}$.

2.2 Definitions, Examples, and First Properties

As we discussed above, for distribution functions F on $\mathbb{R}$, Dümbgen et al. (2017) introduced a shape constraint they called *bi-log-concavity* by requiring that both F and $1 - F$ be log-concave.

In this chapter, we generalize the bi-log-concave distribution functions by introducing and studying bi-s^*-concave distributions defined as follows.

Definition 2.1. *A distribution function F is bi-s^*-concave if it is continuous on $\mathbb{R}$ and satisfies the following*

properties on $J(F)$:

- *For $-\infty < s^* < 0$, both $x \mapsto F^{s^*}(x)$ and $x \mapsto (1 - F(x))^{s^*}$ are convex functions on $J(F)$.*
- *For $s^* = 0$, both $x \mapsto \log(F(x))$ and $x \mapsto \log(1 - F(x))$ are concave functions on $J(F)$.*
- *For $0 < s^* \leq 1$, both $x \mapsto F^{s^*}(x)$ and $x \mapsto (1 - F(x))^{s^*}$ are concave functions on $J(F)$.*

The class of bi-s^*-concave distribution functions is denoted by $\mathcal{P}_{s^*}$, i.e.

$$\mathcal{P}_{s^*} := \{F : F \text{ is bi-}s^*\text{-concave}\}.$$

Recall that a density function f is s-concave if f^s is convex for $s < 0$, f^s is concave for $s > 0$, and $\log f$ is concave for $s = 0$. Two basic properties linking s-concave densities and bi-s^*-concave distribution functions are given in the following two propositions. Proposition 2.1 generalizes the case $s = 0$ as noted above, while Proposition 2.2 generalizes the corresponding nestedness property of the classes of s-concave densities; see e.g. Dharmadhikari and Joag-Dev (1988), page 86, and Borell (1975), page 111.

Proposition 2.1. *Suppose a density function f is s-concave with $s \in (-1, \infty)$. Then the corresponding distribution function F is bi-s^*-concave for all $s^* \leq s/(1 + s)$.*

Proposition 2.2. *The bi-s^*-concave classes are nested in the following sense:*

$$\mathcal{P}_{s^*} \subset \mathcal{P}_{t^*}, \quad \text{whenever } t^* \leq s^* \leq 1. \tag{2.1}$$

Moreover, the bi-s^*-concave classes are continuous at $s^* = 0$ in the following sense:

$$\bigcup_{s^*>0} \mathcal{P}_{s^*} = \mathcal{P}_0 = \bigcap_{s^*<0} \mathcal{P}_{s^*}. \tag{2.2}$$

In view of the nesting property (2.1), for each $F \in \mathcal{P}_{s^*}$ for some s^* we define

$$s_0^*(F) := \sup\{s^* : \ F \text{ is bi-}s^*\text{-concave}\}.$$

Similarly if f is s-concave for some s we define

$$s_0(f) := \sup\{s : \ f \text{ is } s\text{-concave}\}.$$

We often drop the subscript 0 if the meaning is clear. For other basic properties of s-concave densities and bi-s^*-concave distribution functions, including results concerning closure under convolution, see Borell (1975), Dharmadhikari and Joag-Dev (1988), and Saumard (2019).

Now we introduce two important parameters, one of which will appear in connection with our characterization of the class of bi-s^*-concave distribution functions in the next chapter and in our examples below.

The Csörgő - Révész constant of a bi-log-concave distribution function F, denoted by $\tilde{\gamma}(F)$, is given by

$$\tilde{\gamma}(F) := \text{esssup}_{x\in J(F)} F(x)(1-F(x))\frac{|f'(x)|}{f^2(x)}, \tag{2.3}$$

provided that F is differentiable on $J(F) := \{x \in \mathbb{R} : 0 < F(x) < 1\}$ with derivative $f := F'$ and f is differentiable almost everywhere on $J(F)$ with derivative $f' = F''$. Here the essential supremum is with respect to Lebesgue measure. Alternatively (and suited for our characterization Theorem 2.1),

$$\gamma(F) := \text{esssup}_{x\in J(F)} \{F(x) \wedge (1-F(x))\}\frac{|f'(x)|}{f^2(x)}. \tag{2.4}$$

Note that since $u \wedge (1-u) \le 2u(1-u) \le 2\{u \wedge (1-u)\}$ it follows that $2^{-1}\gamma(F) \le \tilde{\gamma}(F) \le \gamma(F)$, and hence finiteness of $\gamma(F)$ is equivalent to finiteness of $\tilde{\gamma}(F)$. The Csörgő - Révész constant $\tilde{\gamma}(F)$ arises in the study of quantile processes and transportation distances between empirical distributions and true distributions on $\mathbb{R}$: see Csörgő and Révész (1978), Shorack and Wellner (2009), Barrio et al. (2005), and Bobkov and Ledoux (2019). It follows from the characterization theorem 1(iv) of Dümbgen et al. (2017) that F is bi-log-concave if and only if $\overline{\gamma}(F) \le 1$. We will define $\overline{\gamma}(F) \ge \gamma(F)$ and generalize this to the classes of bi-s^*-concave distribution functions in Chapter 2.3.

Now we consider several examples of s-concave densities and bi-s^*-concave distribution functions.

Example 2.1 (Student-t)**.** Suppose $x \mapsto f_r(x)$ is the density function of the Student-t distribution with r degrees of freedom defined as follows:

$$f_r(x) = \frac{\Gamma((r+1)/2)}{\sqrt{\pi}\Gamma(r/2)}\left(1+\frac{x^2}{r}\right)^{-(r+1)/2} \quad \text{for } x \in \mathbb{R}.$$

It is well-known (see e.g. Borell (1975)) that f_r is s-concave for any $s \le -1/(1+r) = s_0(f_r)$. Note that s takes values in $(-1,0)$ since $r \in (0,\infty)$. It follows from Proposition 2.1 that $F_r^{s^*}$ and $(1-F)^{s^*}$ are convex for $s^* = s/(1+s) = -1/r = s_0^*(F_r) < 0$, and hence F_r is bi-s^*-concave for all $0 < r < \infty$. Direct calculation shows that the Csörgő - Révész constant $\gamma(F_r) = 1 - s^* = 1 + (1/r) \in (1,\infty)$ for $0 < r < \infty$.

In particular, this yields $\gamma(F_1) = \gamma(Cauchy) = 2$. And it suggests that $\gamma(F) \le 1/(1+s) = 1 - s^*$ for all bi-s^*-concave distribution functions F where $1/(1+s)$ varies from 1 to ∞ as s varies from 0 to -1. This is one of the characterizations of the bi-s^*-concave class that we will prove in Chapter 2.3.

Example 2.2 ($F_{a,b}$)**.** Suppose that $f_{a,b}$ is the family of $F-$distributions with "degrees of freedom" $a > 0$ and $b > 0$. (In statistical practice, if T has the density $f_{a,b}$, this would usually be denoted by $T \sim F_{a,b}$, where a is the "numerator degrees of freedom" and b is the "denominator degrees of freedom".) The density is given by

$$f_{a,b}(x) = C_{a,b}\frac{x^{b/2-1}}{(a+bx)^{(a+b)/2}} \text{ for } x \ge 0.$$

(In fact, $C(a,b) = a^{a/2}b^{b/2}\text{Beta}(a/2, b/2)$, and $f_{a,b}(x) \to g_b(x)$ as $a \to \infty$ where g_b is the Gamma density with parameters $b/2$ and $b/2$.) It is well-known (see e.g. Borell (1975)) that $f_{a,b}$ belongs to the class of s-concave densities, if $s \leq -1/(1+a/2) = s_0(f_{a,b})$ when $a \geq 2$ and $b \geq 2$. This implies that $s \in [-1/2, 0)$, and the resulting $s_0^* = s/(1+s) = -2/a$ is in $[-1, 0)$. By Proposition 2.1, it follows that F^{s^*} and $(1-F)^{s^*}$ are convex; i.e. F is bi-s^*-concave.

Example 2.3 (Pareto)**.** Suppose that $f_{a,b} = (a/b)(x/b)^{-(a+1)}1_{[b,\infty)}(x)$, the Pareto distribution with parameters $a > 0$ and $b > 0$. In this case, $f_{a,b}$ is s-concave for each $s \leq -1/(1+a)$ by noting the convexity of $f_{a,b}^{-1/(1+a)} = (x/b) \cdot (b/a)^{1/(1+a)}$.

Thus we take $s = -1/(1+a) \in (-1, 0)$ for $a \in (0, \infty)$ and hence $s^* = s/(1+s)$ equals $-1/a$. Furthermore, it is easily seen that

$$CR_R(x) := (1 - F(x))\frac{-f'(x)}{f^2(x)} = 1 - s^* = 1 + 1/a \text{ for all } x > b\,.$$

($CR_R(\cdot)$ represents the Csörgő - Révész function in the right tail.)

Thus the Pareto distribution is analogous to the exponential distribution in the log-concave case in the sense that $x \mapsto f^s(x) = cx$ (with $c = b^{-1}(b/a)^{1/(1+a)}$) is linear.

Example 2.4. (Symmetrized Beta) Suppose that

$$f_r(x) = C_r(1 - x^2/r)^{r/2}1_{[-\sqrt{r},\sqrt{r}]}(x),$$

where

$$C_r = \Gamma((3+r)/2)/(\sqrt{\pi r}\Gamma(1+r/2))$$

and $r \in (0, \infty)$. Note that f_r is an s-concave density with $s = 2/r \in (0, \infty)$ since

$$f_r^{2/r}(x) = C_r^{2/r}(1 - x^2/r)1_{[-\sqrt{r},\sqrt{r}]}$$

is concave and hence the corresponding distribution function F_r is bi-s^*-concave with $s^* = s/(1+s) = 2/(2+r)$. As $r \to \infty$ it is easily seen that

$$f_r(x) \to (2\pi)^{-1/2}\exp(-x^2/2),$$

the standard normal density. Thus $r = \infty$ corresponds to $s = 0$ and $s^* = 0$. On the other hand,

$$g_r(x) := \sqrt{r}f_r(\sqrt{r}x) = \sqrt{r}C_r(1 - x^2)^{r/2}1_{[-1,1]}(x) \to 2^{-1}1_{[-1,1]}(x)$$

as $r \to 0$. Thus $r = 0$ corresponds to $s = \infty$ and $s^* = 1$.

Note that just as the class of bi-log-concave distributions is considerably larger than the class of log-

concave distributions (as shown by Dümbgen et al. (2017)), the class of bi-s^*-concave distributions is considerably larger than the class of s-concave distributions. In particular, multimodal distributions are allowed in both the bi-log-concave and the bi-s^*-concave classes.

Example 2.5. (Exponential family; exponential tilt of $U(0,1)$) Suppose that

$$f_t(x) = \exp(tx - K(t))1_{[0,1]}(x)$$

where

$$K(t) := \begin{cases} \log(e^t - 1) - \log t, & t > 0, \\ 0, & t = 0, \\ \log(1 - e^t) - \log(-t), & t < 0, \end{cases} \tag{2.5}$$

for $-\infty < t < \infty$ with $K(0) := 0$, and further define $F_t(x) := \int_0^x f_t(y)dy$.

One can verify that f_t is s-concave only for $s \leq 0$ and hence F_t is bi-s^*-concave for $s^* \leq s/(1+s) \leq 0$ by Proposition 2.1. However, this might not be optimal; i.e. there remains the possibility that $F \in \mathcal{P}_{s^*}$ for some $s^* > 0$. In fact, by Theorem 2.1(iv) it follows that $F_t \in \mathcal{P}_{s^*}$ with $s^* = e^{-|t|}$. (For an example involving a power-tilt of $U(0,1)$, see Dharmadhikari and Joag-Dev (1988) (iv), page 95.) This also implies that the converse of Proposition 2.1 does not hold here or in general. The following two examples also illustrate this point.

Example 2.6. (Mixture of Gaussians shifted) (Dümbgen et al. (2017), page 2-3) Suppose that f_δ is the mixture $(1/2)N(-\delta, 1) + (1/2)N(\delta, 1)$ with $\delta > 0$. It is well-known that f_δ is bimodal if $\delta > 1$. Since all s-concave densities are unimodal (see e.g. Dharmadhikari and Joag-Dev (1988) page 86), it follows that f_δ is not s-concave for any $\delta > 1$. Dümbgen et al. (2017) showed (numerically) that the corresponding distribution F_δ is bi-log-concave for $\delta \leq 1.34$ but not for $\delta \geq 1.35$. With $\delta = 1.8$ this example also shows that strict inequality can occur in the second inequality in Corollary 2.1 below.

Example 2.7. (Mixture of shifted Student-t) Now suppose that f is the mixture $(1/2)t_1(\cdot - \delta) + (1/2)t_1(\cdot + \delta)$ with $\delta > 0$ where t_r is the standard Student-t density with r degrees of freedom as in example 1. Since f_δ is bimodal if $\delta > \delta_0 \approx 0.6$ and all s-concave densities are unimodal, it follows that f_δ is not s-concave for any $\delta > \delta_0$. For values of $\delta < \delta_0$, f_δ is s-concave with $s = -1/2$, so proposition 1 applies and shows that F_δ is bi-s^*-concave with $s^* = -1$. By numerical calculation, for $\delta > \delta_0$ the distribution functions F_δ are bi-s^*-concave for some $s^* = s^*(\delta) \in (-\infty, 1]$ which decreases (approximately linearly) for large δ.

Example 2.8 (Lévy with $\alpha = 1/2$)**.** This example is the completely asymmetric $\alpha-$stable (or Lévy) law with $\alpha = 1/2$. It gives the first passage time to the level $a > 0$ for a standard Brownian motion B (started at 0

and with no drift). See e.g. Durrett (2019), pages 372 - 374. The density is given by

$$f_a(t) = \frac{a}{\sqrt{2\pi t^3}} \exp(-a^2/2t) 1_{(0,\infty)}(t),$$

and the distribution function $F_a(t) = 2P(B_t \geq a) = 2(1 - \Phi(a/\sqrt{t}))$. It is easily seen that f_a is s-concave with $s = -2/3$, and hence F_a is bi-s^*-concave with $s^* = -2$. Thus $\gamma(F) = 3$.

The following table summarizes the examples:

Table 2: Summary of examples 1-8

Name	example	density f	d.f. F	s	$s^* = s/(1+s)$	$\overline{\gamma}(F) = 1 - s^*$
student-t	1	$f_r,\ r > 0$	F_r	$-1/(1+r)$	$-1/r$	$1 + (1/r)$
$F_{a,b}$	2	$f_{a,b},\ a, b > 0$	$F_{a,b}$	$-1/(1+a/2)$	$-2/a$	$1 + 2/a$
Pareto(a, b)	3	$f_{a,b},\ a, b > 0$	$F_{a,b}$	$-1/(1+a)$	$-1/a$	$1 + 1/a$
Symmetric Beta	4	$f_r,\ r > 0$	F_r	$2/r$	$2/(r+2)$	$1/(1+2/r) = r/(r+2)$
Expo family Tilted $U(0,1)$	5	$f_t,\ t \in \mathbb{R}$	F_t	0	$e^{-\lvert t\rvert}$	$1 - e^{-\lvert t\rvert}$
Mixture, $N(\delta,1)$, $N(-\delta,1)$	6	f_δ	F_δ	not s-concave for $\delta > 1$	0 for $0 < \delta < 1.34$	1 $0 < \delta < 1.34$
Mixture, $T(\delta,1)$, $T(-\delta,1)$	7	f_δ	F_δ	not s-concave $\delta > .6$	bi-s^*-concave, some s^* $0 < \delta < \infty$	2 δ small
Lévy $\alpha = 1/2$	8	f_a	F_a	$-2/3$	-2	3

Example 5 shows that strict inequality can hold in the inequality $\gamma(F) \leq \overline{\gamma}(F)$.

2.3 Main Theoretical Results

Here is our theorem characterizing bi-s^*-concave distribution functions.

Theorem 2.1. *Let $s^* \leq 1$. For a non-degenerate distribution function F, the following statements are equivalent:*

(i) F is bi-s^*-concave.

(ii) F is continuous on $\mathbb{R}$ and differentiable on $J(F)$ with derivative $f = F'$.

Moreover, for $s^* \neq 0$,

$$F(y) \begin{cases} \leq F(x) \cdot \left(1 + s^* \frac{f(x)}{F(x)}(y-x)\right)_+^{1/s^*} \\ \geq 1 - (1 - F(x)) \cdot \left(1 - s^* \frac{f(x)}{1-F(x)}(y-x)\right)_+^{1/s^*} \end{cases} \tag{2.6}$$

while for $s^* = 0$

$$F(y) \begin{cases} \leq F(x) \cdot \exp\left(\frac{f(x)}{F(x)}(y-x)\right) \\ \geq 1 - (1-F(x)) \cdot \exp\left(-\frac{f(x)}{1-F(x)}(y-x)\right) \end{cases} \tag{2.7}$$

for all $x, y \in J(F)$.

(iii) F *is continuous on* $\mathbb{R}$ *and differentiable on* $J(F)$ *with derivative* $f = F'$ *such that the* s^**-hazard function* $f/(1-F)^{1-s^*}$ *is non-decreasing on* $J(F)$*, and the reverse* s^**-hazard function* f/F^{1-s^*} *is non-increasing on* $J(F)$.

(iv) F *is continuous on* $\mathbb{R}$ *and differentiable on* $J(F)$ *with bounded and strictly positive derivative* $f = F'$*. Furthermore,* f *is differentiable almost everywhere on* $J(F)$ *with derivative* $f' = F''$ *satisfying*

$$-(1-s^*)\frac{f^2}{1-F} \leq f' \leq (1-s^*)\frac{f^2}{F} \text{ almost everywhere on } J(F). \tag{2.8}$$

The following two remarks are immediately consequences of Theorem 2.1. See Chapter 2.6 for a proof of Remark 2.1.

Remark 2.1. (i) The proof of Theorem 2.1(iv) implies that if $s^* > 1$, then not both F^{s^*} and $(1-F)^{s^*}$ can be concave.

(ii) If F is a bi-s^*-concave distribution function for $0 < s^* \leq 1$, then $\inf J(F) > -\infty$ and $\sup J(F) < \infty$.

(iii) If F is a bi-s^*-concave distribution function for some s^*, then F is bi-$s_0^*(F)$-concave.

(iv) If F is a bi-s^*-concave distribution function with $s_0^*(F) < 0$, then $\int |x|^t dF(x) < \infty$ for all $t \in (0, -1/s_0^*(F))$.

Corollary 2.1. *(Connection with the Csörgő - Révész constant.)*

Suppose F *is a bi-*s^**-concave distribution function for* $s^* \leq 1$*. Then with* $\tilde{\gamma}(F)$ *and* $\gamma(F)$ *as defined in (2.3) and (2.4), we have*

$$\frac{1}{2}\gamma(F) \leq \tilde{\gamma}(F) \leq \gamma(F) \leq \overline{\gamma}(F) \leq 1 - s^*, \tag{2.9}$$

where

$$\overline{\gamma}(F) := \max\{\widetilde{CR}(F), \widetilde{CR}(\bar{F})\}, \ \bar{F} := 1 - F,$$

$$\widetilde{CR}(F) := \operatorname{esssup}_{x \in J(F)} \frac{F(x)F''(x)}{(F'(x))^2},$$

and

$$\gamma(F) := \operatorname{esssup}_{x \in J(F)} \frac{\{F(x) \wedge (1-F(x))\}|F''(x)|}{(F'(x))^2} = \operatorname{esssup}_{x \in J(F)} \frac{\{F(x) \wedge (1-F(x))\}|f'(x)|}{(f(x))^2}.$$

Remark 2.2. By Theorem 2.1, one can verify that $\widetilde{CR}(F)$ is well-defined for any $F \in \mathcal{P}_{s^*}$. Note that

$$\widetilde{CR}(\overline{F}) := \operatorname{esssup}_{x \in J(F)} \frac{\overline{F}(x)(-F''(x))}{(F'(x))^2}.$$

The first two inequalities in Corollary 2.1 follow (as we noted before) from $2^{-1}\{u \wedge (1-u)\} \le u(1-u) \le u \wedge (1-u)$ for $0 \le u \le 1$. Thus finiteness of $\tilde{\gamma}(F)$ implies finiteness of $\gamma(F)$ and vice-versa. Examples show that strict inequality may hold in the inner inequalities in (2.9). On the other hand, if f is non-decreasing on $(F^{-1}(0), F^{-1}(1/2))$ and f is non-increasing on $(F^{-1}(1/2), F^{-1}(1))$, then $\gamma = \overline{\gamma}$ by inspection of the proof of $\gamma(F) \le \overline{\gamma}(F)$.

Corollary 2.2 (Bounds for $F \in \mathcal{P}_{s^*}$, where $s^* \neq 0$)**.** *For any $s^* \in (-\infty, 0) \cup (0, 1]$ and $F \in \mathcal{P}_{s^*}$,*

$$F_L(x) \le F(x) \le F_U(x), \tag{2.10}$$

where $F_L(x) := [F^{s^*}(x) - (1-s^*)]/s^*$ and $F_U(x) := [1 - (1-F(x))^{s^*}]/s^*$.
Moreover, $F_U(x)$ is a convex function on $J(F)$, and $F_L(x)$ is a concave function on $J(F)$. For $s^* = 0$ and $F \in \mathcal{P}_0$, (2.10) holds with $F_L(x) = 1 + \log F(x)$ and $F_U(x) = -\log(1 - F(x))$.

2.4 Confidence bands for bi-s^*-concave distribution functions

Our goal in this chapter is to define confidence bands for F which exploit the shape constraint $F \in \mathcal{P}_{s_0^*}$. We start with some known unconstrained nonparametric bands and define new bands under the assumption that the true distribution function F satisfies the shape constraint $F \in \mathcal{P}_{s_0^*}$ where s_0^* is known.

2.4.1 Definitions and Basic Properties

Let $X_1, \ldots, X_n$ be i.i.d. random variables with continuous distribution function F. A $(1-\alpha)$-confidence band, denoted by (L_n, U_n), for F means that both L_n and U_n are monotonically non-decreasing functions on $\mathbb{R}$ depending on α and $X_1, \ldots, X_n$ only, moreover, L_n and U_n have to satisfy $L_n < 1$, $U_n > 0$ and

$$P\left(L_n(x) \le F(x) \le U_n(x) \text{ for all } x \in \mathbb{R}\right) = 1 - \alpha.$$

The following two bands are discussed in Dümbgen et al. (2017) and we briefly restate them here.

Example 2.9 (Komogorov-Smirnov band)**.** A Komogorov-Smirnov band (L_n, U_n) is given by

$$[L_n(x), U_n(x)] := \left[\mathbb{F}_n(x) - \frac{\kappa_{\alpha,n}^{KS}}{\sqrt{n}}, \mathbb{F}_n(x) + \frac{\kappa_{\alpha,n}^{KS}}{\sqrt{n}}\right] \cap [0,1],$$

where $\mathbb{F}_n$ is the empirical distribution function and $\kappa_{\alpha,n}^{KS}$ denotes the $(1-\alpha)$-quantile of $\sup_{x \in \mathbb{R}} n^{1/2} |\mathbb{F}_n(x) -$

$F(x)|$, see Shorack and Wellner (2009) Note that $\kappa_{\alpha,n}^{KS} \leq \sqrt{\log(2/\alpha)/2}$ by Massart's (1990) inequality, see Massart (1990).

Example 2.10 (Weighted Komogorov-Smirnov band)**.** A Weighted Komogorov-Smirnov band (L_n, U_n) is as follows: for any $\gamma \in [0, 1/2)$,

$$[L_n(x), U_n(x)] := \left[t_i - \frac{\kappa_{\alpha,n}^{WKS}}{\sqrt{n}} (t_i(1-t_i))^{\gamma}, t_{i+1} + \frac{\kappa_{\alpha,n}^{WKS}}{\sqrt{n}} (t_{i+1}(1-t_{i+1}))^{\gamma} \right] \cap [0,1],$$

for $i \in \{0, 1, \ldots, n\}$ and $x \in [X_{(i)}, X_{(i+1)})$, where $\{X_{(i)}\}_{i=1}^n$ denotes the order statistics of $\{X_i\}_{i=1}^n$, $X_{(0)} := -\infty$, $X_{(n+1)} := \infty$, $t_i := i/(n+1)$ for $i = 1, \ldots, n$, and $\kappa_{\alpha,n}^{WKS}$ denotes the $(1-\alpha)$-quantile of the following test statistics

$$\sqrt{n} \max_{i=1,\ldots,n} \frac{|F(X_{(i)}) - t_i|}{(t_i(1-t_i))^{\gamma}}.$$

Note that $\kappa_{\alpha,n}^{WKS} = O(1)$.

Example 2.11 (Owen's band refined)**.** This confidence band which is proposed by Owen (1995) and refined by Dümbgen and Wellner (2014) is as follows:

$$\begin{aligned}
&L_n(x) := 0 \text{ for } x < X_{(1)},\\
&L_n(x) := \min\{p \in (0, t_j] : K(t_j, p) \leq \gamma_n(t_j)\} \text{ for } 1 \leq j \leq n, X_{(j)} \leq x < X_{(j+1)},\\
&U_n(x) := \max\{p \in [t_j, 1) : K(t_j, p) \leq \gamma_n(t_j)\} \text{ for } 1 \leq j \leq n, X_{(j-1)} \leq x < X_{(j)},\\
&U_n(x) := 1 \text{ for } x \geq X_{(n)},
\end{aligned}$$

where $\{X_{(j)}\}_{j=1}^n$ denotes the order statistics of $\{X_j\}_{j=1}^n$, $t_j := j/(n+1)$ for $j = 1, \ldots, n$,

$$K(\hat{p}, p) := \hat{p} \log \frac{\hat{p}}{p} + (1-\hat{p}) \log \frac{1-\hat{p}}{1-p}, \gamma_n(t) := \frac{C(t) + \nu D(t) + \kappa_{n,\alpha}^{ODW}}{n+1},$$
$$C(t) := \log(1 + \text{logit}(t)^2/2)/2, D(t) := \log(1 + C(t)^2/2)/2, \text{logit}(t) := \log\left(t/(1-t)\right),$$

and $\kappa_{n,\alpha}^{ODW}$ denotes the $(1-\alpha)$-quantile of the following statistics:

$$\max_{j=1,\ldots,n} \left((n+1) K\left(t_j, F(X_{(j)})\right) - C(t_j) - \nu D(t_j) \right).$$

Note that $\kappa_{n,\alpha}^{ODW}$ is also $O(1)$.

Now we turn to confidence bands for bi-s^*-distribution functions. Our approach will be to refine the three unconstrained bands given in the three examples.

Suppose F is a bi-s^*-concave distribution function. A nonparametric $(1-\alpha)$ confidence band (L_n, U_n) for

F may be refined as follows:

$$\begin{aligned} L_n^o(x) &:= \inf\{G(x) : G \in \mathcal{P}_{s^*}, L_n \le G \le U_n\}, \\ U_n^o(x) &:= \sup\{G(x) : G \in \mathcal{P}_{s^*}, L_n \le G \le U_n\}. \end{aligned}$$

If there is no bi-s^*-concave distribution function F fitting into the band (L_n, U_n), we set $L_n^o := 1$ and $U_n^o := 0$ and we conclude with confidence $1-\alpha$ that F is not bi-s^*-concave. But in the case of $F \in \mathcal{P}_{s^*}$, this happens with probability at most α.

The following lemma implies two properties of our shape-constrained band (L_n^o, U_n^o). The first one is that both L_n^o and U_n^o are Lipschitz continuous on $\mathbb{R}$, unless $\inf\{x \in \mathbb{R} : L_n(x) > 0\} \ge \sup\{x \in \mathbb{R} : U_n(x) < 1\}$. The second one is that $L_n^o(x)$ converges polynomially fast to 0 as $x \to -\infty$ and $U_n^o(x)$ converges polynomially fast to 1 as $x \to \infty$ as long as $\lim_{x\to\infty} L_n(x) > \lim_{x\to-\infty} U_n(x)$.

Lemma 2.1. *For real numbers $a < b$, $0 < u < v < 1$ and $s^* \in (-\infty, 0) \cup (0, 1]$, define*

$$\gamma_1 := \frac{\frac{1}{s^*}\left(v^{s^*} - u^{s^*}\right)}{b-a} \text{ and } \gamma_2 := \frac{\frac{-1}{s^*}\left((1-v)^{s^*} - (1-u)^{s^*}\right)}{b-a}.$$

(i) If $L_n(a) \ge u$ and $U_n(b) \le v$, then L_n^o and U_n^o are Lipschitz-continuous on $\mathbb{R}$ with Lipschitz constant $\max\{\gamma_1, \gamma_2\}$.
(ii) If $U_n(a) \le u$ and $L_n(b) \ge v$, then

$$U_n^o(x) \le \left(u^{s^*} + s^*\gamma_1(x-a)\right)_+^{1/s^*} \text{ for } x \le a,$$

$$1 - L_n^o(x) \le \left((1-v)^{s^*} - s^*\gamma_2(x-b)\right)_+^{1/s^*} \text{ for } x \ge b.$$

The following theorem implies the consistency of our proposed confidence band (L_n^o, U_n^o).

Theorem 2.2. *Suppose that the original confidence band (L_n, U_n) is consistent in the sense that for any fixed $x \in \mathbb{R}$, both $L_n(x)$ and $U_n(x)$ tend to $F(x)$ in probability.*
(i) Suppose that $F \notin \mathcal{P}_{s^*}$. Then $P(L_n^o \le U_n^o) \to 0$.
(ii) Suppose that $F \in \mathcal{P}_{s^*}$ with $s^* \neq 0$. Then $P(L_n^o \le U_n^o) \ge 1-\alpha$, and

$$\sup_{G\in\mathcal{P}_{s^*}: L_n \le G \le U_n} \|G - F\|_\infty \to_p 0, \tag{2.11}$$

where $\sup(\emptyset) := 0$. Moreover, for any compact interval $K \subset J(F)$,

$$\sup_{G\in\mathcal{P}_{s^*}: L_n \le G \le U_n} \|h_G - h_F\|_{K,\infty} \to_p 0, \tag{2.12}$$

where h_G stands for any of the three functions G', $\left(G^{s^*}\right)'$, and $\left((1-G)^{s^*}\right)'$. Finally, for any fixed $x_1 \in J(F)$

and $0 < b_1 < f(x_1)/F^{1-s^*}(x_1)$,

$$P\left(U_n^o(x) \le \left(U_n^{s^*}(x') + s^* b_1 (x - x')\right)_+^{1/s^*} \text{ for } x \le x' \le x_1\right) \to 1, \tag{2.13}$$

while for any fixed $x_2 \in J(F)$ *and* $0 < b_2 < f(x_2)/(1 - F(x_2))^{1-s^*}$,

$$P\left(1 - L_n^o(x) \le \left((1 - L_n(x'))^{s^*} - s^* b_2 (x - x')\right)_+^{1/s^*} \text{ for } x \ge x' \ge x_2\right) \to 1. \tag{2.14}$$

The following result provides the consistency of confidence bands for functionals $\int \phi dF$ of F with well-behaved integrands $\phi : \mathbb{R} \to \mathbb{R}$.

Corollary 2.3. *Suppose that the original confidence band* (L_n, U_n) *is consistent, and let* $F \in \mathcal{P}_{s^*}$ *with* $s^* < 0$. *Let* $\phi : \mathbb{R} \mapsto \mathbb{R}$ *be absolutely continuous with a continuous derivative* ϕ' *satisfying the following constraint: there exist constants* $a > 0$ *and* $k < -1/s^*$ *such that*

$$|\phi'(x)| \le a|x|^{k-1}.$$

Then

$$\sup_{G: L_n^o \le G \le U_n^o} \left| \int \phi dG - \int \phi dF \right| \to_p 0.$$

The following theorem provides rates of convergence, with the following condition on the original confidence band (L_n, U_n):

Condition (*): For certain constants $\gamma \in [0, 1/2)$ and $\kappa, \lambda > 0$,

$$\max\{\mathbb{F}_n - L_n, U_n - \mathbb{F}_n\} \le \kappa n^{-1/2} \left(\mathbb{F}_n(1 - \mathbb{F}_n)\right)^\gamma$$

on the interval $\{\lambda n^{-1/(2-2\gamma)} \le \mathbb{F}_n \le 1 - \lambda n^{-1/(2-2\gamma)}\}$.

As stated in Dümbgen et al. (2017), this condition is satisfied with $\gamma = 0$ in the case of the Kolmogorov-Smirnov band. In the case of the weighted Kolmogorov-Smirnov band, it is satisfied for the given value of $\gamma \in [0, 1/2)$. For the refined version of Owen's band, it is satisfied for any fixed number $\gamma \in (0, 1/2)$.

Theorem 2.3. *Suppose that* $F \in \mathcal{P}_{s^*}$ *with* $s^* < 0$ *and let* (L_n, U_n) *satisfy Condition (*). Let* $\phi : \mathbb{R} \mapsto \mathbb{R}$ *be absolutely continuous with a continuous derivative* ϕ'.

Suppose that $|\phi'(x)| = O(|x|^{k-1})$ *as* $|x| \to \infty$ *for some numbers* $k < -1/s^*$. *Then*

$$\sup_{G: L_n^o \le G \le U_n^o} \left| \int \phi dG - \int \phi dF \right| = O_p\left(n^{-\frac{1}{2}\left(1 \wedge \frac{ks^*+1}{1-\gamma}\right)}\right). \tag{2.15}$$

Remark 2.3. (i) From (2.15), one can verify that the convergence rate is $n^{-1/2}$ as long as $k < \gamma/(-s^*)$.

(ii) From (2.15), one can verify that when $\gamma = 0$, the convergence rate is $n^{-1/2+k/(-s^*)}$ and we have a

"power deficit" (or "polynomial rate deficit") relative to $n^{-1/2}$.

2.4.2 Implementation and illustration of the confidence bands

In this chapter, we discuss the implementation of confidence bands for bi-s^*-concave distribution functions. This extends the treatment of Dümbgen et al. (2017) from $s^* = 0$ to general values $s^* \in (-\infty, 1]$.

Recall the procedure ConcInt$(\cdot,\cdot)$ developed in Dümbgen et al. (2017). Given any finite set $\mathcal{T} = \{t_0, ..., t_m\}$ of real numbers $t_0 < t_1 < \cdots < t_m$ and any pair (l, u) of functions $l, u : \mathcal{T} \to [-\infty, \infty)$ with $l < u$ pointwise and $l(t) > -\infty$ for at least two different points $t \in \mathcal{T}$, this procedure computes the pair (l^o, u^o) where

$$\begin{aligned} l^o(x) &:= \inf\{g(x) : \ g \text{ is concave on } \mathbb{R},\ l \le g \le u \text{ on } \mathcal{T}\}, \\ u^o(x) &:= \sup\{g(x) : \ g \text{ is concave on } \mathbb{R},\ l \le g \le u \text{ on } \mathcal{T}\}. \end{aligned}$$

First note that l^o is the smallest concave majorant of l on $\mathcal{T}$; thus it may be computed by a version of the pool-adjacent-violators algorithm; see for example Robertson et al. (1988). Then we obtain indices $0 \le j(0) < j(1) < \cdots < j(b) \le m$ such that

$$l^o \begin{cases} := -\infty \text{ on } \mathbb{R}\backslash[t_{j(0)}, t_{j(b)}], \\ \text{is linear on } [t_{j(a-1)}, t_{j(a)}] \text{ for } 1 \le a \le b, \\ \text{change slope at } t_{j(a)} \text{ if } 1 \le a \le b. \end{cases}$$

With l^o in hand, we then check to see if $l^o \le u$ on $\mathcal{T}$. If this fails, then there is no concave function lying between l and u, and the procedure returns an error message. If this test succeeds, then we compute $u^o(x)$ as

$$\min\left\{u(s) + \frac{u(s) - l^o(r)}{s - r}(x - s) : r \in \mathcal{T}_o, r < s \le x \text{ or } x \le s < r\right\},$$

where $\mathcal{T}_o = \{t_{j(0)}, t_{j(1)}, \ldots, t_{j(b)}\}$. (The rest of the description of the procedure ConcInt$(\cdot,\cdot)$ is just as in Dümbgen et al. (2017).)

When $s^* < 0$, let $g(v; s^*) := g(v) := -v^{s^*}$ and $h(v; s^*) := h(v) := (-v)^{1/s^*}$. (This is the most important new case. When $s = s^* = 0$, $g(v) := \log(v), h(v) := \exp(v)$. When $s^* > 0$, $g(v) := v^{s^*}$ and $h(v) := v^{1/s^*}$.) Here is pseudocode for the computation of (L_n^o, U_n^o).

$$\begin{aligned} &(L_n^o, U_n^o) \leftarrow (L_n, U_n) \\ &(l^o, u^o) \leftarrow \text{ConcInt}(g(L_n^o), g(U_n^o)) \\ &(\tilde{L}_n^o, \tilde{U}_n^o) \leftarrow (h(l^o), h(u^o)) \end{aligned}$$

$(l^o, u^o) \leftarrow \text{ConcInt}(g(1-\tilde{U}_n^o), g(1-\tilde{L}_n^o))$

$(\tilde{L}_n^o, \tilde{U}_n^o) \leftarrow (1-h(u^o), 1-h(l^o))$

while $(\tilde{L}_n^o, \tilde{U}_n^o) \neq (L_n^o, U_n^o)$ do

$(L_n^o, U_n^o) \leftarrow (\tilde{L}_n^o, \tilde{U}_n^o)$

$(l^o, u^o) \leftarrow \text{ConcInt}(g(L_n^o), g(U_n^o))$

$(\tilde{L}_n^o, \tilde{U}_n^o) \leftarrow (h(l^o), h(u^o))$

$(l^o, u^o) \leftarrow \text{ConcInt}(g(1-\tilde{U}_n^o), g(1-\tilde{L}_n^o))$

$(\tilde{L}_n^o, \tilde{U}_n^o) \leftarrow (1-h(u^o), 1-h(l^o))$

end while.

Illustration of the confidence bands

To get some feeling for the new confidence bands in a setting in which s_0^* is known, we generated a sample of size $n = 100$ from the Student-t distribution with $r = 1$ degrees of freedom. This distribution belongs to $\mathcal{P}_{s^*}$ for every $s^* \leq -1 := s_0^*$. We constructed Kolmogorov-Smirnov (KS) and weighted Kolmogorov-Smirnov (WKS) bands with $\gamma = 0.4$ as the initial starting bands (L_n, U_n). We then computed and plotted our shape constrained confidence bands (L_n^0, U_n^0) under the (correct) assumption that $s^* = -1$ and the (incorrect) assumption that $s^* = 0$ for both the KS and WKS bands as initial nonparametric bands with for $\alpha = 0.05$; see Figure 3 and Figure 4. To see the components of Figures 1 and 2 separately, see the Supplementary file, Figures 1-2 and 3-4 respectively.

Note that when $s^* = 0$, s^* is miss-specified and the resulting bands are not guaranteed to have coverage probabiltiy $.95$. An indication of this is that the shape constrained bands computed under the assumption $s^* = 0$ do not contain the empirical distribution.

From these two plots, an immediate observation is that the confidence bands for smaller s^* are wider than those with larger s^*. This is a direct consequence of the nested property of the bi-s^*-concave classes; see Proposition 2.2. Also note that the shape constrained band with $s^* = -1$ does improve on the KS band, especially in the tail.

An Application

Dümbgen et al. (2017) gave an application of bi-log-concave confidence bands to a dataset from Wooldridge (2000). It contains approximate annual salaries of the CEOs of 177 randomly chosen companies in the U.S. The salary is rounded to multiples of 1000 USD. We denote the i-th observed approximate salary by $Y_{i,raw}$.

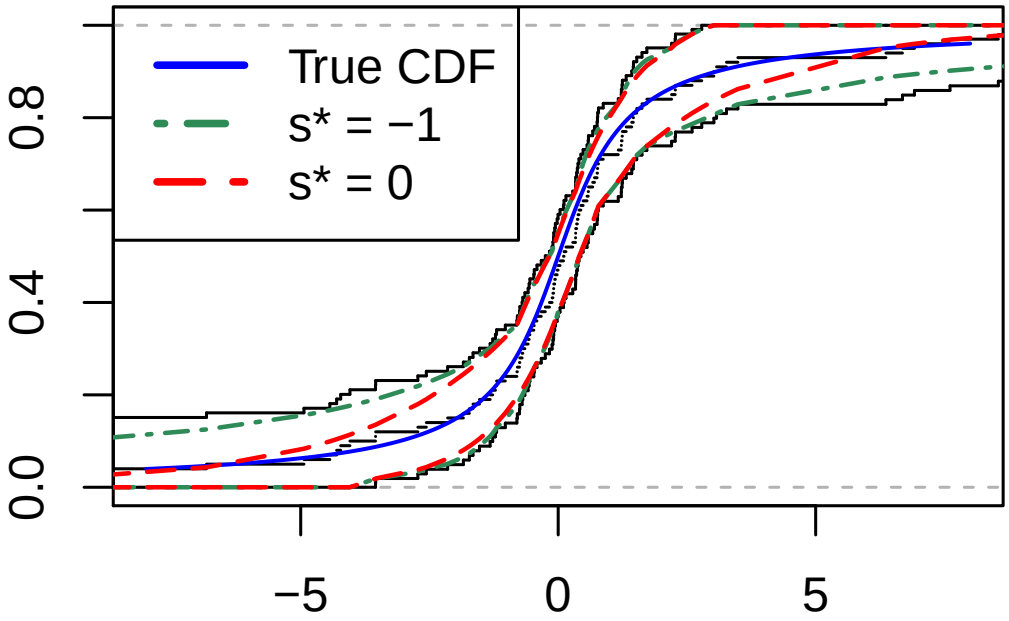

Figure 3: Confidence bands for bi-s^*-concave distribution functions based on KS bands. The black curve is the distribution function of the Student-t distribution with 1 degree of freedom. The two gray-black lines give the KS band and lines in other colors are refined confidence bands under the bi-s^*-concave assumption. The step function in the middle is the empirical distribution function.

Dümbgen et al. (2017) assume that the unobserved true salary $Y_{i,true}$ lies within $(Y_{i,raw} - 1, Y_{i,raw} + 1)$. Let us assume that G_{true} is the unknown distribution of Y_{true}. For income data it is sometimes assumed that $\log_{10} Y_{true}$ is Gaussian (see Kleiber and Kotz (2003)). Since Gaussian densities are all log-concave and hence have bi-log-concave distribution functions (by Proposition 2.1), it is natural to consider replacing the Gaussian assumption by the assumption of bi-log-concavity. Dümbgen et al. (2017) therefore assumed that $X = \log_{10} Y_{true}$ is bi-log-concave and constructed 95% confidence bands (L_n, U_n) (see Figure 4 of Dümbgen et al. (2017)) where L_n is computed with the empirical distribution of $\log_{10}(Y_{i,raw} - 1)_{i=1}^n$ and U_n is computed with that of $\log_{10}(Y_{i,raw} + 1)_{i=1}^n$.

Here we assume that the distribution of X is bi-s^*-concave for some s^* and compute confidence bands

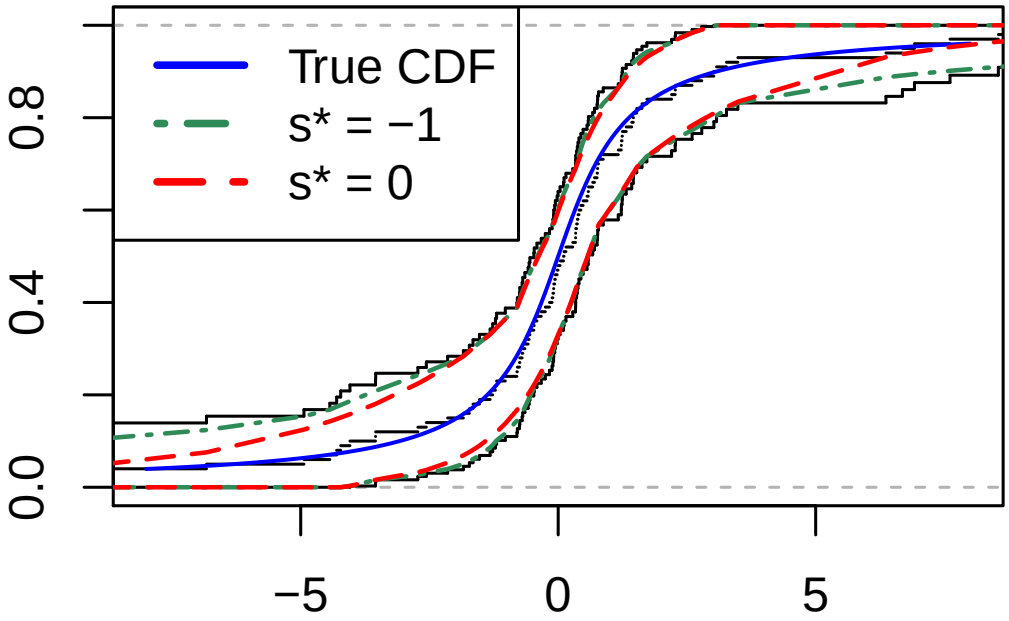

Figure 4: Confidence bands for bi-s^*-concave distribution functions based on WKS bands. The black curve is the distribution function of the Student-t distribution with 1 degree of freedom. The two gray-black lines give the WKS band and lines in other colors are refined confidence bands under the bi-s^*-concave assumption. The step function in the middle is the empirical distribution function.

for different values of s^*. Now we are confronted with the issue of choosing s^*: if we want narrower confidence bands we would assume some value of $s^* \in (0, 1]$, while if we are not willing to assume $s^* = 0$ (the choice made by Dümbgen et al. (2017), then we would assume some value of $s^* < 0$ (leading to the larger classes $\mathcal{P}_{s^*}$ with $s^* < 0$. It is of some interest to know if the CEO data could be modeled by use of the bi-s^* classes with $s^* \in (0, 1]$ since this would result in still narrower confidence bands. But it is also of interest to try to use the data to choose s^*.

Choosing s^*

Since F can be a member of $\mathcal{P}_{s^*}$ for various values of s^*, each s^* leads to a different set of bands. However, due to the nesting property of $\mathcal{P}_{s^*}$, a larger s^* always yields a narrower confidence band. Thus, it

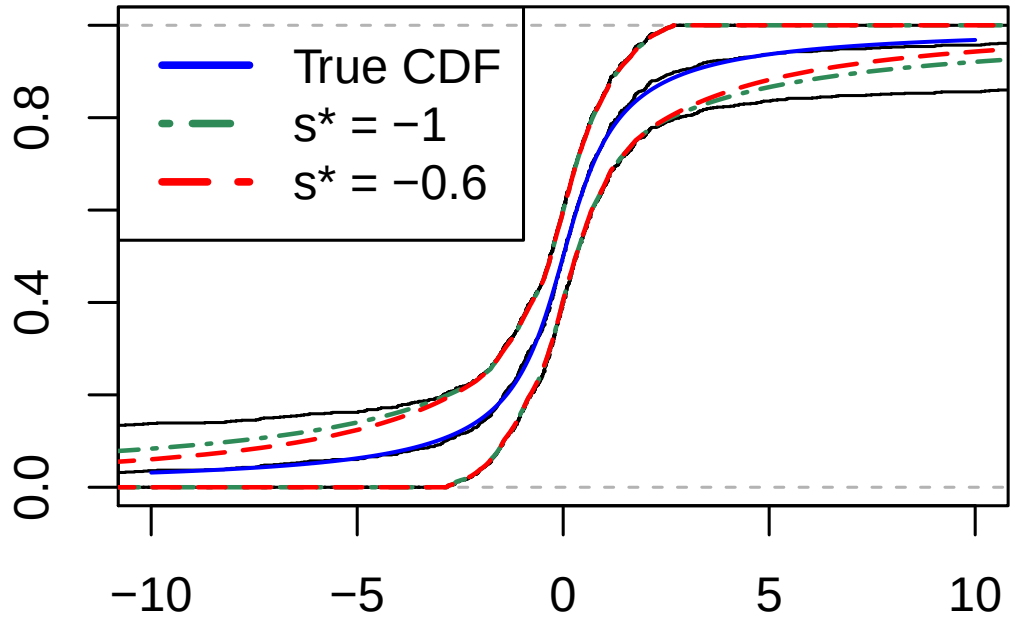

Figure 5: Confidence Bands for bi-s^*-concave distribution functions from KS bands based on a sample of size 1000 from the Student-t distribution with 1 degree of freedom. The two gray-black lines give the initial bands, lines in other colors are refined confidence bands under the bi-s^*-concave assumption. The step function (black) in the middle is the empirical distribution function.

is of interest to estimate

$$s_0^*(F) := \sup\{s^* \in (-\infty, 1] : F \in \mathcal{P}_{s^*}\}$$

since $s^* = s_0^*$ generates the narrowest bands at a given confidence level. If F is not bi-s^*-concave for any $s^* \leq 1$, then we set $s_0^*(F) = -\infty$. Now s_0^* is connected to the Csörgő - Révész constant since $s^* = s_0^*$ when $\overline{\gamma}(F) = 1 - s^*$ and $F \in \mathcal{P}_s^*$. For example, the Student-t distribution with r "degree of freedom" has $s_0^* = -1/r$. However, this connection cannot be easily exploited for practical estimation purposes due to difficulties in estimating $\gamma(F)$ or $\overline{\gamma}(F)$. So we take an alternative route to making inference about s_0^*.

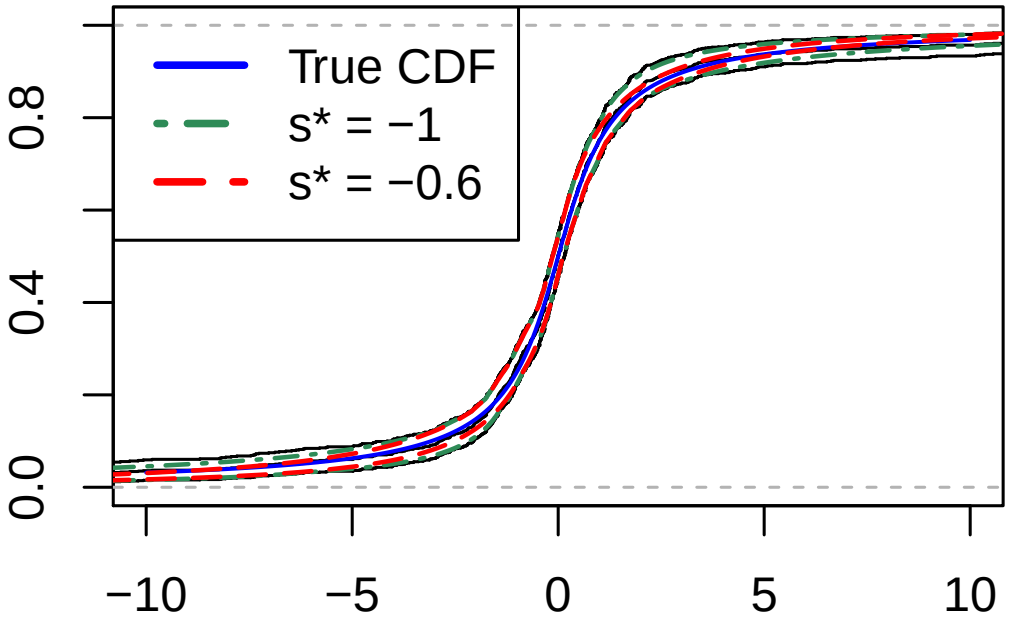

Figure 6: Confidence Bands for bi-s^*-concave distribution functions from WKS bands based on a sample of size 1000 from the Student-t distribution with one degree of freedom. The two gray-black lines give the initial bands, lines in other colors are refined confidence bands under the bi-s^*-concave assumption. The step function (black) in the middle is the empirical distribution function.

Starting from an initial $1-\alpha$ band (L_n, U_n), a bound on s_0^* is given by

$$\overline{s}_n^* = \sup\{s^* \in (-\infty, 1] : \ (L_n, U_n) \text{ contains some d.f. } F \in \mathcal{P}_{s^*}\}.$$

Clearly, for $s^* > \overline{s}_n^*$, there is no bi-s^*-concave distribution function fitting into the band (L_n, U_n). Since this happens with probability at most $\alpha \in (0,1)$ when the true distribution function $F \in \mathcal{P}_{s^*}$, it follows that $(-\infty, \overline{s}_n^*]$ is a confidence set for s_0^* with coverage probability at least $1-\alpha$. Our simulations suggest that $\overline{s}_n^*$ is generally considerably larger than s_0^*, and hence not suitable as an estimator, especially for $\alpha = 0.05$.

Instead, we propose an estimator of s_0^* based on the $\mathbb{F}_n$ measure of the set where the empirical measure remains between the shape-constrained band for s^*. More formally, let $L_n^o(s^*)$ and $U_n^o(s^*)$ denote the $1-\alpha$

level bi-s^*-concave confidence bands based on the initial bands L_n and U_n and the assumption $F \in \mathcal{P}_{s^*}$. Define

$$\begin{aligned}\omega(s^*) &:= n^{-1}\sum_{i=1}^{n} 1\{L_n^o(s^*)(X_i) \le \mathbb{F}_n(X_i) \le U_n^o(s^*)(X_i)\} \\ &\qquad \cdot 1\{L_n^o(s^*)(X_i) \le U_n^o(s^*)(X_i)\} \\ &= \mathbb{F}_n\left(\{L_n^o(s^*) \le \mathbb{F}_n \le U_n^o(s^*)\} \cap \{L_n^o(s^*) \le U_n^o(s^*)\}\right).\end{aligned}$$

A higher value of $\omega(s^*)$ indicates that $(L_n^o(s^*), U_n^o(s^*))$ contains a greater portion of $\mathbb{F}_n$. Since the bands $(L_n(s^*), U_n(s^*))$ become narrower as s^* increases, $\omega(s^*)$ decreases in s^*, and eventually becomes zero when $s^* > \overline{s}_n^*$. A plausible estimator of s_0^* is therefore given by

$$\widehat{s}_n^* = \min\{s^* \in (-\infty, \overline{s}_n^*] : \omega(s^*) > \rho\}, \tag{2.16}$$

where ρ is a threshold taking values in $(0, 1)$. The calculation of $\widehat{s}_n^*$ thus depends on α and ρ.

In the case of the CEO data, $\overline{s}_n^* \approx 0.23$ for the KS initial band, and $\overline{s}_n^* \approx 0.18$ for the WKS band. Taking $\alpha = 0.05$ and $\rho = .95$, leads to $\widehat{s}_n^* = 0.12$, while taking $\alpha = 0.05$ and $\rho = 0.95$, leads to $\widehat{s}_n^* = .12$. The resulting bands are given in Figures 7 and 8. Also see the Supplementrary file, Figures 9 - 10 and Figures 11-12 for the steps in constructing Figures 7 and 8.

We should emphasize that our current theory says little about the coverage probabilities of the bands $(L_n^o(s^*), U_n^o(s^*))$. Discussion of the consistency of $\widehat{s}_n^*$ is beyond the scope of the present chapter, but this and further issues concerning inference for both s^* and $F \in \mathcal{P}_{s^*}$ seem to be interesting directions for future research.

2.5 Summary and further problems

In this chapter we have:

• Defined new classes of shape-constrained distribution functions, the bi-s^*-classes extending the bi-log-concave class of distribution functions defined by Dümbgen et al. (2017).

• Characterized the new classes and connected our characterization to an important parameter, the Csörgő - Révész constant associated with a distribution function F.

• Used the new bi-s^*-concave classes to define refined confidence bands for distribution functions which exploit the shape constraint, thereby producing more accurate (narrower) bands with honest coverage when the shape constraint holds.

Thus we have shown that if we know the parameter $s^* \in (-\infty, 1]$ determining the class, we can construct refined confidence bands which improve on any given nonparametric confidence bands if the given value of

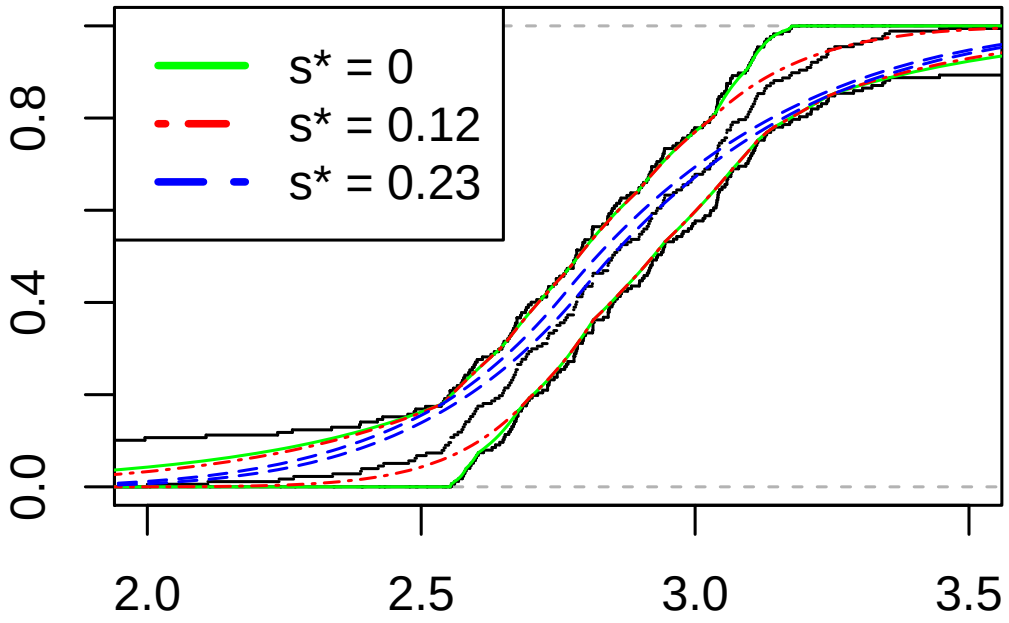

Figure 7: Confidence Bands from an initial KS band for the CEO salary data. The step function in the middle is the empirical distribution function. The two gray-black lines give the KS band and lines in other colors are refined confidence bands under the bi-s^*-concave assumption.

s^* is correct. It follows from the construction of our bands that they have conservative coverage probabilities under the (null) hypothesis that the true distribution function is in $\mathcal{P}_{s^*}$ and that s^* is correctly specified.

• What if we do not know s^*? Can we estimate it from the data? As becomes clear from the discussion of the CEO data via Figures 5 and 6, our methods provide one-sided confidence bounds for the true s^* of the form $(-\infty, \overline{s}_n^*]$ under the assumption that $F \in \mathcal{P}_{s^*}$ for some s^*. It remains to develop inference methods for s^* and (s^*, F) jointly. It will also be of interest to have a more complete understanding of the power behavior of tests related to $\overline{s}_n^*$ and $\widehat{s}_n^*$.

• The stable laws are known to be unimodal; see e.g. Hall (1984) for some history. In connection with Example 2.8 we have the following:

Conjecture: the $\alpha-$stable laws are $s-$concave with $s = -1/(1+\alpha)$ for $0 < \alpha < 2$.

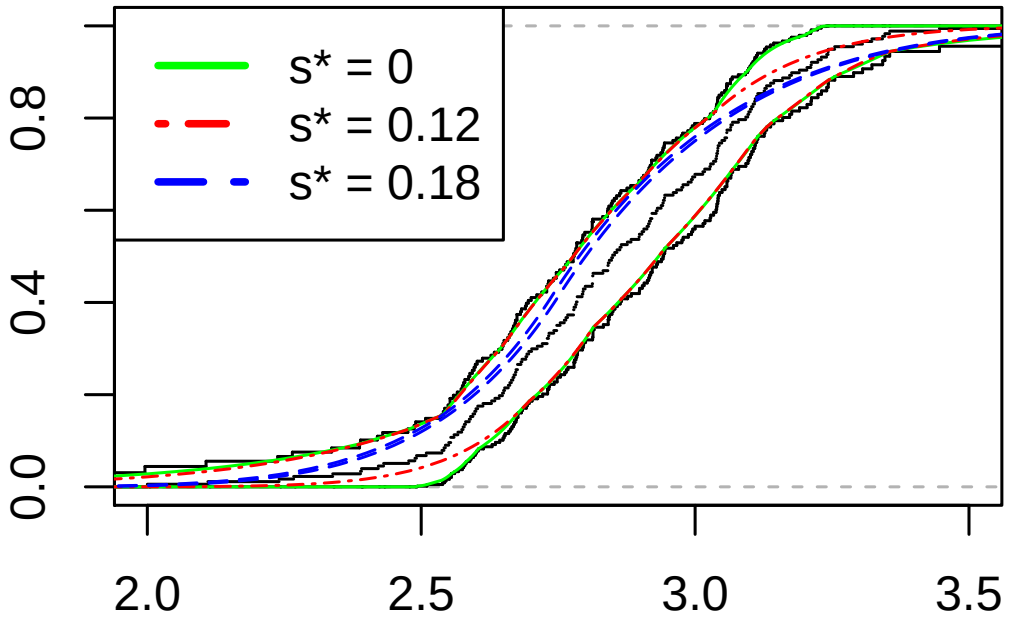

Figure 8: Confidence Bands from an initial WKS band for the CEO salary data. The step function in the middle is the empirical distribution function. The two gray-black lines give the WKS band and lines in other colors are refined confidence bands under the bi-s^*-concave assumption.

2.6 Proofs

Proof of Theorem 2.1. Throughout our proof we will denote $\inf J(F)$ and $\sup J(F)$ by a and b respectively. Note that the case $s^* = 0$ is proved by Dümbgen et al. (2017).

Part (a): $s^* < 0$**.**

(i) implies (ii):

Suppose $F \in \mathcal{P}_{s^*}$. To prove that F is continuous on $\mathbb{R}$, we first note that $x \mapsto F^{s^*}(x)$ and $x \mapsto (1 - F(x))^{s^*}(x)$ are convex functions on $\mathbb{R}$. By theorem 10.1 (page 82) of Rockafellar (1970), F^{s^*} and $(1 - F(x))^{s^*}$ are continuous on any open convex sets in their effective domains. In particular, F^{s^*} and $(1 - F)^{s^*}$ are continuous on (a, ∞) and $(-\infty, b)$ respectively. This implies that F is continuous on (a, ∞) and $(-\infty, b)$, or equivalently,

on $(a,\infty)\cup(-\infty,b)=(-\infty,\infty)$ since F is non-degenerate.

To prove that F is differentiable on $J(F)$, note that $J(F)=(a,b)$ since F is continuous on $\mathbb{R}$. By theorem 23.1 (page 213) of Rockafellar (1970), for any $x\in J(F)$, the convexity of F^{s^*} on $J(F)$ implies the existence of $\left(F^{s^*}\right)'_+(x)$ and $\left(F^{s^*}\right)'_-(x)$. Moreover, $\left(F^{s^*}\right)'_-(x)\le\left(F^{s^*}\right)'_+(x)$ by theorem 24.1 (page 227) in Rockafellar (1970). Since $F=\left(F^{s^*}\right)^{1/s^*}$ on $J(F)$, the chain rule guarantees the existence of $F'_\pm(x)$ and

$$F'_\pm(x)=\frac{1}{s^*}\left(F^{s^*}\right)^{1/s^*-1}(x\pm)\left(F^{s^*}\right)'_\pm(x).$$

Since F is continuous on $J(F)$, then

$$F'_\pm(x)=\frac{1}{s^*}\left(F^{s^*}\right)^{1/s^*-1}(x)\left(F^{s^*}\right)'_\pm(x).$$

Hence $F'_-(x)\ge F'_+(x)$ by noting that $\left(F^{s^*}\right)'_-(x)\le\left(F^{s^*}\right)'_+(x)$ and $s^*<0$.

Similarly, one can prove $F'_-(x)\le F'_+(x)$ by the convexity of $(1-F)^{s^*}$ on $J(F)$.

Thus $F'_-(x)=F'_+(x)=F'(x)$ for any $x\in J(F)$, or equivalently, F is differentiable on $J(F)$. The derivative of F is denoted by f, i.e. $f:=F'$.

To prove (2.6), note that the convexity of $x\mapsto F^{s^*}(x)$ on $J(F)$ implies that, for any $x,y\in J(F)$,

$$F^{s^*}(y)-F^{s^*}(x)\ge(y-x)\left(F^{s^*}\right)'(x)=(y-x)s^*F^{s^*-1}(x)f(x),$$

or, with $x_+=\max\{x,0\}$,

$$\frac{F^{s^*}(y)}{F^{s^*}(x)}\ge\left(1+s^*\frac{f(x)}{F(x)}(y-x)\right)_+.$$

Hence,

$$\frac{F(y)}{F(x)}\le\left(1+s^*\frac{f(x)}{F(x)}(y-x)\right)_+^{1/s^*},$$

or, equivalently,

$$F(y)\le F(x)\left(1+s^*\frac{f(x)}{F(x)}(y-x)\right)_+^{1/s^*}.$$

Analogously, the convexity of $(1-F(x))^{s^*}$ on $J(F)$ implies that

$$(1-F(y))^{s^*}-(1-F(x))^{s^*}\ge-(y-x)s^*(1-F(x))^{s^*-1}f(x),$$

or, equivalently,

$$\left(\frac{1-F(y)}{1-F(x)}\right)^{s^*}\ge\left(1-s^*\frac{f(x)}{1-F(x)}(y-x)\right)_+,$$

which yields

$$F(y)\ge1-(1-F(x))\left(1-s^*\frac{f(x)}{1-F(x)}(y-x)\right)_+^{1/s^*}.$$

The proof of (2.6) is complete.

(ii) implies (iii):

Applying (2.6) yields that for any $x, y \in J(F)$ with $x < y$,

$$\frac{F^{s^*}(x)}{F^{s^*}(y)} \geq 1 + s^* \frac{f(y)}{F(y)}(x - y),$$

and

$$\frac{F^{s^*}(y)}{F^{s^*}(x)} \geq 1 + s^* \frac{f(x)}{F(x)}(y - x),$$

or, equivalently,

$$F^{s^*}(x) \geq F^{s^*}(y) + s^* \frac{f(y)}{F^{1-s^*}(y)}(x - y),$$

and

$$F^{s^*}(y) \geq F^{s^*}(x) + s^* \frac{f(x)}{F^{1-s^*}(x)}(y - x).$$

By defining $h := f/F^{1-s^*}$ on $J(F)$, it follows that

$$F^{s^*}(x) \geq F^{s^*}(y) + s^* h(y)(x - y),$$

and

$$F^{s^*}(y) \geq F^{s^*}(x) + s^* h(x)(y - x).$$

After summing up the last two inequalities, it follows that

$$F^{s^*}(x) + F^{s^*}(y) \geq F^{s^*}(y) + s^* h(y)(x - y) + F^{s^*}(x) + s^* h(x)(y - x),$$

or, equivalently,

$$0 \geq s^* \left(h(x) - h(y)\right)(y - x).$$

Hence $h(x) \geq h(y)$, or equivalently, $h(\cdot)$ is a monotonically non-increasing function on $J(F)$.

The proof of the monotonicity of $\tilde{h} := f/(1 - F)^{1-s^*}$ is similar and hence is omitted.

(iii) implies (iv):

If (iii) holds, it immediately follows that $f > 0$ on $J(F) = (a, b)$. If not, suppose that $f(x_0) = 0$ for some $x_0 \in J(F)$. It follows that $h(x_0) = f(x_0)/F^{1-s^*}(x_0) = 0$. Since h is monotonically non-increasing on $J(F)$, $h(x) = 0$ for all $x \in [x_0, b)$, or, equivalently, $f = 0$ on $[x_0, b)$. Similarly, the non-decreasing monotonicity of $x \mapsto \tilde{h}(x)$ on $J(F)$ implies that $f = 0$ on $(a, x_0]$. Then $f = 0$ on $J(F)$, which violates the continuity assumption in (iii) and hence $f > 0$ on $J(F)$.

To prove f is bounded on $J(F)$, note that the monotonicities of h and $\tilde{h}$ imply that for any $x, x_0 \in J(F)$,

$$f(x) = \begin{cases} F^{1-s^*}(x)h(x) \leq h(x) \leq h(x_0), & \text{if } x \geq x_0, \\ (1-F(x))^{1-s^*}\tilde{h}(x) \leq \tilde{h}(x) \leq \tilde{h}(x_0), & \text{if } x \leq x_0. \end{cases}$$

Hence $f(x) \leq \max\{h(x_0), \tilde{h}(x_0)\}$ for any $x, x_0 \in J(F)$.

To prove that f is differentiable on $J(F)$ almost everywhere, we first prove that f is Lipschitz continuous on (c, d) for any $c, d \in J(F)$ with $c < d$.

By the non-increasing monotonicity of h on $J(F)$, the following arguments yield an upper bound of $(f(y) - f(x))/(y - x)$ for any $x, y \in (c, d)$:

$$\begin{aligned} \frac{f(y)-f(x)}{y-x} &= \frac{F^{1-s^*}(y)h(y) - F^{1-s^*}(x)h(x)}{y-x} \\ &= h(y)\frac{F^{1-s^*}(y) - F^{1-s^*}(x)}{y-x} + F^{1-s^*}(x)\frac{h(y)-h(x)}{y-x} \\ &\leq h(y)\frac{F^{1-s^*}(y) - F^{1-s^*}(x)}{y-x} \\ &= h(y)(1-s^*)f(z)F^{-s^*}(z), \end{aligned}$$

where the last equality follows from the mean value theorem and z is between x and y.

Since $-s^* > 0$, it follows that $F^{-s^*} < 1$ and hence

$$\frac{f(y)-f(x)}{y-x} < (1-s^*)f(z)h(y) \leq (1-s^*)\max\{h(x_0), \tilde{h}(x_0)\}h(c)$$

for $x, y \in (c, d)$.

Similar arguments imply that

$$\begin{aligned} \frac{f(y)-f(x)}{y-x} &= \frac{\bar{F}^{1-s^*}(y)\tilde{h}(y) - \bar{F}^{1-s^*}(x)\tilde{h}(x)}{y-x} \\ &= \tilde{h}(y)\frac{\bar{F}^{1-s^*}(y) - \bar{F}^{1-s^*}(x)}{y-x} + \bar{F}^{1-s^*}(x)\frac{\tilde{h}(y)-\tilde{h}(x)}{y-x} \\ &\geq \tilde{h}(y)\frac{\bar{F}^{1-s^*}(y) - \bar{F}^{1-s^*}(x)}{y-x} \\ &= -\tilde{h}(y)(1-s^*)\bar{F}^{-s^*}(z)f(z) \\ &\geq -(1-s^*)\max\{h(x_0), \tilde{h}(x_0)\}\tilde{h}(d). \end{aligned}$$

Hence

$$\left|\frac{f(y)-f(x)}{y-x}\right| \leq (1-s^*)\max\{h(x_0), \tilde{h}(x_0)\}\max\{h(c), \tilde{h}(d)\}.$$

The last display shows that f is Lipschitz continuous on (c, d).

By proposition 4.1(iii) of Shorack (2017), page 82, f is absolutely continuous on (c, d), and hence f is

differentiable on (c, d) almost everywhere.

Since (c, d) is an arbitrary interval in (a, b), the differentiability of f on (c, d) implies the differentiability of f on (a, b) and hence f is differentiable on (a, b) with $f' = F''$ almost everywhere.

Since f is differentiable almost everywhere, the non-increasing monotonicity of h on $J(F)$ implies that

$$h'(x) \leq 0 \text{ almost everywhere on } J(F),$$

or, equivalently,

$$\log(h)'(x) \leq 0 \text{ almost everywhere on } J(F).$$

Straight-forward calculation yields that the last display is equivalent to

$$\frac{f'}{f} - (1 - s^*)\frac{f}{F} \leq 0 \text{ almost everywhere on } J(F),$$

or,

$$f' \leq (1 - s^*)\frac{f^2}{F} \text{ almost everywhere on } J(F),$$

which is the right hand side of (2.8).

Similarly, the non-decreasing monotonicity of $\tilde{h}$ implies the left hand side of (2.8).

(iv) implies (i):

Since F is continuous on $\mathbb{R}$, it suffices to prove that F^{s^*} is convex on $J(F)$ by Definition 2.1. Since we assume that F is differentiable on $J(F)$ with derivative $f = F'$, the convexity of F^{s^*} on $J(F)$ can be proved by the increasing monotonicity of the first derivative of F^{s^*} on $J(F)$. Since f is differentiable almost everywhere on $J(F)$, the increasing monotonicity of $\left(F^{s^*}\right)'$ on $J(F)$ can be proved by the non-negativity of $\left(F^{s^*}\right)''$ on $J(F)$ almost everywhere, which follows from

$$\left(F^{s^*}\right)''(x) = s^* F^{s^*-1}(x)\left(-(1 - s^*)\frac{f^2(x)}{F(x)} + f'(x)\right) \geq 0,$$

where $f = F', f' = F''$. The last inequality follows from the right hand side of (2.8).

Similarly, the convexity of $(1 - F(x))^{s^*}$, or $\bar{F}^{s^*}$, on $J(F)$ can be proved by the following arguments:

$$\left(\bar{F}^{s^*}\right)''(x) = s^* \bar{F}^{s^*-1}(x)\left(-(1 - s^*)\frac{f^2(x)}{\bar{F}(x)} - f'(x)\right) \geq 0,$$

where the last inequality follows from the left part of (2.8).

Part (b): $0 < s^* \leq 1$**.**

(i) implies (ii):

Suppose $F \in \mathcal{P}_{s^*}$. To prove that F is continuous on $\mathbb{R}$, we first note that $x \mapsto F^{s^*}(x)$ and $x \mapsto (1 - F(x))^{s^*}(x)$ are concave functions on (a, ∞) and $(-\infty, b)$ respectively. By theorem 10.1 (page 82) in Rockafellar (1970),

F^{s^*} and $(1-F(x))^{s^*}$ are continuous on any open convex sets in their effective domains. In particular, F^{s^*} and $(1-F)^{s^*}$ are continuous on (a,∞) and $(-\infty,b)$ respectively. This yields that F is continuous on (a,∞) and $(-\infty,b)$, or equivalently, on $(a,\infty)\cup(-\infty,b)=(-\infty,\infty)$ since F is non-degenerate.

To prove that F is differentiable on $J(F)$, note that $J(F)=(a,b)$ since F is continuous on $\mathbb{R}$. By theorem 23.1 (page 213) in Rockafellar (1970), for any $x\in J(F)$, the concavity of F^{s^*} on $J(F)$ implies the existence of $\left(F^{s^*}\right)'_+(x)$ and $\left(F^{s^*}\right)'_-(x)$. Moreover, $\left(F^{s^*}\right)'_-(x)\geq\left(F^{s^*}\right)'_+(x)$ by theorem 24.1 (page 227) in Rockafellar (1970). Since $F=\left(F^{s^*}\right)^{1/s^*}$ on $J(F)$, the chain rule guarantees the existence of $F'_\pm(x)$ and

$$F'_\pm(x)=\frac{1}{s^*}\left(F^{s^*}\right)^{1/s^*-1}(x\pm)\left(F^{s^*}\right)'_\pm(x).$$

Since F is continuous on $J(F)$, then

$$F'_\pm(x)=\frac{1}{s^*}\left(F^{s^*}\right)^{1/s^*-1}(x)\left(F^{s^*}\right)'_\pm(x).$$

Hence $F'_-(x)\geq F'_+(x)$ by $\left(F^{s^*}\right)'_-(x)\geq\left(F^{s^*}\right)'_+(x)$.

Similarly, one can prove $F'_-(x)\leq F'_+(x)$ by the concavity of $(1-F)^{s^*}$ on $J(F)$.

Thus $F'_-(x)=F'_+(x)=F'(x)$ for any $x\in J(F)$, or equivalently, F is differentiable on $J(F)$. The derivative of F is denoted by f, i.e. $f:=F'$.

To prove (2.6), note that the concavity of $x\mapsto F^{s^*}(x)$ on $J(F)$ implies that, for any $x,y\in J(F)$,

$$F^{s^*}(y)-F^{s^*}(x)\leq(y-x)\left(F^{s^*}\right)'(x)=(y-x)s^*F^{s^*-1}(x)f(x),$$

or, with $x_+=\max\{x,0\}$,

$$\frac{F^{s^*}(y)}{F^{s^*}(x)}\leq\left(1+s^*\frac{f(x)}{F(x)}(y-x)\right)_+.$$

Hence

$$\frac{F(y)}{F(x)}\leq\left(1+s^*\frac{f(x)}{F(x)}(y-x)\right)_+^{1/s^*},$$

or, equivalently,

$$F(y)\leq F(x)\left(1+s^*\frac{f(x)}{F(x)}(y-x)\right)_+^{1/s^*}.$$

Analogously, the convexity of $(1-F(x))^{s^*}$ on $J(F)$ implies that for any $x,y\in J(F)$

$$(1-F(y))^{s^*}-(1-F(x))^{s^*}\leq-(y-x)s^*(1-F(x))^{s^*-1}f(x),$$

or, equivalently,

$$\left(\frac{1-F(y)}{1-F(x)}\right)^{s^*}\leq\left(1-s^*\frac{f(x)}{1-F(x)}(y-x)\right)_+,$$

which yields

$$F(y) \geq 1 - (1 - F(x)) \left(1 - s^* \frac{f(x)}{1 - F(x)}(y - x)\right)_+^{1/s^*}.$$

The proof of (2.6) is complete.

(ii) implies (iii):

Applying (2.6) yields that for any $x, y \in J(F)$ with $x < y$,

$$\frac{F^{s^*}(x)}{F^{s^*}(y)} \leq 1 + s^* \frac{f(y)}{F(y)}(x - y),$$

and

$$\frac{F^{s^*}(y)}{F^{s^*}(x)} \leq 1 + s^* \frac{f(x)}{F(x)}(y - x),$$

or, equivalently,

$$F^{s^*}(x) \leq F^{s^*}(y) + s^* \frac{f(y)}{F^{1-s^*}(y)}(x - y),$$

and

$$F^{s^*}(y) \leq F^{s^*}(x) + s^* \frac{f(x)}{F^{1-s^*}(x)}(y - x).$$

By defining $h := f/F^{1-s^*}$ on $J(F)$, it follows that

$$F^{s^*}(x) \leq F^{s^*}(y) + s^* h(y)(x - y),$$

and

$$F^{s^*}(y) \leq F^{s^*}(x) + s^* h(x)(y - x).$$

After summing up the last two inequalities, it follows that

$$F^{s^*}(x) + F^{s^*}(y) \leq F^{s^*}(y) + s^* h(y)(x - y) + F^{s^*}(x) + s^* h(x)(y - x),$$

or, equivalently,

$$0 \leq s^* \left(h(x) - h(y)\right)(y - x).$$

Hence $h(x) \geq h(y)$, or equivalently, $h(\cdot)$ is a monotonically non-increasing function on $J(F)$.

The proof of the monotonicity of $\tilde{h} := f/(1 - F)^{1-s^*}$ is similar and hence is omitted.

(iii) implies (iv):

If (iii) holds, it immediately follows that $f > 0$ on $J(F) = (a, b)$. If not, suppose that $f(x_0) = 0$ for some $x_0 \in J(F)$. It follows that $h(x_0) = f(x_0)/F^{1-s^*}(x_0) = 0$. Since h is monotonically non-increasing on $J(F)$, $h(x) = 0$ for all $x \in [x_0, b)$, or, equivalently, $f = 0$ on $[x_0, b)$. Similarly, the non-decreasing monotonicity of $x \mapsto \tilde{h}(x)$ on $J(F)$ implies that $f = 0$ on $(a, x_0]$. Then $f = 0$ on $J(F)$, which violates the continuity assumption in (iii) and hence $f > 0$ on $J(F)$.

To prove f is bounded on $J(F)$, note that the monotonicities of h and $\tilde{h}$ imply that for any $x, x_0 \in J(F)$,

$$f(x) = \begin{cases} F^{1-s^*}h(x) \leq h(x) \leq h(x_0), & \text{if } x \geq x_0, \\ (1-F(x))^{1-s^*}\tilde{h}(x) \leq \tilde{h}(x) \leq \tilde{h}(x_0), & \text{if } x \leq x_0. \end{cases}$$

Hence $f(x) \leq \max\{h(x_0), \tilde{h}(x_0)\}$ for any $x, x_0 \in J(F)$.

To prove that f is differentiable on $J(F)$ almost every, we first prove that f is Lipschitz continuous on (c, d) for any $c, d \in J(F)$ with $c < d$.

By noticing the non-increasing monotonicity of h on $J(F)$, the following arguments yield an upper bound of $(f(y) - f(x))/(y - x)$ for $x, y \in (c, d)$:

$$\begin{aligned} \frac{f(y)-f(x)}{y-x} &= \frac{F^{1-s^*}(y)h(y) - F^{1-s^*}(x)h(x)}{y-x} \\ &= h(y)\frac{F^{1-s^*}(y) - F^{1-s^*}(x)}{y-x} + F^{1-s^*}(x)\frac{h(y)-h(x)}{y-x} \\ &\leq h(y)\frac{F^{1-s^*}(y) - F^{1-s^*}(x)}{y-x} \\ &= h(y)(1-s^*)f(z)F^{-s^*}(z), \end{aligned}$$

where the last equality follows from the mean value theorem and z is between x and y.

Since $-s^* < 0$, it follows that $F^{-s^*}(z) < F^{-s^*}(c)$ and hence

$$\frac{f(y)-f(x)}{y-x} \leq (1-s^*)f(z)h(y)F^{-s^*}(z) \leq (1-s^*)\max\{h(x_0), \tilde{h}(x_0)\}h(c)F^{-s^*}(c),$$

for $x, y \in (c, d)$.

Similar arguments imply that

$$\begin{aligned} \frac{f(y)-f(x)}{y-x} &= \frac{\bar{F}^{1-s^*}(y)\tilde{h}(y) - \bar{F}^{1-s^*}(x)\tilde{h}(x)}{y-x} \\ &= \tilde{h}(y)\frac{\bar{F}^{1-s^*}(y) - \bar{F}^{1-s^*}(x)}{y-x} + \bar{F}^{1-s^*}(x)\frac{\tilde{h}(y)-\tilde{h}(x)}{y-x} \\ &\geq \tilde{h}(y)\frac{\bar{F}^{1-s^*}(y) - \bar{F}^{1-s^*}(x)}{y-x} \\ &= -\tilde{h}(y)(1-s^*)\bar{F}^{-s^*}(z)f(z) \\ &\geq -(1-s^*)\max\{h(x_0), \tilde{h}(x_0)\}\tilde{h}(d)\bar{F}^{-s^*}(d). \end{aligned}$$

Hence

$$\left|\frac{f(y)-f(x)}{y-x}\right| \leq (1-s^*)\max\{h(x_0), \tilde{h}(x_0)\}\max\{h(c)F^{-s^*}(c), \tilde{h}(d)\bar{F}^{-s^*}(d)\}.$$

The last display shows that f is Lipschitz continuous on (c, d).

By proposition 4.1(iii) of Shorack (2017), page 82, f is absolutely continuous on (c, d), and hence f is

differentiable on (c, d) almost everywhere.

Since (c, d) is an arbitrary interval in (a, b), the differentiability of f on (c, d) implies the differentiability of f on (a, b) and hence f is differentiable on (a, b) with $f' = F''$ almost everywhere.

Since f is differentiable almost everywhere, the non-increasing monotonicity of h on $J(F)$ implies that

$$h'(x) \leq 0 \text{ almost everywhere on } J(F),$$

or, equivalently,

$$\log(h)'(x) \leq 0 \text{ almost everywhere on } J(F).$$

Straight-forward calculation yields that the last display is equivalent to

$$\frac{f'}{f} - (1 - s^*)\frac{f}{F} \leq 0 \text{ almost everywhere on } J(F),$$

or,

$$f' \leq (1 - s^*)\frac{f^2}{F} \text{ almost everywhere on } J(F),$$

which is the right hand side of (2.8).

Similarly, the non-decreasing monotonicity of $\tilde{h}$ implies the left hand side of (2.8).

(iv) implies (i):

Since F is continuous on $\mathbb{R}$, it suffices to prove that F^{s^*} is convex on $J(F)$ by Definition 2.1. Since we assume that F is differentiable on $J(F)$ with derivative $f = F'$, the concavity of F^{s^*} on $J(F)$ can be proved by the non-increasing monotonicity of the first derivative of F^{s^*} on $J(F)$. Since f is differentiable almost everywhere on $J(F)$, the non-increasing monotonicity of $\left(F^{s^*}\right)'$ on $J(F)$ can be proved by the non-positivity of $\left(F^{s^*}\right)''$ on $J(F)$ almost everywhere, which follows from

$$\left(F^{s^*}\right)''(x) = s^* F^{s^*-1}(x)\left(-(1 - s^*)\frac{f^2(x)}{F(x)} + f'(x)\right) \leq 0,$$

where $f = F', f' = F''$. The last inequality follows from the right hand side of (2.8).

Similarly, the concavity of $(1 - F(x))^{s^*}$, or $\bar{F}^{s^*}$, on $J(F)$ can be proved by the following arguments:

$$\left(\bar{F}^{s^*}\right)''(x) = s^* \bar{F}^{s^*-1}(x)\left(-(1 - s^*)\frac{f^2(x)}{\bar{F}(x)} - f'(x)\right) \leq 0,$$

where the last inequality follows from the left part of (2.8). □

Proof of Proposition 2.1. First some background and definitions:

- Let $a, b \geq 0$ and $\theta \in (0,1)$. The generalized mean of order $s \in \mathbb{R}$ is defined by

$$M_s(a,b;\theta) = \begin{cases} ((1-\theta)a^s + \theta b^s)^{1/s}, & \text{if } \pm s \in (0,\infty), \\ a^{1-\theta}b^{\theta}, & \text{if } s = 0, \\ \max\{a,b\}, & \text{if } s = \infty, \\ \min\{a,b\}, & \text{if } s = -\infty, \end{cases}$$

- Let (M,d) be a metric space with Borel $\sigma-$field $\mathcal{M}$. A measure μ on $\mathcal{M}$ is called t-concave if for nonempty sets $A, B \in \mathcal{M}$ and $0 < \theta < 1$ we have

$$\mu_* \left((1-\theta)A + \theta B\right) \geq M_t \left(\mu_*(A), \mu_*(B); \theta\right)$$

  where μ_* is the inner measure corresponding to μ (which is needed in general in view of examples noted by Erdős and Stone (1970)).

- A non-negative real-valued function h on (M,d) is called s-concave if for $x, y \in M$ and $0 < \theta < 1$ we have

$$h\left((1-\theta)x + \theta y\right) \geq M_s\left(h(x), h(y); \theta\right).$$

  See Chapter 3.3 in Dharmadhikari and Joag-Dev (1988) for more details of the definitions of $M_s(a,b;\theta)$, t-concave and s-concave.

- Suppose $(M,d) = (\mathbb{R}^k, |\cdot|)$, $k-$dimensional Euclidean space with the usual Euclidean metric and suppose that f is an s-concave density function with respect to Lebesgue measure λ on $\mathcal{B}_k$, and consider the probability measure μ on $\mathcal{B}_k$ defined by

$$\mu(B) = \int_B f d\lambda \text{ for all } B \in \mathcal{B}_k.$$

  Then by a theorem of Borell (1975), Brascamp and Lieb (1976) and Rinott (1976), the measure μ is s^*-concave where $s^* = 1/(1+ks)$ if $s \in (-1/k, \infty)$ and $s^* = 0$ if $s = 0$.

- Here we are in the case $k = 1$. Thus for $s \in (-1, \infty)$ the measure μ is s^* concave: for $s \in (-1, \infty)$, $A, B \in \mathcal{B}_1$, and $0 < \theta < 1$,

$$\mu_* \left((1-\theta)A + \theta B\right) \geq M_{s^*} \left(\mu_*(A), \mu_*(B); \theta\right); \tag{2.17}$$

  here μ_* denotes inner measure corresponding to μ.

With this preparation we can give our proof of Proposition 2.1: if $A = (-\infty, x]$ and $B = (-\infty, y]$ for $x, y \in$

$J(F)$, it is easily seen that

$$\begin{aligned}(1-\theta)A+\theta B &= \{(1-\theta)x'+\theta y' : x' \le x, y' \le y\}\\ &\subset \{(1-\theta)x'+\theta y' : (1-\theta)x'+\theta y' \le (1-\theta)x+\theta y\}\\ &= (-\infty,(1-\theta)x+\theta y].\end{aligned}$$

Therefore, with the second inequality following from (2.17),

$$\begin{aligned}F\left((1-\theta)x+\theta y\right) &= \mu\left((-\infty,(1-\theta)x+\theta y]\right)\\ &\ge \mu\left((1-\theta)(-\infty,x]+\theta(-\infty,y]\right)\\ &\ge M_{s^*}\left(\mu((-\infty,x]),\mu((-\infty,y]);\theta\right) = M_{s^*}(F(x),F(y);\theta);\end{aligned}$$

i.e. F is s^*-concave. Similarly, taking $A=(x,\infty)$ and $B=(y,\infty)$ it follows that $1-F$ is s^*-concave.

Note that this argument contains the case $s^*=0$. □

Proof of Proposition 2.2. By Theorem 2.1, for any $F \in \mathcal{P}_{s^*}$, F is continuous on $\mathbb{R}$ and differentiable on $J(F)$ with derivative $f=F'$. Furthermore, f is differentiable almost everywhere on $J(F)$ with derivative $f'=F''$ satisfying (2.8).

For any $t^* \le s^*$, by noting that $1-s^* \le 1-t^*$ and $-(1-s^*) \ge -(1-t^*)$, it follows that

$$-(1-t^*)\frac{f^2}{1-F} \le -(1-s^*)\frac{f^2}{1-F} \le f' \le (1-s^*)\frac{f^2}{F} \le (1-t^*)\frac{f^2}{F},$$

almost everywhere on $J(F)$. Hence $F \in \mathcal{P}_t^*$ by Theorem 2.1. This proves (2.1).

To prove (2.2), note that for any $F \in \cup_{s^*>0}\mathcal{P}_{s^*}$, F is continuous on $\mathbb{R}$ and differentiable on $J(F)$ with derivative $f=F'$. Furthermore, f is differentiable almost everywhere on $J(F)$ with derivative $f'=F''$ satisfying (2.8), i.e.

$$-(1-s^*)\frac{f^2}{1-F} \le f' \le (1-s^*)\frac{f^2}{F} \text{ almost everywhere on } J(F),$$

for all $s^*>0$. By taking $s^* \to 0$, it follows that

$$-(1-0)\frac{f^2}{1-F} \le f' \le (1-0)\frac{f^2}{F} \text{ almost everywhere on } J(F).$$

The last display is equivalent to $F \in \mathcal{P}_0$ by Theorem 2.1. This proves that the left hand side of (2.2) holds. Similarly, one can prove the right hand side of (2.2); the details are omitted. □

Proof of Corollary 2.1. To prove the right part of (2.9), note that (2.8) implies that

$$1-s^* \ge \frac{Ff'}{f^2} \text{ and } 1-s^* \ge -\frac{(1-F)f'}{f^2}$$

almost everywhere on $J(F)$, or equivalently,

$$1 - s^* \geq \max\left\{\operatorname{esssup}_{x\in J(F)} \frac{Ff'}{f^2}, \operatorname{esssup}_{x\in J(F)} -\frac{(1-F)f'}{f^2}\right\}.$$

Replacing $\operatorname{esssup}_{x\in J(F)} Ff'/f^2$ and $\operatorname{esssup}_{x\in J(F)} -(1-F)f'/f^2$ by $\widetilde{CR}(F)$ and $\widetilde{CR}(\overline{F})$, it follows that

$$1 - s^* \geq \max\{\widetilde{CR}(F), \widetilde{CR}(\overline{F})\} = \overline{\gamma}(F).$$

One can prove the left two inequalities of (2.9) by the following arguments:

$$\begin{aligned}
&\overline{\gamma}(F) = \max\{\widetilde{CR}(F), \widetilde{CR}(\overline{F})\} \\
&= \max\left\{\operatorname{esssup}_{x\in J(F)} \frac{F(x)f'(x)}{f(x)^2}, \operatorname{esssup}_{x\in J(F)} -\frac{(1-F(x))f'(x)}{f(x)^2}\right\} \\
&= \max\left\{\operatorname{esssup}_{x\in J(F)} \frac{F(x)f'(x)}{f(x)^2}1_{[f'(x)\geq 0]}, \operatorname{esssup}_{x\in J(F)} -\frac{(1-F(x))f'(x)}{f(x)^2}1_{[f'(x)\leq 0]}\right\} \\
&= \max\left\{\operatorname{esssup}_{x\in J(F)} \frac{F(x)|f'(x)|}{f(x)^2}1_{[f'(x)\geq 0]}, \operatorname{esssup}_{x\in J(F)} \frac{(1-F(x))|f'(x)|}{f(x)^2}1_{[f'(x)\leq 0]}\right\} \\
&\geq \max\left\{\operatorname{esssup}_{x\in J(F)} \frac{F(x)\wedge(1-F(x))|f'(x)|}{f(x)^2}1_{[f'(x)\geq 0]},\right. \\
&\qquad \left.\operatorname{esssup}_{x\in J(F)} \frac{F(x)\wedge(1-F(x))|f'(x)|}{f(x)^2}1_{[f'(x)\leq 0]}\right\} \\
&= \operatorname{esssup}_{x\in J(F)} \frac{F(x)\wedge(1-F(x))|f'(x)|}{f(x)^2} \\
&= \gamma(F) \geq \tilde{\gamma}(F)
\end{aligned}$$

where the last inequality holds since $u \wedge (1-u) \geq u(1-u)$ for $0 \leq u \leq 1$. □

Proof of Corollary 2.2. Note that for $s^* < 0$ and $y > -1$, we have $(1+y)^{s^*} \geq 1 + s^* y$. Replacing y by $-F(x)$, where $x \in J(F)$, it follows that

$$(1 - F(x))^{s^*} \geq 1 - s^* F(x),$$

or, by rearranging,

$$F(x) \leq \frac{1}{s^*}\left(1 - (1-F(x))^{s^*}\right) = F_U(x),$$

where F_U is a convex function on $J(F)$ if $F \in \mathcal{P}_{s^*}$. This proves the right hand side of (2.10) for $s^* < 0$. Similarly, replacing y by $-(1-F(x))$, where $x \in J(F)$, by rearranging terms, it follows that

$$F(x) \geq \frac{1}{s^*}\left(F^{s^*}(x) - (1-s^*)\right) = F_L(x),$$

which proves the left hand side of (2.10) for $s^* < 0$.

Similarly, for $1 \geq s^* > 0$ and $y > -1$, we have $(1+y)^{s^*} \leq 1 + s^* y$. Replacing y by $-F(x)$, where $x \in J(F)$,

it follows that

$$(1-F(x))^{s^*} \leq 1 - s^* F(x),$$

or, by rearranging,

$$F(x) \leq \frac{1}{s^*}\left(1-(1-F(x))^{s^*}\right) = F_U(x),$$

where F_U is a convex function on $J(F)$ if $F \in \mathcal{P}_{s^*}$. This proves the right hand side of (2.10) for $s^* > 0$. Similarly, replacing y by $-(1-F(x))$, where $x \in J(F)$, by rearranging terms, it follows that

$$F(x) \geq \frac{1}{s^*}\left(F^{s^*}(x) - (1-s^*)\right) = F_L(x),$$

which proves the left hand side of (2.10) for $s^* > 0$. □

Proof of Lemma 2.1. If there is no $G \in \mathcal{P}_{s^*}$ fitting in between L_n and U_n, it follows that $L_n^o := 1$ and $U_n^o := 0$ and assertions in both (i) and (ii) are trivial. In the following proof, we let $G \in \mathcal{P}_{s^*}$ such that $L_n \leq G \leq U_n$.

(i) It suffices to prove that for any $x \in J(G)$ the density function $g = G'$ satisfies $g(x) \leq \max\{\gamma_1, \gamma_2\}$, because this is equivalent to Lipschitz-continuity of G with the latter constant, and this property carries over to the pointwise infimum L_n^o and supremum U_n^o.

To prove $g(x) \leq \max\{\gamma_1, \gamma_2\}$, note that g/G^{1-s^*} is monotonically non-increasing on $J(G)$ (see Theorem 2.1(iii)), it follows that for $x \geq b$

$$\begin{aligned}
\frac{g(x)}{G^{1-s^*}(x)} &\leq \frac{g(b)}{G^{1-s^*}(b)} = \left(\frac{1}{s^*}G^{s^*}\right)'(b) \\
&\leq \frac{\frac{1}{s^*}G^{s^*}(b) - \frac{1}{s^*}G^{s^*}(a)}{b-a} \\
&\leq \frac{\frac{1}{s^*}\left(v^{s^*} - u^{s^*}\right)}{b-a} = \gamma_1.
\end{aligned}$$

The last inequality follows from noting that $x \mapsto (1/s^*)x^{s^*}$ is a monotonically non-decreasing function for all $s^* \neq 0$, $G(b) \leq U_n(b) \leq v$ and $G(a) \geq L_n(a) \geq u$. Hence

$$g(x) \leq G^{1-s^*}(x)\gamma_1 \leq \gamma_1 \text{ for } x \geq b.$$

Similarly, by noting that $g/(1-G)^{1-s^*}$ is monotonically non-decreasing on $J(G)$ (see Theorem 2.1(iii)), it follows that for $x \leq a$

$$\begin{aligned}
\frac{g(x)}{(1-G(x))^{1-s^*}} &\leq \frac{g(a)}{(1-G(a))^{1-s^*}} \\
&= \left(\frac{-1}{s^*}(1-G)^{s^*}\right)'(a) \\
&\leq \frac{\frac{-1}{s^*}(1-G(b))^{s^*} - \frac{-1}{s^*}(1-G(a))^{s^*}}{b-a}
\end{aligned}$$

$$\leq \quad \frac{\frac{-1}{s^*}\left((1-v)^{s^*}-(1-u)^{s^*}\right)}{b-a} = \gamma_2.$$

The last inequality follows from noting that $x \mapsto -(1/s^*)(1-x)^{s^*}$ is a monotonically non-decreasing function for all $s^* \neq 0$, $G(b) \leq v$ and $G(a) \geq u$. Hence

$$g(x) \leq (1-G(x))^{1-s^*}\gamma_2 \leq \gamma_2 \text{ for } x \leq a.$$

For $a < x < b$, analogously, we get two following inequalities

$$\begin{aligned} g(x) &= G^{1-s^*}(x)\frac{g(x)}{G^{1-s^*}(x)} \\ &\leq G^{1-s^*}(x)\frac{\frac{1}{s^*}G^{s^*}(x)-\frac{1}{s^*}G^{s^*}(a)}{x-a} \\ &= \frac{1}{s^*}\frac{1}{x-a}\left(G(x)-G^{s^*}(a)G^{1-s^*}(x)\right) \end{aligned}$$

and

$$\begin{aligned} g(x) &= (1-G(x))^{1-s^*}\frac{g(x)}{(1-G(x))^{1-s^*}} \\ &\leq (1-G(x))^{1-s^*}\frac{\frac{-1}{s^*}(1-G(b))^{s^*}-\frac{-1}{s^*}(1-G(x))^{s^*}}{b-x} \\ &= \frac{1}{s^*}\frac{1}{b-x}\left(1-G(x)-(1-G(b))^{s^*}(1-G(x))^{1-s^*}\right). \end{aligned}$$

The former inequality times $(x-a)$ plus the latter inequality times $(b-x)$ yields

$$g(x) \quad \leq \quad \frac{1}{s^*}\frac{1-G^{s^*}(a)G^{1-s^*}(x)-(1-G(b))^{s^*}(1-G(x))^{1-s^*}}{b-a} = \frac{h(G(x))}{b-a},$$

where

$$h(y) := \frac{1}{s^*}\left(1-G^{s^*}(a)y^{1-s^*}-(1-G(b))^{s^*}(1-y)^{1-s^*}\right) \text{ for } y \in (0,1).$$

Since

$$h''(y) = (1-s^*)\left(G^{s^*}(a)y^{-s^*-1}+(1-G(b))^{s^*}(1-y)^{-s^*-1}\right) \geq 0,$$

it follows that $h(y)$ is convex on $(0,1)$ and hence

$$g(x) \leq \max_{y\in\{G(a),G(b)\}}\frac{h(y)}{b-a} = \max\left\{\frac{h(G(a))}{b-a},\frac{h(G(b))}{b-a}\right\}.$$

Note that

$$\frac{h(G(a))}{b-a} = (1-G(a))^{1-s^*}\frac{\frac{-1}{s^*}(1-G(b))^{s^*}-\frac{-1}{s^*}(1-G(a))^{s^*}}{b-a} \leq \gamma_2$$

and

$$\frac{h(G(b))}{b-a} = G(b)^{1-s^*}\frac{\frac{1}{s^*}G^{s^*}(b)-\frac{1}{s^*}G^{s^*}(a)}{b-a} \leq \gamma_1.$$

Hence $g(x) \le \max\{\gamma_1, \gamma_2\}$ for $a < x < b$.

(ii) By Theorem 2.1(ii), it follows that for $x \le a$

$$G(x) \le G(a)\left(1 + s^* \frac{g(a)}{G(a)}(x-a)\right)_+^{1/s^*} = \left(G^{s^*}(a) + s^* \frac{g(a)}{G^{1-s^*}(a)}(x-a)\right)_+^{1/s^*}.$$

By Theorem 2.1(iii), the non-increasing monotonicity of g/G^{1-s^*} implies that

$$\frac{g(a)}{G^{1-s^*}(a)} = \left(\frac{1}{s^*}G^{s^*}\right)'(a) \ge \frac{\frac{1}{s^*}G^{s^*}(b) - \frac{1}{s^*}G^{s^*}(a)}{b-a} \ge \frac{\frac{1}{s^*}v^{s^*} - \frac{1}{s^*}u^{s^*}}{b-a} = \gamma_1.$$

The last inequality follows from noting that $G(a) \le U_n(a) \le u$ and $G(b) \ge L_n(b) \ge v$. Since $x - a \le 0$, it follows that

$$\begin{aligned} G(x) &\le \left(G^{s^*}(a) + s^* \frac{g(a)}{G^{1-s^*}(a)}(x-a)\right)_+^{1/s^*} \\ &\le \left(G^{s^*}(a) + s^*\gamma_1(x-a)\right)_+^{1/s^*} \\ &\le \left(u^{s^*} + s^*\gamma_1(x-a)\right)_+^{1/s^*}. \end{aligned}$$

The last inequality follows from noting that $G(a) \le u$.

On the other hand, by Theorem 2.1(ii), it follows that for $x \ge b$

$$\begin{aligned} 1 - G(x) &\le (1-G(b))\left(1 - s^* \frac{g(b)}{1-G(b)}(x-b)\right)_+^{1/s^*} \\ &= \left((1-G(b))^{s^*} - s^* \frac{g(b)}{(1-G(b))^{1-s^*}}(x-b)\right)_+^{1/s^*} \\ &\le \left((1-v)^{s^*} - s^* \frac{g(b)}{(1-G(b))^{1-s^*}}(x-b)\right)_+^{1/s^*}. \end{aligned}$$

The last inequality follows from noting that $1 - G(b) \le 1 - v$. By Theorem 2.1(iii), the non-decreasing monotonicity of $g/(1-G)^{1-s^*}$ implies that

$$\begin{aligned} \frac{g(b)}{(1-G)^{1-s^*}(b)} &= (\frac{-1}{s^*}(1-G)^{s^*})'(b) \\ &\ge \frac{\frac{-1}{s^*}(1-G(b))^{s^*} - \frac{-1}{s^*}(1-G(a))^{s^*}}{b-a} \\ &= \frac{\frac{1}{s^*}\left((1-G(a))^{s^*} - (1-G(b))^{s^*}\right)}{b-a} \\ &\ge \frac{\frac{1}{s^*}\left((1-u)^{s^*} - (1-v)^{s^*}\right)}{b-a} \\ &= \gamma_2\,. \end{aligned}$$

The last inequality follows from noting that $G(a) \le U_n(a) \le u$ and $G(b) \ge L_n(b) \ge v$. Since $x - b \ge 0$, it

follows that

$$1 - G(x) \leq \Big((1-v)^{s^*} - s^* \gamma_2 (x-b) \Big)_+^{1/s^*}.$$

□

Proof of Theorem 2.2. The following proof is analogous to the proof of theorem 3 in Dümbgen et al. (2017), in which they proved the result in the case $s^* = 0$. In the following proof we assume that $s^* \neq 0$.

(i) Suppose $s^* > 0$. Since F is not bi-s^*-concave, it follows that F^{s^*} or $(1-F)^{s^*}$ is not concave. Without loss of generality, we assume that F^{s^*} is not concave and hence there exist real numbers $x_0 < x_1 < x_2$ such that $F^{s^*}(x_1) < (1-\lambda)F^{s^*}(x_0) + \lambda F^{s^*}(x_2)$, where $\lambda := (x_1 - x_0)/(x_2 - x_0) \in (0,1)$. By the consistency of L_n and U_n, it follows that, with probability tending to one, $U_n^{s^*}(x_1) < (1-\lambda)L_n^{s^*}(x_0) + \lambda L_n^{s^*}(x_2)$ and hence

$$G^{s^*}(x_1) < (1-\lambda)G^{s^*}(x_0) + \lambda G^{s^*}(x_2),$$

for any G such that $L_n \leq G \leq U_n$. Therefore, there are no bi-s^*-concave distribution functions fitting between L_n and U_n and hence $L_n^o = 1$ and $U_n^o = 0$ with probability tending to one.

The proof of the case $s^* < 0$ is similar and hence is omitted.

(ii) Suppose $F \in \mathcal{P}_{s^*}$. Note that since (L_n, U_n) is a $(1-\alpha)$ confidence band for F, it follows that $P(L_n^o \leq U_n^o) \geq P(L_n \leq F \leq U_n) \geq 1 - \alpha$.

If $\{G \in \mathcal{P}_{s^*} : L_n \leq G \leq U_n\}$ is empty, it follows that $L_n^o = 1$ and $U_n^o = 0$ and hence the assertions are trivial. In the following proof, we assume that $\{G \in \mathcal{P}_{s^*} : L_n \leq G \leq U_n\}$ is not empty.

To prove (2.11), we first prove that $\|L_n - F\|_\infty \to_p 0$ and $\|U_n - F\|_\infty \to_p 0$. By the continuity of F, for any $m \in \mathbb{N}^+$ with $m \geq 2$, there exist real numbers $\{x_i\}_{i=1}^{m-1}$ such that $F(x_i) = i/m$, $i = 1, \ldots, m-1$. Furthermore, define $x_0 = -\infty$ and $x_m = \infty$.

By the non-decreasing monotonicity of L_n and F, it follows that for $x \in [x_{i-1}, x_i]$

$$L_n(x) - F(x) \leq L_n(x_i) - F(x_{i-1}) = L_n(x_i) - (F(x_i) - \frac{1}{m}) = L_n(x_i) - F(x_i) + \frac{1}{m},$$

and

$$\begin{aligned} L_n(x) - F(x) &\geq L_n(x_{i-1}) - F(x_i) \\ &= L_n(x_{i-1}) - (F(x_{i-1}) + \frac{1}{m}) = L_n(x_{i-1}) - F(x_{i-1}) - \frac{1}{m}. \end{aligned}$$

Hence

$$|L_n(x) - F(x)| \leq \max_{i=1,\ldots,m-1} |L_n(x_i) - F(x_i)| + \frac{1}{m}$$

for $x \in [x_{i-1}, x_i]$. Note that

$$\|L_n - F\|_\infty = \sup_{x\in\mathbb{R}} |L_n(x) - F(x)| = \max_{i=1,\ldots,m} \sup_{x\in[x_{i-1},x_i]} |L_n(x) - F(x)|,$$

it follows that

$$\|L_n - F\|_\infty \le \max_{i=1,\ldots,m-1} |L_n(x_i) - F(x_i)| + \frac{1}{m},$$

and hence pointwise convergence implies uniform convergence. An analogous proof shows that $\|U_n - F\|_\infty \to_p 0$ and is omitted.

Combining $\|L_n - F\|_\infty \to_p 0$ and $\|U_n - F\|_\infty \to_p 0$ implies that

$$\sup_{G\in\mathcal{P}_{s^*}: L_n \le G \le U_n} \|G - F\|_\infty \le \|L_n - F\|_\infty + \|U_n - F\|_\infty \to_p 0.$$

To prove (2.12) in the case that $h_G = \left(G^{s^*}\right)'$, it suffices to prove that

$$\sup_{G\in\mathcal{P}_{s^*}: L_n \le G \le U_n} \left\| \left(G^{s^*}/s^*\right)' - \left(F^{s^*}/s^*\right)' \right\|_{K,\infty} \to_p 0. \tag{2.18}$$

Note that $h_G/s^* = G'/G^{1-s^*}$. Since K is a compact interval in $J(F)$ and $h_F/s^* = f/F^{1-s^*}$ is continuous and non-increasing on $J(F)$, for any fixed $\epsilon > 0$ there exist points $a_0 < a_1 < \cdots < a_m < a_{m+1}$ in $J(F)$ such that $K \subset [a_1, a_m]$ and

$$0 \le \frac{1}{s^*} h_F(a_{i-1}) - \frac{1}{s^*} h_F(a_i) \le \epsilon \text{ for } 1 \le i \le m+1.$$

For $G \in \mathcal{P}_{s^*}$ with $L_n \le G \le U_n$, for any $x \in K$ it follows from the monotonicity of h_F/s^* and h_G/s^* that

$$\begin{aligned}
\sup_{x\in K} \left(\frac{1}{s^*} h_G(x) - \frac{1}{s^*} h_F(x) \right) &\le \max_{i=1,\ldots,m-1} \left(\frac{1}{s^*} h_G(a_i) - \frac{1}{s^*} h_F(a_{i+1}) \right) \\
&\le \max_{i=1,\ldots,m-1} \left(\frac{\frac{1}{s^*} G^{s^*}(a_i) - \frac{1}{s^*} G^{s^*}(a_{i-1})}{a_i - a_{i-1}} \right. \\
&\qquad \left. - \frac{1}{s^*} h_F(a_{i+1}) \right) \\
&\le \max_{i=1,\ldots,m-1} \left(\frac{\frac{1}{s^*} U_n^{s^*}(a_i) - \frac{1}{s^*} L_n^{s^*}(a_{i-1})}{a_i - a_{i-1}} \right. \\
&\qquad \left. - \frac{1}{s^*} h_F(a_{i+1}) \right) \\
&= \max_{i=1,\ldots,m-1} \left(\frac{\frac{1}{s^*} F^{s^*}(a_i) - \frac{1}{s^*} F^{s^*}(a_{i-1})}{a_i - a_{i-1}} \right. \\
&\qquad \left. - \frac{1}{s^*} h_F(a_{i+1}) \right) + o_p(1) \\
&\le \max_{i=1,\ldots,m-1} \left(\frac{1}{s^*} h_F(a_{i-1}) - \frac{1}{s^*} h_F(a_{i+1}) \right) \\
&\quad + o_p(1)
\end{aligned}$$

$$\leq \quad 2\epsilon + o_p(1).$$

Analogously,

$$\begin{aligned}\sup_{x\in K}\left(\frac{1}{s^*}h_F(x) - \frac{1}{s^*}h_G(x)\right) &\leq \max_{i=1,\ldots,m-1}\left(\frac{1}{s^*}h_F(a_i) - \frac{1}{s^*}h_F(a_{i+2})\right) + o_p(1)\\ &\leq 2\epsilon + o_p(1).\end{aligned}$$

Since $\epsilon > 0$ is arbitrarily small, this shows that (2.18) holds.

The proof of (2.12) in the case that $h_G = \left((1-G)^{s^*}\right)'$ is similar and hence is omitted.

Since $G' = G^{1-s^*}\left(G^{s^*}/s^*\right)'$, it follows from (2.18) that (2.12) holds in the case that $h_G = G'$.

Finally, let $x_1 < \sup J(F)$ and $b_1 < f(x_1)/F^{1-s^*}(x_1)$. As in the proof of Lemma 2.1(ii) an analogous argument implies that for any $x_1' > x_1$, $x_1' \in J(F)$,

$$U_n^o(x) \quad \leq \quad \left(U_n^{s^*}(x') + s^*\frac{\frac{1}{s^*}L_n^{s^*}(x_1') - \frac{1}{s^*}U_n^{s^*}(x_1)}{x_1' - x_1}(x - x')\right)_+^{1/s^*}$$

for all $x \leq x' \leq x_1$.

Note that by the consistency of L_n and U_n and letting $x_1' \downarrow x_1$, it follows that.

$$\frac{\frac{1}{s^*}L_n^{s^*}(x_1') - \frac{1}{s^*}U_n^{s^*}(x_1)}{x_1' - x_1} \to_p \frac{\frac{1}{s^*}F^{s^*}(x_1') - \frac{1}{s^*}F^{s^*}(x_1)}{x_1' - x_1} > b_1.$$

Hence with probability tending to one,

$$U_n^o(x) \leq \left(U_n^{s^*}(x') + s^* b_1(x - x')\right)_+^{1/s^*},$$

for all $x \leq x' \leq x_1$. The proof of (2.14) is similar and hence is omitted. □

Proof of Remark 2.1. (ii) By Theorem 2.1(ii), if $s^* > 0$ and $\inf J(F) = -\infty$, it follows that for arbitrary $x \in J(F)$,

$$F(y) \leq F(x)\cdot\left(1 + s^*\frac{f(x)}{F(x)}(y - x)\right)_+^{1/s^*} = 0$$

for small enough y such that

$$1 + s^*\frac{f(x)}{F(x)}(y - x) < 0.$$

This violates the assumption that $\inf J(F) = -\infty$ and hence $\inf J(F) > -\infty$.

The finiteness of $\sup J(F)$ can be proved similarly and hence is omitted.

(iv) Note that the claim holds automatically if $\inf J(F) > -\infty$ and $\sup J(F) < \infty$.

In the following proof, we focus on the case that $\inf J(F) = -\infty$ and $\sup J(F) < \infty$, and it suffices to show that $\int |x|^t dF(x)$ is finite for $t \in (0, (-1)/s^*)$.

Note that

$$\begin{aligned}\int |x|^t dF(x) &= E|X|^t = \int_0^\infty P(|X|^t > a)da = \int_0^\infty P(|X| > a^{1/t})da \\ &= \int_0^\infty ta^{t-1}P(|X| > a)da \\ &= \int_0^\infty ta^{t-1}P(X > a)da + \int_0^\infty ta^{t-1}P(X < -a)da.\end{aligned}$$

Since $\sup J(F)$ is finite, the first term of the last display is finite and hence it suffices to prove that $ta^{t-1}P(X < -a)$ is integrable for $t < (-1)/s^*$.

It follows from Theorem 2.1(ii) that for any a large enough and $x \in J(F)$,

$$P(X < -a) \le F(x)\left(1 + \frac{s^* f(x)(-a-x)}{F(x)}\right)_+^{1/s^*} = F(x)\left(1 + \frac{-s^* f(x)(a+x)}{F(x)}\right)_+^{1/s^*}.$$

Thus $ta^{t-1}P(X < -a)$ is integrable for $t < (-1)/s^*$, since

$$\begin{aligned}ta^{t-1}P(X < -a) &\le tF(x)a^{t-1}\left(1 + \frac{-s^* f(x)(a+x)}{F(x)}\right)_+^{1/s^*} \\ &= tF(x)\left(\frac{-s^* f(x)}{F(x)}\right)^{1/s^*} a^{t-1}\left(a + x + \frac{F(x)}{-s^* f(x)}\right)_+^{1/s^*} \\ &\le 2tF(x)\left(\frac{-s^* f(x)}{F(x)}\right)^{1/s^*} a^{t+1/s^*-1}\end{aligned}$$

for a large enough and a^{t+1/s^*-1} is integrable for $t < (-1)/s^*$.

For other cases, the proof is similar and hence is omitted. □

Proof of Corollary 2.3. Suppose that x_0 is a point in $J(F)$. Notice that for any $z \in \mathbb{R}$,

$$\phi(z) - \phi(x_0) = \int_{\mathbb{R}} \left(1_{[x_0 \le x < z]} - 1_{[z \le x < x_0]}\right) \phi'(x)dx,$$

and hence by Fubini's theorem, it follows that

$$\int_{\mathbb{R}} \phi dG = \phi(x_0) + \int_{\mathbb{R}} \phi'(x)\left(1_{[x \ge x_0]} - G(x)\right)dx, \tag{2.19}$$

provided that

$$\int_{\mathbb{R}} |\phi'(x)|\left|1_{[x \ge x_0]} - G(x)\right|dx < \infty.$$

To prove the last display, note that for any $b_1 \in (0, T_1(F))$ and $b_2 \in (0, T_2(F))$, there exist points $x_1, x_2 \in J(F)$ with $x_1 \le x_0 \le x_2$ and

$$\frac{f}{F^{1-s^*}}(x_1) > b_1, \quad \frac{f}{(1-F)^{1-s^*}}(x_2) > b_2.$$

Then it follows from Theorem 2.2(ii) that with probability tending to one,

$$U_n^o(x) \le \left(U_n^{s^*}(x_1) + s^* b_1(x - x_1)\right)_+^{1/s^*} \text{ for } x \le x_1,$$

and

$$1 - L_n^o(x) \le \left((1 - L_n(x_2))^{s^*} - s^* b_2(x - x_2)\right)_+^{1/s^*} \text{ for } x \ge x_2.$$

Hence for any $c > \max\{|x_1|, |x_2|\}$, it follows that

$$\begin{aligned}
\int_{-\infty}^{x_1 - c} |\phi'(x)| \left|1_{[x \ge x_0]} - G(x)\right| dx &= \int_{-\infty}^{x_1 - c} |\phi'(x)| \, G(x) dx \\
&\le \int_{-\infty}^{x_1 - c} |\phi'(x)| \, U_n^o(x) dx \\
&\le \int_{-\infty}^{x_1 - c} |\phi'(x)| \left(U_n^{s^*}(x_1) + s^* b_1(x - x_1)\right)_+^{1/s^*} dx \\
&= \int_{-\infty}^{x_1 - c} |\phi'(x)| \left(U_n^{s^*}(x_1) + s^* b_1(x - x_1)\right)^{1/s^*} dx.
\end{aligned}$$

Since $|\phi'(x)| \le a|x|^{k-1}$, it follows that the last display is no larger than

$$\int_{-\infty}^{x_1 - c} a|x|^{k-1} \left(U_n^{s^*}(x_1) + s^* b_1(x - x_1)\right)^{1/s^*} dx,$$

which is finite by noting that $k - 1 + 1/s^* < -1$. Analogously, one can prove that for $c > \max\{|x_1|, |x_2|\}$,

$$\int_{x_2 + c}^{\infty} |\phi'(x)| \left|1_{[x \ge x_0]} - G(x)\right| dx \le \int_{x_2 + c}^{\infty} |\phi'(x)| \, |1 - L_n^o(x)| \, dx < \infty.$$

Since ϕ' is continuous on $\mathbb{R}$, it follows that for any $c > \max\{|x_1|, |x_2|\}$,

$$\int_{x_1 - c}^{x_2 + c} |\phi'(x)| \left|1_{[x \ge x_0]} - G(x)\right| dx < \infty$$

and hence

$$\int_{\mathbb{R}} |\phi'(x)| \left|1_{[x \ge x_0]} - G(x)\right| dx < \infty.$$

By (2.19), it follows that

$$\sup_{G: L_n^o \le G \le U_n^o} \left| \int \phi dG - \int \phi dF \right| = \sup_{G: L_n^o \le G \le U_n^o} \left| \int \phi'(x)(F - G)(x) dx \right|,$$

which is not larger than

$$\begin{aligned}
&\sup_{G: L_n^o \le G \le U_n^o} \|G - F\|_\infty \int_{x_1 - c}^{x_2 + c} |\phi'(x)| dx \\
&\quad + \int_{-\infty}^{x_1 - c} |\phi'(x)| (F + U_n^o)(x) dx + \int_{x_2 + c}^{\infty} |\phi'(x)| (1 - F + 1 - L_n^o)(x) dx \\
\le \quad & o_p(1) + 2 \int_{-\infty}^{x_1 - c} |\phi'(x)| U_n^o(x) dx + 2 \int_{x_2 + c}^{\infty} |\phi'(x)| (1 - L_n^o(x)) dx.
\end{aligned}$$

Note that the last two terms go to zero as c goes to infinity by their integrability and hence

$$\sup_{G:L_n^o \leq G \leq U_n^o} \left| \int \phi dG - \int \phi dF \right| = o_p(1).$$

□

Proof of Theorem 2.3. It follows from the proof of Corollary 2.3 that

$$\sup_{G:L_n^o \leq G \leq U_n^o} \left| \int \phi dG - \int \phi dF \right| = \sup_{G:L_n^o \leq G \leq U_n^o} \left| \int \phi'(x)(F-G)(x)dx \right|$$

and hence

$$\sup_{G:L_n^o \leq G \leq U_n^o} \left| \int \phi dG - \int \phi dF \right| \leq \sup_{G:L_n^o \leq G \leq U_n^o} \int |\phi'(x)| \, |(G-F)(x)| \, dx.$$

It suffices to bound $|G-F|$ on $\mathbb{R}$, where G is between L_n^o and U_n^o.

It follows from $G \leq U_n^o \leq U_n$ and Condition (*) that on the interval $\{\lambda n^{-1/(2-2\gamma)} \leq \mathbb{F}_n \leq 1 - \lambda n^{-1/(2-2\gamma)}\}$,

$$G - F \leq U_n^o - F \leq U_n - F \leq U_n - \mathbb{F}_n + \mathbb{F}_n - F \leq \kappa n^{-1/2} \left(\mathbb{F}_n \left(1 - \mathbb{F}_n \right) \right)^{\gamma} + |\mathbb{F}_n - F|$$

To bound $|\mathbb{F}_n - F|$, it follows from theorem 3.7.1, page 141, Shorack and Wellner (2009) that

$$\left\| \frac{\sqrt{n} \left(\mathbb{F}_n - F \right) - \mathbb{U} \circ F}{\left(F(1-F) \right)^{\gamma}} \right\| \to_p 0$$

by verifying that $q(t) := (t(1-t))^{\gamma}$ with $0 \leq \gamma < 1/2$ is monotonically increasing on $[0, 1/2]$, symmetric about $1/2$ and $\int_0^1 q^{-2}(t)dt < \infty$, where $\mathbb{U}$ is Brownian bridge on $[0,1]$.

Hence for any fixed $\epsilon \in (0,1)$ there exists a constant $\kappa_\epsilon > 0$ such that with probability at least $1-\epsilon$,

$$|\mathbb{F}_n - F| \leq \kappa_\epsilon n^{-1/2} \left(F(1-F) \right)^{\gamma}$$

on $\mathbb{R}$. Thus, it follows that on the interval $\{\lambda n^{-1/(2-2\gamma)} \leq \mathbb{F}_n \leq 1 - \lambda n^{-1/(2-2\gamma)}\}$,

$$G - F \leq \kappa n^{-1/2} \left(\mathbb{F}_n \left(1 - \mathbb{F}_n \right) \right)^{\gamma} + \kappa_\epsilon n^{-1/2} \left(F(1-F) \right)^{\gamma}.$$

To bound $\mathbb{F}_n \left(1 - \mathbb{F}_n \right)$ by $F(1-F)$, note that

$$\begin{aligned}
\mathbb{F}_n \left(1 - \mathbb{F}_n \right) &= \left(\mathbb{F}_n - F + F \right) \left(1 - F + F - \mathbb{F}_n \right) \\
&= \left(\mathbb{F}_n - F \right) \left(1 - F \right) + F(1-F) - \left(\mathbb{F}_n - F \right)^2 - F \left(\mathbb{F}_n - F \right) \\
&= F(1-F) + \left(\mathbb{F}_n - F \right) \left(1 - 2F \right) - \left(\mathbb{F}_n - F \right)^2 \\
&\leq F(1-F) + |\mathbb{F}_n - F| \, |1 - 2F| + |\mathbb{F}_n - F|^2 \\
&\leq F(1-F) + |\mathbb{F}_n - F| + |\mathbb{F}_n - F|
\end{aligned}$$

$$
\begin{aligned}
&= F(1-F)\cdot\left(1+\frac{2\left|\mathbb{F}_n-F\right|}{F(1-F)}\right)\\
&\leq F(1-F)\cdot\left(1+\frac{4\left|\mathbb{F}_n-F\right|}{\min\{F,1-F\}}\right)\\
&\qquad\qquad \text{since } F(1-F)\geq \min\{F,1-F\}/2,\\
&\leq F(1-F)\cdot\left(1+\frac{4\kappa_\epsilon n^{-1/2}\left(F(1-F)\right)^\gamma}{\min\{F,1-F\}}\right).
\end{aligned}
$$

For a constant $\lambda_\epsilon > 0$ to be specified later, it follows from $\lambda_\epsilon n^{-1/(2-2\gamma)} \leq F \leq 1-\lambda_\epsilon n^{-1/(2-2\gamma)}$ and $\gamma\in[0,1/2)$ that

$$
\frac{(F(1-F))^\gamma}{F} = F^{\gamma-1}(1-F)^\gamma \leq \lambda_\epsilon^{\gamma-1} n^{-(\gamma-1)/(2-2\gamma)} = \lambda_\epsilon^{\gamma-1} n^{1/2}
$$

and

$$
\frac{(F(1-F))^\gamma}{1-F} = F^{\gamma}(1-F)^{\gamma-1} \leq \lambda_\epsilon^{\gamma-1} n^{-(\gamma-1)/(2-2\gamma)} = \lambda_\epsilon^{\gamma-1} n^{1/2}.
$$

Hence

$$
\mathbb{F}_n\left(1-\mathbb{F}_n\right) \leq F(1-F)\cdot\left(1+4\kappa_\epsilon n^{-1/2}\lambda_\epsilon^{\gamma-1}n^{1/2}\right) = F(1-F)(1+4\kappa_\epsilon\lambda_\epsilon^{\gamma-1}).
$$

Thus, on the interval

$$
\{\lambda n^{-1/(2-2\gamma)} \leq \mathbb{F}_n \leq 1-\lambda n^{-1/(2-2\gamma)}\}\cap\{\lambda_\epsilon n^{-1/(2-2\gamma)} \leq F \leq 1-\lambda_\epsilon n^{-1/(2-2\gamma)}\},
$$

$$
\begin{aligned}
G-F &\leq \kappa n^{-1/2}\left(F(1-F)(1+4\kappa_\epsilon\lambda_\epsilon^{\gamma-1})\right)^\gamma + \kappa_\epsilon n^{-1/2}\left(F(1-F)\right)^\gamma\\
&= \nu_\epsilon n^{-1/2}\left(F(1-F)\right)^\gamma,
\end{aligned}
$$

where $\nu_\epsilon = \kappa(1+4\kappa_\epsilon\lambda_\epsilon^{\gamma-1})^\gamma + \kappa_\epsilon$.

The following arguments show that for a large enough λ_ϵ, the interval $\{\lambda_\epsilon n^{-1/(2-2\gamma)} \leq F \leq 1-\lambda_\epsilon n^{-1/(2-2\gamma)}\}$ is a subset of $\{\lambda n^{-1/(2-2\gamma)} \leq \mathbb{F}_n \leq 1-\lambda n^{-1/(2-2\gamma)}\}$. To see this, note that

$$
\begin{aligned}
\mathbb{F}_n &= F+\mathbb{F}_n-F\\
&\geq \left(1-\frac{\left|\mathbb{F}_n-F\right|}{F}\right)F\\
&\geq \left(1-\kappa_\epsilon n^{-1/2}F^{\gamma-1}(1-F)^\gamma\right)F\\
&\geq \left(1-\kappa_\epsilon n^{-1/2}\lambda_\epsilon^{\gamma-1}n^{1/2}\right)\lambda_\epsilon n^{-1/(2-2\gamma)}\\
&= \left(\lambda_\epsilon-\kappa_\epsilon\lambda_\epsilon^\gamma\right)n^{-1/(2-2\gamma)}
\end{aligned}
$$

and analogously,

$$
1-\mathbb{F}_n \geq \left(\lambda_\epsilon-\kappa_\epsilon\lambda_\epsilon^\gamma\right)n^{-1/(2-2\gamma)},
$$

it follows that by choosing a λ_ϵ large enough such that $\lambda_\epsilon - \kappa_\epsilon \lambda_\epsilon^\gamma > \lambda$, the interval $\{\lambda_\epsilon n^{-1/(2-2\gamma)} \le F \le 1 - \lambda_\epsilon n^{-1/(2-2\gamma)}\}$ is a subset of $\{\lambda n^{-1/(2-2\gamma)} \le \mathbb{F}_n \le 1 - \lambda n^{-1/(2-2\gamma)}\}$ and hence on the interval

$$\{\lambda_\epsilon n^{-1/(2-2\gamma)} \le F \le 1 - \lambda_\epsilon n^{-1/(2-2\gamma)}\},$$

$$G - F \quad \le \quad \nu_\epsilon n^{-1/2} \left(F(1-F)\right)^\gamma .$$

Define x_{n1} and x_{n2}, such that $F(x_{n1}) = \lambda_\epsilon n^{-1/(2-2\gamma)}$ and $F(x_{n2}) = 1 - \lambda_\epsilon n^{-1/(2-2\gamma)}$. Analogously, one can prove that $F - G \le \nu_\epsilon n^{-1/2} \left(F(1-F)\right)^\gamma$ on $[x_{n1}, x_{n2}]$ and hence

$$|G - F| \le \nu_\epsilon n^{-1/2} \left(F(1-F)\right)^\gamma \tag{2.20}$$

on $[x_{n1}, x_{n2}]$. Thus for G between L_n^o and U_n^o,

$$\begin{aligned}
\sup_{G: L_n^o \le G \le U_n^o} \left| \int \phi d(G - F) \right| \quad &= \quad \sup_{G: L_n^o \le G \le U_n^o} \left| \int \phi'(x) \left(F(x) - G(x)\right) dx \right| \\
&\le \quad \nu_\epsilon n^{-1/2} \int_{x_{n1}}^{x_{n2}} |\phi'(x)| \, F^\gamma(x) (1 - F(x))^\gamma dx \\
&\qquad + \int_{-\infty}^{x_{n1}} |\phi'(x)| \left(F(x) + U_n^o(x)\right) dx \\
&\qquad + \int_{x_{n2}}^{\infty} |\phi'(x)| \left(2 - F(x) - L_n^o(x)\right) dx.
\end{aligned}$$

From here, we can see that if $F \in \mathcal{P}_{s^*}$ with $s^* > 0$, it follows from remark 1(i) that $J(F)$ is bounded and hence

$$\sup_{G: L_n^o \le G \le U_n^o} \left| \int \phi d(G - F) \right| = O_p(n^{-1/2})$$

as long as ϕ' is bounded on $J(F)$.

The similar argument works if $F \in \mathcal{P}_{s^*}$ with $s^* < 0$ and $J(F)$ is bounded. In the following proof, we get back to our case that $F \in \mathcal{P}_{s^*}$ with $s^* < 0$ and without loss of generality, we assume $J(F) = (-\infty, \infty)$.

As in the proof of Corollary 2.3, for $x_0 \in J(F)$, $b_1 \in (0, T_1(F))$ and $b_2 \in (0, T_2(F))$, there exist points $x_1, x_2 \in J(F)$ with $x_1 < x_0 < x_2$ such that $f(x_1)/F^{1-s^*}(x_1) > b_1$ and $f(x_2)/(1 - F(x_2))^{1-s^*} > b_2$. Then it follows from Theorem 2.2(ii) that with asymptotic probability one,

$$\begin{aligned}
U_n^o(x) \quad &\le \quad \left(U_n^{s^*}(x_1) + s^* b_1 (x - x_1)\right)_+^{1/s^*} \\
&= \quad U_n(x_1) \left(1 + \frac{s^* b_1}{U_n^{s^*}(x_1)} (x - x_1)\right)^{1/s^*} \quad \text{for } x \le x_1,
\end{aligned} \tag{2.21}$$

and

$$1 - L_n^o(x) \quad \le \quad \left((1 - L_n(x_2))^{s^*} - s^* b_2 (x - x_2)\right)_+^{1/s^*}$$

$$= (1-L_n(x_2))\left(1-\frac{s^*b_2}{(1-L_n(x_2))^{s^*}}(x-x_2)\right)^{1/s^*} \quad \text{for } x \geq x_2.$$

Similarly, it follows from Theorem 2.1(ii) that

$$\begin{aligned} F(x) &\leq F(x_1)\left(1+s^*\frac{f(x_1)}{F(x_1)}(x-x_1)\right)_+^{1/s^*} \\ &\leq F(x_1)\left(1+\frac{s^*b_1}{F^{s^*}(x_1)}(x-x_1)\right)^{1/s^*} \quad \text{for } x \leq x_1, \end{aligned} \tag{2.22}$$

and

$$\begin{aligned} 1-F(x) &\leq (1-F(x_2))\left(1-s^*\frac{f(x_2)}{1-F(x_2)}(x-x_2)\right)_+^{1/s^*} \\ &\leq (1-F(x_2))\left(1-\frac{s^*b_2}{(1-F(x_2))^{s^*}}(x-x_2)\right)^{1/s^*} \quad \text{for } x \geq x_2. \end{aligned}$$

For large enough n, one can have $[x_1, x_2] \subset [x_{n1}, x_{n2}]$ and hence

$$\sup_{G: L_n^o \leq G \leq U_n^o} \left| \int \phi d(G-F) \right| \leq I_{n0} + I_{n1} + I'_{n1} + I_{n2} + I'_{n2},$$

where

$$\begin{aligned} I_{n0} &:= \nu_\epsilon n^{-1/2} \int_{x_1}^{x_2} |\phi'(x)|\, F^\gamma(x)(1-F(x))^\gamma dx, \\ I_{n1} &:= \nu_\epsilon n^{-1/2} \int_{x_n 1}^{x_1} |\phi'(x)|\, F^\gamma(x)(1-F(x))^\gamma dx, \\ I_{n2} &:= \nu_\epsilon n^{-1/2} \int_{x_2}^{x_{n2}} |\phi'(x)|\, F^\gamma(x)(1-F(x))^\gamma dx, \\ I'_{n1} &:= \int_{-\infty}^{x_{n1}} |\phi'(x)|\, (F(x) + U_n^o(x)) dx, \\ I'_{n2} &:= \int_{x_{n2}}^{\infty} |\phi'(x)|\, (2 - F(x) - L_n^o(x)) dx. \end{aligned}$$

Note that $I_{n0} \leq \nu_\epsilon n^{-1/2} \int_{x_1}^{x_2} |\phi'(x)|\, dx = O(n^{-1/2})$. For the other terms, first note that $F(x_{n1}) = \lambda_\epsilon n^{-1/(2-2\gamma)}$ and hence it follows from (2.22) that

$$x_{n1} \geq x_1 - \frac{F^{s^*}(x_1)}{s^*b_1} + \frac{\lambda_\epsilon^{s^*}}{s^*b_1} n^{-s^*/(2-2\gamma)} = O(1) + \frac{\lambda_\epsilon^{s^*}}{s^*b_1} n^{-s^*/(2-2\gamma)}.$$

Analogously, one can prove that

$$x_{n2} \leq x_2 - \frac{(1-F(x_2))^{s^*}}{s^*b_2} - \frac{\lambda_\epsilon^{s^*}}{s^*b_2} n^{-s^*/(2-2\gamma)} = O(1) + \frac{\lambda_\epsilon^{s^*}}{s^*b_1} n^{-s^*/(2-2\gamma)}.$$

Thus, it follows from (2.22) and the upper bound of $|\phi'|$ that

$$I_{n1} \leq \nu_\epsilon n^{-1/2} \int_{x_n 1}^{x_1} |\phi'(x)|\, F^\gamma(x) dx$$

$$
\begin{aligned}
&\leq O\left(n^{-1/2}\int_{x_{n1}}^{x_1}|\phi'(x)|\left(1+\frac{s^*b_1}{F^{s^*}(x_1)}(x-x_1)\right)^{\gamma/s^*}dx\right)\\
&\leq O\left(n^{-1/2}\int_{0}^{O\left(n^{-s^*/(2-2\gamma)}\right)}|\phi'(x)|\left(1+\frac{-s^*b_1}{F^{s^*}(x_1)}x\right)^{\gamma/s^*}dx\right)\\
&= O\left(n^{-1/2}\int_{0}^{O\left(n^{-s^*/(2-2\gamma)}\right)}|\phi'(x)|\,x^{\gamma/s^*}dx\right)\\
&\leq O\left(n^{-1/2}\int_{0}^{O\left(n^{-s^*/(2-2\gamma)}\right)}x^{k-1}x^{\gamma/s^*}dx\right)\\
&= O\left(n^{-1/2}n^{-(k+\gamma/s^*)s^*/(2-2\gamma)}\right)\\
&= O\left(n^{-\frac{1}{2}\left(\frac{1+s^*k}{1-\gamma}\right)}\right).
\end{aligned}
$$

Analogously, one could show that

$$
I_{n2}\leq O\left(n^{-\frac{1}{2}\left(\frac{1+s^*k}{1-\gamma}\right)}\right).
$$

To bound I'_{n1}, note that for $x\leq x_{n1}$, it follows from an analogous proof of (2.22) that

$$
F(x)\leq\left(F^{s^*}(x_{n1})+s^*b_1(x-x_{n1})\right)^{1/s^*}=\left(\lambda_\epsilon^{s^*}n^{-s^*/(2-2\gamma)}+s^*b_1(x-x_{n1})\right)^{1/s^*}.
$$

Analogously, it follows that for $x\leq x_{n1}$,

$$
U_n^o(x)\leq\left(U_n^{s^*}(x_{n1})+s^*b_1(x-x_{n1})\right)^{1/s^*}.
$$

Note that it follows from (2.20) that

$$
\begin{aligned}
U_n(x_{n1}) &= U_n(x_{n1})-F(x_{n1})+F(x_{n1})\\
&\leq \nu_\epsilon n^{-1/2}\left(F(x_{n1})(1-F(x_{n1}))\right)^\gamma+F(x_{n1})\\
&\leq \nu_\epsilon n^{-1/2}F^\gamma(x_{n1})+F(x_{n1})\\
&= (\nu_\epsilon\lambda_\epsilon^\gamma+\lambda_\epsilon)n^{-1/(2-2\gamma)}
\end{aligned}
$$

and hence for $x\leq x_{n1}$,

$$
U_n^o(x) \leq \left((\nu_\epsilon\lambda_\epsilon^\gamma+\lambda_\epsilon)^{s^*}n^{-s^*/(2-2\gamma)}+s^*b_1(x-x_{n1})\right)^{1/s^*}.
$$

Thus,

$$
\begin{aligned}
I'_{n1} &= \int_{-\infty}^{x_{n1}}|\phi'(x)|\left(F(x)+U_n^o(x)\right)dx\\
&= O\left(\int_{-\infty}^{x_{n1}}|\phi'(x)|\left(n^{-s^*/(2-2\gamma)}+s^*b_1(x-x_{n1})\right)^{1/s^*}dx\right)
\end{aligned}
$$

$$
\begin{aligned}
&= O\left(\int_{-\infty}^{0} |\phi'(x+x_{n1})| \left(n^{-s^*/(2-2\gamma)} + s^* b_1 x\right)^{1/s^*} dx\right) \\
&= O\left(\int_{-\infty}^{0} |x+x_{n1}|^{k-1} \left(n^{-s^*/(2-2\gamma)} + s^* b_1 x\right)^{1/s^*} dx\right) \\
&= O\left(n^{-1/(2-2\gamma)} \int_{-\infty}^{0} |x+x_{n1}|^{k-1} \left(1 + s^* b_1 x / n^{-s^*/(2-2\gamma)}\right)^{1/s^*} dx\right) \\
&= O\left(n^{-1/(2-2\gamma)} n^{-s^*/(2-2\gamma)}\right. \\
&\qquad \left. \cdot \int_{-\infty}^{0} \left|x n^{-s^*/(2-2\gamma)} + x_{n1}\right|^{k-1} (1 + s^* b_1 x)^{1/s^*} dx\right) \\
&= O\left(n^{-1/(2-2\gamma)} n^{-s^*/(2-2\gamma)}\right. \\
&\qquad \left. \cdot \int_{-\infty}^{0} \left|x n^{-s^*/(2-2\gamma)} + n^{-s^*/(2-2\gamma)}\right|^{k-1} (1 + s^* b_1 x)^{1/s^*} dx\right) \\
&= O\left(n^{-1/(2-2\gamma)} n^{-ks^*/(2-2\gamma)} \int_{-\infty}^{0} |x|^{k-1} |x|^{1/s^*} dx\right) \\
&= O\left(n^{-(ks^*+1)/(2-2\gamma)}\right).
\end{aligned}
$$

Analogously, one could show that

$$
I'_{n2} \leq O\left(n^{-(ks^*+1)/(2-2\gamma)}\right).
$$

Hence

$$
\begin{aligned}
\sup_{G: L_n^o \leq G \leq U_n^o} \left|\int \phi d(G-F)\right| &\leq I_{n0} + I_{n1} + I'_{n1} + I_{n2} + I'_{n2} \\
&\leq O(n^{-1/2}) + O\left(n^{-(ks^*+1)/(2-2\gamma)}\right).
\end{aligned}
$$

□

3 Fisher-Pitman permutation tests based on nonparametric Poisson mixtures with application to single cell genomics

3.1 Introduction

Considering an experiment with multiple samples drawn from multiple populations, distinguishing possible difference among them in one or more dimensions is a fundamental statistical task. In the classical test of the null hypothesis of no mean differences, one-way analysis of variance (ANOVA, cf. Fisher (1925)) F-test is perhaps the most commonly used tool, and is the uniformly most powerful invariant one under additional normal assumption, c.f. Scheffé (1959, page 50).

Despite its popularity, one-way ANOVA has its competing alternatives. In the context of randomized experiments, Fisher (Fisher, 1935) initialized an ingenious permutation approach as an alternative to performing ANOVA F-test. This idea was later developed further by Pitman (Pitman, 1938). The resulting procedures, often termed the Fisher-Pitman permutation tests in literature, achieve the appealing property of being exactly distribution-free and have been suggested in various contexts as, e.g., when the distributional assumptions of F-tests no longer hold (Marascuilo and McSweeney, 1977; Still and White, 1981; Berry and Mielke, 1983). Robustness properties have been further studied empirically (Boik, 1987) and theoretically (Chung and Romano, 2013); power analyses were also performed in Hoeffding (1952) and Robinson (1973).

Although being originally defined in Euclidean spaces, it is by now well understood that the ANOVA F-tests and especially their permutation-type alternatives are able to adapt to an arbitrary metric space. This is via the approach of "interpoint" distance functions (Mielke Jr et al., 1976; Mielke Jr, 1984) that uses an alternative representation of the F statistic as a function of between- and within-group pairwise distances. Thus, through replacing the original Euclidean distance by any properly defined distance function, the idea of Fisher-Pitman permutation tests is now implementable in many complicated metric spaces beyond the Euclidean (Anderson, 2001; Mielke and Berry, 2007; Petersen and Müller, 2019).

Our study of Fisher-Pitman-type permutation tests stems from the analysis of single-cell RNA-seq (scRNA-seq) data, and particularly, a framework that was recently promoted in Sarkar and Stephens (2021). There, the authors described how a separation of measurement and expression models is able to clarify confusion in modeling scRNA-seq data, and accordingly advocated using the terminology of Poisson mixtures to unify many existing models (cf. Table 1 in Sarkar and Stephens (2021)). In detail, thinking about $X_{ij}^{(k)}$ to be the absolute expression of a specific gene in cell $i \in [N_{jk}] := \{1, 2, \ldots, N_{jk}\}$ of subject $j \in [n_k]$ of population $k \in [K]$, we are interested in studying the following model of $X_{ij}^{(k)}$ that is a slight simplification

to Sarkar and Stephens's Equation (1):

$$X_{ij}^{(k)} \mid \lambda_{ij}^{(k)} \sim \text{Poisson}\big(r_{ij}^{(k)}\lambda_{ij}^{(k)}\big); \qquad \text{(measurement model)} \tag{3.1}$$

$$\lambda_{ij}^{(k)} \sim Q_j^{(k)}. \qquad \text{(expression model)} \tag{3.2}$$

Here $r_{ij}^{(k)} > 0$ adjusts the cell "read depth" (cf. Zhang et al. (2020, Page 1)), often set to be the scaled number of the cell's total reads across all genes (Sarkar and Stephens, 2021, Equation (1)), and in this chapter is assumed to be known; $Q_j^{(k)}$ is a properly defined distribution that describes the "expression level" of the gene in population k and is assumed to have a compact support on the nonnegative real line. Adopting the statistical terminology, for each $k \in [K]$ and $j \in [n_k]$, $\{X_{ij}^{(k)}, i = 1, \ldots, N_{jk}\}$ then independently follow Poisson mixture distributions of point mass functions (PMFs)

$$h_{ij}^{(k)}(x) := \int_0^\infty e^{-\lambda r_{ij}^{(k)}} \frac{\{\lambda r_{ij}^{(k)}\}^x}{x!} \mathrm{d}Q_j^{(k)}(\lambda), \quad x = 0, 1, 2, \ldots$$

and a mixing distribution $Q_j^{(k)}$ that has to be characterized by a nonparametric model; see Sarkar and Stephens (2021, Section "Modeling scRNA-seq data") for a discussion of why a nonparametric model of $Q_j^{(k)}$ is preferred in single-cell genomics, though Sarkar and Stephens (2021) did not employ such Poisson mixtures for individual level differential expression testing.

Based on the observations $\big\{X_{ij}^{(k)}, i \in [N_{jk}], j \in [n_k], k \in [K]\big\}$ as well as the measure/expression models (3.1)-(3.2), a natural question to ask is whether there exists any population-level gene expression difference among the K groups. For this, we propose to leverage a Fisher-Pitman-type permutation test based on consistent estimators $\big\{\tilde{Q}_j^{(k)}, j \in [n_k], k \in [K]\big\}$ of the mixing distributions $\big\{Q_j^{(k)}, j \in [n_k], k \in [K]\big\}$ under Wasserstein metrics, which have received much attention in recent mixture distribution estimation literature (see, among many others, Nguyen (2013), Tian et al. (2017), Vinayak et al. (2019), Wu and Yang (2020a), and the references therein). Particularly appealing choices to us include the NPMLE $\hat{Q}_j^{(k)}$ and its Poisson-smoothed one $h_{\hat{Q}_j^{(k)}}$ (notation to be introduced by the end of this chapter); see Chapter 3.2 ahead for the detailed description of the testing procedure.

Many methods have been developed for differential expression analysis of scRNA-seq data (Chen et al., 2019). However, their focus is differential expression between two groups of cells instead of two groups of individuals. For individual level testing, a standard approach is to add up gene expression across all the cells (of a particular cell type) of an individual to create a pseudo-bulk sample, and then apply the methods for differential expression analysis using bulk RNA-seq data, such as DESeq2 (Love et al., 2014). The novelty of our proposed procedure is that we assess differential expression across individuals using cell level data instead of pseudo-bulk data. Furthermore, the proposed tests are shown to be consistent

against their ANOVA-type alternatives, i.e., they are able to asymptotically distinguish the null from any fixed alternative where the "between-group" variation is larger than the "within-group" variation, a result that sheds insight to the power of the developed tests and is in line with classic observations (Hoeffding, 1952; Robinson, 1973)[1].

As a byproduct of our theoretical study, this chapter further justifies the use of NPMLEs via establishing their rate-optimality in estimating the Poisson mixing distribution under the Wasserstein-1 (W_1) metric. Although the consistency of the NPMLEs has been established in the literature for different nonparametric mixture models (cf. Simar (1976) for nonparametric Poisson mixtures; and Chen (2017) and the references therein for more general models), NPMLEs' rates of convergence and their matching to a minimax lower bound are long standing until very recently. Built on the breakthroughs in binomial (Tian et al., 2017; Vinayak et al., 2019) and Gaussian mixtures (Wu and Yang, 2020a) (see also Jiang and Zhang (2019) for a related study on the nonparametric likelihood ratio test) as well as the new analytical techniques devised in Jiao et al. (2015), Wu and Yang (2016), Jiao et al. (2018), and Han and Shiragur (2021), we are now able to further the optimality of NPMLEs to the nonparametric Poisson mixtures under minimal assumptions on the true mixing distribution function. These results yield additional theoretical support for the use of NPMLEs in our developed tests.

The rest of this chapter is organized as follows. Chapter 3.2 describes the model setup and studies the size and power of the proposed permutation tests. Chapter 3.3 discusses implementation of the developed test. The finite-sample performance of the developed (smoothed or not) NPMLE-based permutation tests is investigated in Chapter 3.4. Chapter 3.5 applies the studied tests to a real scRNA-seq data containing single brain nuclei from autism subjects and healthy controls (Velmeshev et al., 2019) and discover significantly differentially expressed genes that cannot be detected using the benchmark DESeq2 method applied on pseudo-bulk data (Love et al., 2014). In Chapter 3.6, we justify the use of NPMLEs in the permutation tests outlined in Chapter 3.2 by providing minimax optimality results for the NPMLE for nonparametric mixture of Poissons. All the proofs are relegated to a supplement.

Notation. For any two distributions P, Q on the real line, the Wasserstein-1 distance is defined to be $W_1(P,Q) := \sup_{\ell \in \mathrm{Lip}_1} \int \ell(\mathrm{d}P - \mathrm{d}Q)$, where Lip_1 represents all 1-Lipschitz functions. For any distribution Q on the nonnegative real line, we define its Poisson smoothed version as

$$h_Q(x) := \int_0^\infty e^{-\lambda} \frac{\lambda^x}{x!} \mathrm{d}Q(\lambda), \ x = 0, 1, 2, ...$$

For any two constants a, b, we denote $a \vee b := \max\{a, b\}$ and $a \wedge b := \min\{a, b\}$.

[1] In addition to developing a more flexible non-parametric model, another route to boost the power of differential expression analysis is to de-noise the scRNA-seq data; see Zhang et al. (2022) for a proposal along that track.

3.2 Permutation tests

3.2.1 Setup

Throughout this chapter, it is assumed that the observations are heterogeneous count data $\{X_{ij}^{(k)}, i \in [N_{jk}], j \in [n_k], k \in [K]\}$ with $N_{jk} = N_{jk,n} \to \infty$ and $n_k = n_{k,n} \to \infty$ as $n := \sum n_k \to \infty$. In contrast, $K \geq 2$ is assumed to be a fixed integer. It is further assumed that the probability measures $Q_j^{(k)}$'s in (3.1) have a common support $[0, B]$ for some $B > 0$ that is known a priori (cf. appendix Chapter 3.8 for a real implementation) and kept to be fixed in this chapter; later in Chapter 3.6 we will explore a more general setting where $B = B_n$ is allowed to increase with n.

To facilitate the approach to distinguishing differences among the K groups, in addition to the measurement model (3.1) and the expression model (3.2), a third-layer "population model" is introduced to encourage independent and identically distributed (i.i.d.) randomness among each n_k within-group expression models:

$$\text{for each } k \in [K]: \quad Q_1^{(k)}, \ldots, Q_{n_k}^{(k)} \overset{i.i.d.}{\sim} \mathcal{Q}_k. \quad \text{(population model)} \tag{3.3}$$

Here $\mathcal{Q}_k$ is understood to be a probability measure over the Prohorov-metric topology of the space of probability measures that are defined on the Borel σ-field of $[0, B]$; details about constructing Prohorov-metric topology are referred to Pages 72-73 in Billingsley (1999). Following the discussions in Sarkar and Stephens (2021, Section "Modeling scRNA-seq data"), we do not specify $\mathcal{Q}_k$ except for assuming boundedness and well-definedness.

To wrap up, the model considered in this manuscript, summarizing the three layers ((3.1), (3.2), (3.3)), is:

$$\begin{aligned} &\left\{X_{ij}^{(k)}, i \in [N_{jk}], j \in [n_k], k \in [K]\right\} \text{ are independently distributed with PMFs} \\ &\int \Big[\int_0^B e^{-\lambda r_{ij}^{(k)}} \frac{\{\lambda r_{ij}^{(k)}\}^x}{x!} \mathrm{d}Q(\lambda) \Big] \mathrm{d}\mathcal{Q}_k(Q), \quad x = 0, 1, 2, \ldots. \end{aligned} \tag{3.4}$$

Under the above model, it is understood that $\mathcal{Q}_1, \ldots, \mathcal{Q}_K$ and $K \geq 2$ are fixed, all of which won't change with n. Besides $\mathcal{Q}_1, \ldots, \mathcal{Q}_K$ and accordingly the random measures $Q_j^{(k)}$'s, the observations $X_{ij}^{(k)}$'s also depend on the read depths $r_{ij}^{(k)} = r_{ij,n}^{(k)}$'s that are allowed to change with n. We are hence faced with a triangular array of possibly highly heterogeneous observations.

3.2.2 Tests

Under Model (3.4), we are interested in testing the following null hypothesis,

$$H_0 : \mathcal{Q}_1 = \mathcal{Q}_2 = \cdots = \mathcal{Q}_K, \tag{3.5}$$

and aim to detect any population-level difference between groups. Note that here, due to the incorporation of read depths $r_{ij}^{(k)}$'s, the measurements themselves even within each group are generally not identically distributed; thus, a naive empirical distribution function based test could be substantially biased.

The main interest of this chapter is to explore how robust a Fisher-Pitman-type test can be when each unobserved subject-level random measure $Q_j^{(k)}$ is replaced by a plug-in-type estimate $\tilde{Q}_j^{(k)}$ and its Poisson-smoothed version $h_{\tilde{Q}_j^{(k)}}$ calculated from the measurements $X_{1j}^{(k)}, \ldots, X_{N_{jkj}}^{(k)}$. To this end, let's regulate $\tilde{Q}_j^{(k)}$ as follows.

Definition 3.1. *For any $j \in [n_k]$ and any $k \in [K]$, an estimator $\tilde{Q}_j^{(k)}$ of $Q_j^{(k)}$ is said to be subject-specific conditionally W_1-consistent (shorthanded as "conditionally W_1-consistent") if it is (i) a function of $X_{1j}^{(k)}, \ldots, X_{N_{jkj}}^{(k)}$; (ii) of support $[0, B]$; and (iii) satisfying*

$$E\Big\{W_1\Big(\tilde{Q}_j^{(k)}, Q_j^{(k)}\Big) \;\Big|\; Q_j^{(k)}\Big\} \to 0 \text{ as } N_{jk} = N_{jk,n} \to \infty \tag{3.6}$$

for almost all $Q_j^{(k)}$ with regard to the measure $\mathcal{Q}_k$.

We next consider the Poisson-smoothed mixing distribution estimator

$$h_{\tilde{Q}_j^{(k)}} := \int_0^\infty e^{-\lambda} \frac{\lambda^x}{x!} \mathrm{d}\tilde{Q}_j^{(k)}(\lambda)$$

based on any conditionally W_1-consistent estimator $\tilde{Q}_j^{(k)}$. It justifies the use of smoothed NPMLEs as an alternative to $\tilde{Q}_j^{(k)}$. Note that in the classic setting when read depths are all forced to be equal, Proposition 3.1 in Lambert and Tierney (1984) showed that some $h_{\tilde{Q}_j^{(k)}}$ can approximate $h_{Q_j^{(k)}}$ polynomially fast. Accordingly, although the differences between $Q_j^{(k)}$'s can be larger than those between $h_{Q_j^{(k)}}$'s, the differences between $\tilde{Q}_j^{(k)}$'s can be smaller than those between $h_{\tilde{Q}_j^{(k)}}$'s. See also Han et al. (2021) for some similar observations.

Theorem 3.1. *Suppose $\tilde{Q}_j^{(k)}$ is conditionally W_1-consistent. Then*

$$E\Big\{W_1\Big(h_{\tilde{Q}_j^{(k)}}, h_{Q_j^{(k)}}\Big) \;\Big|\; Q_j^{(k)}\Big\} \to 0 \text{ as } N_{jk} = N_{jk,n} \to \infty$$

for almost all $Q_j^{(k)}$ with regard to the measure $\mathcal{Q}_k$.

A particularly appealing candidate estimator of the mixing distribution is the following NPMLE $\hat{Q}_j^{(k)}$ with

read depth incorporated:

$$\hat{Q}_j^{(k)} \in \underset{Q \text{ of support } [0,B]}{\operatorname{argmax}} \sum_{i \in [N_{jk}]} \log \int_0^\infty e^{-\lambda r_{ij}^{(k)}} \frac{\{\lambda r_{ij}^{(k)}\}^{X_{ij}^{(k)}}}{X_{ij}^{(k)}!} \mathrm{d}Q(\lambda). \tag{3.7}$$

Note that here $\hat{Q}_j^{(k)}$ may not be unique due to read depths, and if there are multiple choices, pick any one of them (cf. Remark 3.4). We shall discuss the calculation of $\hat{Q}_j^{(k)}$ in Chapter 3.3. The next theorem shows that NPMLEs are conditionally W_1-consistent under no further assumptions on the population measures $\mathcal{Q}_k$'s except for the already imposed bounded support one.

Theorem 3.2 (Conditionally W_1-consistency of NPMLEs)**.** *Assume $N_{jk} = N_{jk,n} \to \infty$ as $n \to \infty$, $r_{ij}^{(k)} = r_{ij,n}^{(k)} \in [\gamma_0, \gamma_1]$ are uniformly upper and lower bounded by two positive universal constants γ_0, γ_1, and $\mathcal{Q}_k$'s have a common fixed support $[0, B]$. We then have the NPMLEs $\hat{Q}_j^{(k)}$'s are all conditionally W_1-consistent.*

Remark 3.1. In the literature, consistency of NPMLEs of mixing distributions under the classical i.i.d. mixture distribution setup (corresponding to the case with all read depths identical to each other) has been studied in depth. Notable results include Kiefer and Wolfowitz (1956), Simar (1976), Pfanzagl (1988); note also the survey by Chen (Chen, 2017). However, although arising naturally from single-cell genomics modeling, read-depth-incorporated nonparametric mixture distributions have not received much attention in mathematical statistics and, to our knowledge, Theorem 3.2 delivers the first consistency result for NPMLEs under this heterogeneous setting.

Based on any conditionally W_1-consistent estimators $\{\tilde{Q}_j^{(k)}\}$ of $\{Q_j^{(k)}\}$ and their Poisson-smoothed versions $h_{\tilde{Q}_j^{(k)}}$'s, the proposed ANOVA-type (pseudo-F) test statistics are

$$\tilde{F} := \frac{\frac{1}{n} \sum_{k_1,k_2 \in [K]} \sum_{j_1 \in [n_{k_1}], j_2 \in [n_{k_2}]} W_1\left(\tilde{Q}_{j_1}^{(k_1)}, \tilde{Q}_{j_2}^{(k_2)}\right)^2 - \sum_{k \in [K]} \frac{1}{n_k} \sum_{j_1,j_2 \in [n_k]} W_1\left(\tilde{Q}_{j_1}^{(k)}, \tilde{Q}_{j_2}^{(k)}\right)^2}{\sum_{k \in [K]} \frac{1}{n_k} \sum_{j_1,j_2 \in [n_k]} W_1\left(\tilde{Q}_{j_1}^{(k)}, \tilde{Q}_{j_2}^{(k)}\right)^2}$$

and

$$\tilde{F}_h := \frac{\frac{1}{n} \sum_{k_1,k_2 \in [K]} \sum_{j_1 \in [n_{k_1}], j_2 \in [n_{k_2}]} W_1\left(h_{\tilde{Q}_{j_1}^{(k_1)}}, h_{\tilde{Q}_{j_2}^{(k_2)}}\right)^2 - \sum_{k \in [K]} \frac{1}{n_k} \sum_{j_1,j_2 \in [n_k]} W_1\left(h_{\tilde{Q}_{j_1}^{(k)}}, h_{\tilde{Q}_{j_2}^{(k)}}\right)^2}{\sum_{k \in [K]} \frac{1}{n_k} \sum_{j_1,j_2 \in [n_k]} W_1\left(h_{\tilde{Q}_{j_1}^{(k)}}, h_{\tilde{Q}_{j_2}^{(k)}}\right)^2}.$$

It is ready to check that these two test statistics both reduce to the original one-way ANOVA statistic if the examined space is the real space equipped with the Euclidean norm. The studied statistics then generalize the one-way ANOVA statistics to the W_1-metric measure space with different inputs (mixing distribution smoothed or not); similar generalizations have been made in various other (non-) Euclidean spaces (Anderson, 2001; Mielke and Berry, 2007; Petersen and Müller, 2019).

We then move on to introduce the corresponding permuted ANOVA-type test statistics. To this end, for each permutation $\pi : [n] \to [n]$, let $\Pi^{j,k} = (\Pi_1^{j,k}, \Pi_2^{j,k}) := \pi^{\uparrow}(j,k)$ represent the original subject and population indices corresponding to "the j-th subject in the k-th group" after permutation π. The permuted test statistics are

$$\tilde{F}^{\pi} := \frac{\frac{1}{n}\sum_{k_1,k_2\in[K]}\sum_{j_1\in[n_{k_1}],j_2\in[n_{k_2}]} W_1\left(\tilde{Q}_{j_1}^{(k_1)}, \tilde{Q}_{j_2}^{(k_2)}\right)^2 - \sum_{k\in[K]}\frac{1}{n_k}\sum_{j_1,j_2\in[n_k]} W_1\left(\tilde{Q}_{\Pi_1^{j_1,k}}^{(\Pi_2^{j_1,k})}, \tilde{Q}_{\Pi_1^{j_2,k}}^{(\Pi_2^{j_2,k})}\right)^2}{\sum_{k\in[K]}\frac{1}{n_k}\sum_{j_1,j_2\in[n_k]} W_1\left(\tilde{Q}_{\Pi_1^{j_1,k}}^{(\Pi_2^{j_1,k})}, \tilde{Q}_{\Pi_1^{j_2,k}}^{(\Pi_2^{j_2,k})}\right)^2}$$

and

$$\tilde{F}_h^{\pi} := \frac{\frac{1}{n}\sum_{k_1,k_2\in[K]}\sum_{j_1\in[n_{k_1}],j_2\in[n_{k_2}]} W_1\left(h_{\tilde{Q}_{j_1}^{(k_1)}}, h_{\tilde{Q}_{j_2}^{(k_2)}}\right)^2 - \sum_{k\in[K]}\frac{1}{n_k}\sum_{j_1,j_2\in[n_k]} W_1\left(h_{\tilde{Q}_{\Pi_1^{j_1,k}}^{(\Pi_2^{j_1,k})}}, h_{\tilde{Q}_{\Pi_1^{j_2,k}}^{(\Pi_2^{j_2,k})}}\right)^2}{\sum_{k\in[K]}\frac{1}{n_k}\sum_{j_1,j_2\in[n_k]} W_1\left(h_{\tilde{Q}_{\Pi_1^{j_1,k}}^{(\Pi_2^{j_1,k})}}, h_{\tilde{Q}_{\Pi_1^{j_2,k}}^{(\Pi_2^{j_2,k})}}\right)^2}.$$

The following are the Fisher-Pitman-type permutation tests with nominal level α:

$$\tilde{T}_{\alpha} := \begin{cases} 1, & \text{if } P(\tilde{F}^{\pi} < \tilde{F} \mid \tilde{Q}_j^{(k)}\text{'s}) \geq 1-\alpha, \\ 0, & \text{otherwise,} \end{cases}$$

and

$$\tilde{T}_{h,\alpha} := \begin{cases} 1, & \text{if } P(\tilde{F}_h^{\pi} < \tilde{F}_h \mid \tilde{Q}_j^{(k)}\text{'s}) \geq 1-\alpha, \\ 0, & \text{otherwise,} \end{cases}$$

where the probability here is only with respect to the random permutation π.

As the (Poisson smoothed-)NPMLEs are chosen, the corresponding tests $\tilde{T}_{\alpha}$ and $\tilde{T}_{h,\alpha}$ are specified as $\hat{T}_{\alpha}$ and $\hat{T}_{h,\alpha}$.

3.2.3 Theory

This subchapter provides the necessary theoretical support on the presented tests $\tilde{F}^{\pi}$ and $\tilde{F}_h^{\pi}$. Particular focus is on the asymptotic size and consistency against Robinson-type ANOVA alternatives (cf. Theorem 3 in Robinson (1973)). To minimize assumptions and for presentation clearness, we are focused on the following balanced design case:

Assumption 3.1. *The design is balanced so that $n_k = n/K$ and $N_{jk} = N$ for $j \in [n_k]$, $k \in [K]$. In addition, it is assumed that the sets $\{r_{ij}^{(k)}, i \in [N]\}$ are invariant with respect to both j and k.*

Remark 3.2. We note that Assumption 3.1 can be weakened in a straightforward manner to allow for $n_k/n \to 1/K$, N_{jk}'s asymptotically comparable, and the sets $\{r_{ij}^{(k)}, i \in [N_{jk}]\}$ all weakly converge to a

same probability measure that does not depend on the particular choice of j and k (see Shi et al. (2020, Proposition 2.2) as well as Deb and Sen (2021) for a similar setup in the recent independence testing literature). We however do not pursue these tracks but rather leave them to the readers of interest to verify.

Our first result concerns with the sizes of proposed tests, is still valid with a finite sample size, and is a direct consequence of a long line of literature on permutation-based tests.

Theorem 3.3 (Size validity)**.** *We have, for any finite N and n, as long as H_0 in (3.5) and Assumption 3.1 hold,*

$$P(\tilde{T}_\alpha = 1|H_0) \leq \alpha \quad \text{and} \quad P(\tilde{T}_{h,\alpha} = 1|H_0) \leq \alpha.$$

In the following, we are focused on asymptotic results with the balanced design and let $N = N_n \to \infty$ as $n \to \infty$. The next theorem is the main result of this subchapter.

Theorem 3.4 (Test consistency)**.** *Consider $\tilde{Q}_j^{(k)}$'s to be conditionally W_1-consistent estimators of $Q_j^{(k)}$'s. If Assumption 3.1 holds, then the following two statements are true.*

(a) Under any fixed alternative regarding $\mathcal{Q}_1, \ldots, \mathcal{Q}_K$ such that

$$H_1 : \frac{1}{K} \sum_{k\in[K]} E\Big\{W_1\Big(Q_1^{(k)}, Q_2^{(k)}\Big)^2\Big\} < \sum_{k_1\neq k_2\in[K]} \frac{E\{W_1(Q_1^{(k_1)}, Q_1^{(k_2)})^2\}}{K(K-1)}, \tag{3.8}$$

we have $\lim_{n\to\infty} P(\tilde{T}_\alpha = 1|H_1) = 1$ for each $\alpha \in (0,1)$.

(b) Under any fixed alternative regarding $\mathcal{Q}_1, \ldots, \mathcal{Q}_K$ such that

$$H_{1,h} : \frac{1}{K} \sum_{k\in[K]} E\Big\{W_1\Big(h_{Q_1^{(k)}}, h_{Q_2^{(k)}}\Big)^2\Big\} < \sum_{k_1\neq k_2\in[K]} \frac{E\Big\{W_1\Big(h_{Q_1^{(k_1)}}, h_{Q_1^{(k_2)}}\Big)^2\Big\}}{K(K-1)}, \tag{3.9}$$

we have $\lim_{n\to\infty} P(\tilde{T}_{h,\alpha} = 1|H_{1,h}) = 1$ for each $\alpha \in (0,1)$.

Remark 3.3. Assumption (3.8) states that the average of the Wasserstein distances within groups is less than the average of the Wasserstein distances between groups. If then, our theory suggested that one is able to test the differences between groups using the permuted ANOVA-type test statistics. Assumption (3.9) is analogous, while since the test statistics is based on the smoothed NPMLEs, assumptions have to be made on $(h_{Q_1^{(k)}}, h_{Q_2^{(k)}})$'s.

Specific to (smoothed-)NPMLEs, the following theorem is a direct consequence of Theorems 3.2-3.4.

Corollary 3.1. *Suppose Assumption 3.1 and all conditions in Theorem 3.2 hold. Then the following are true for any $\alpha \in (0,1)$.*

(a) For any finite N and n, as long as H_0 in (3.5) holds, we have

$$P(\hat{T}_\alpha = 1|H_0) \le \alpha \quad \text{and} \quad P(\hat{T}_{h,\alpha} = 1|H_0) \le \alpha.$$

(b) Concerning any fixed alternative H_1 (or H_{1h}), we have

$$\lim_{n\to\infty} P(\hat{T}_\alpha = 1 \mid H_1) = 1 \quad \text{and} \quad \lim_{n\to\infty} P(\hat{T}_{h,\alpha} = 1 \mid H_{1,h}) = 1.$$

3.3 Algorithms

This chapter presents three algorithms to calculate (3.7),

(1) the vertex direction method (VDM), cf. Fedorov (1972), Simar (1976), Wu (1978a), Wu (1978b), Böhning (1982), and Lindsay (1983a);

(2) the vertex exchange method (VEM), cf. Böhning (1985) and Böhning (1986);

(3) the intra simplex direction method (ISDM), cf. Lesperance and Kalbfleisch (1992).

To simplify the notation, in this chapter we remove j, k from the subscript and use $\{X_i, i \in [N]\}$ and $\{r_i, i \in [N]\}$ to denote the sample points and the corresponding read-depths. Moreover, we use $\hat{Q}$ to denote the NPMLE defined in (3.7) based on $\{X_i, i \in [N]$ and $\{r_i, i \in [N]\}$. For a discrete measure G on $[0, B]$ with support points $\{\lambda_m, m \in [M]\}$, let $G(\lambda_m)$ stand for the mass G assigned at λ_m for each $m \in [M]$. We define

$$\Phi(G) := \frac{1}{N} \sum_{i\in[N]} \log \left(\sum_{m\in[M]} G(\lambda_m) e^{-\lambda_m r_i} (\lambda_m r_i)^{X_i} \right)$$

and its directional derivative from G to δ_λ as

$$\Phi'(G, \delta_\lambda) := \lim_{\epsilon\to 0^+} \epsilon^{-1} \Big\{ \Phi\{(1-\epsilon)G \oplus \epsilon\delta_\lambda\} - \Phi(G) \Big\} = \frac{1}{N} \sum_{i\in[N]} \frac{e^{-\lambda r_i}(\lambda r_i)^{X_i}}{\sum_{m\in[M]} G(\lambda_m) e^{-\lambda_m r_i} (\lambda_m r_i)^{X_i}} - 1.$$

Here δ_λ represents the unit measure at $\lambda \in [0, B]$. Lastly, for any two signed measures ν_1 and ν_2 on the real line, we denote $\nu_1 \oplus \nu_2$ as the sum of ν_1 and ν_2, and $\nu_1 \ominus \nu_2$ as the sum of ν_1 and $-\nu_2$.

With these notation, we are now ready to present the VDM, VEM, and ISDM algorithms for calculating $\hat{Q}$.

The VDM Algorithm

Step 0 (*Initialization*). Select a point $\lambda_1 \in (0, B]$. Let $G_1 = \delta_{\lambda_1}$ be the initial value. Set the loop index $L = 1$.

Step 1 If $\max_{\lambda\in[0,B]} \Phi'(G_L, \delta_\lambda) = 0$, then stop and return G_L. Otherwise, find $\lambda_{\mathsf{max}} = \operatorname{argmax}_{\lambda\in[0,B]} \Phi'(G_L, \delta_\lambda)$.

Step 2 Find $\alpha_{\max} = \text{argmax}_{\alpha\in[0,1]}\, \Phi\Big\{(1-\alpha)G_L \oplus \alpha\delta_{\lambda_{\mathsf{max}}}\Big\}$.

Step 3 Set $G_{L+1} = (1-\alpha)G_L \oplus \alpha_{\max}\delta_{\lambda_{\mathsf{max}}}$. Set $L = L+1$ and go to Step 1.

The VEM Algorithm

Step 0 (*Initialization*). Select a point $\lambda_1 \in (0, B]$. Let $G_1 = \delta_{\lambda_1}$ be the initial value. Set the loop index $L = 1$.

Step 1 If $\max\limits_{\lambda\in[0,B]} \Phi'(G_L, \delta_\lambda) = 0$, then stop and return G_L. Otherwise, find $\lambda_{\mathsf{max}} = \operatorname{argmax}\limits_{\lambda\in[0,B]} \Phi'(G_L, \delta_\lambda)$ and $\lambda_{\mathsf{min}} = \operatorname{argmin}\limits_{\lambda\in\mathsf{supp}(G_L)} \Phi'(G_L, \delta_\lambda)$, where $\mathsf{supp}(G_L)$ stands for the support of G_L.

Step 2 Find $\alpha_{\max} = \text{argmax}_{\alpha\in[0,1]}\, \Phi\Big\{G_L \oplus \Big(\alpha G_L(\lambda_{\mathsf{min}})(\delta_{\lambda_{\mathsf{max}}} \ominus \delta_{\lambda_{\mathsf{min}}})\Big)\Big\}$.

Step 3 Set $G_{L+1} = G_L \oplus \Big(\alpha_{\max} G_L(\lambda_{\mathsf{min}})(\delta_{\lambda_{\mathsf{max}}} \ominus \delta_{\lambda_{\mathsf{min}}})\Big)$. Set $L = L+1$ and go to Step 1.

The ISDM Algorithm

Step 0 (*Initialization*). Select a point $\lambda_1 \in (0, B]$. Let $G_1 = \delta_{\lambda_1}$ be the initial value. Set the loop index $L = 1$.

Step 1 If $\max\limits_{\lambda\in[0,B]} \Phi'(G_L, \delta_\lambda) = 0$, then stop and return G_L. Otherwise, find all local maxima $\lambda_{\max,1}, \ldots, \lambda_{\max,\mathcal{N}}$ of $\lambda \mapsto \Phi'(G_L, \delta_\lambda)$ on $[0, B]$, where $\mathcal{N}$ represents the number of local maxima.

Step 2 Find $(\alpha_{\max,0}, \ldots, \alpha_{\max,\mathcal{N}}) = \operatorname{argmax}\limits_{\alpha_0,\ldots,\alpha_{\mathcal{N}}} \Phi\Big\{(1-\alpha_0)G_L \oplus \alpha_1\delta_{\lambda_{\max,1}} \oplus \cdots \oplus \alpha_{\mathcal{N}}\delta_{\lambda_{\max,\mathcal{N}}}\Big\}$ subject to $\alpha_0 \geq 0, \alpha_1 \geq 0, \cdots, \alpha_{\mathcal{N}} \geq 0$ and $\alpha_0 + \alpha_1 + \cdots + \alpha_{\mathcal{N}} = 1$.

Step 3 Set $G_{L+1} = (1-\alpha_{\max,0})G_L \oplus \alpha_{\max,1}\delta_{\lambda_{\max,1}} \oplus \cdots \oplus \alpha_{\max,\mathcal{N}}\delta_{\lambda_{\max,\mathcal{N}}}$. Set $L = L+1$ and go to Step 1.

The convergence of VDM, VEM, and ISDM is guaranteed by the following theorem.

Theorem 3.5. *Assuming $r_i > 0$ for each $i \in [N]$. For each of VDM, VEM and ISDM, if it stops for some L, then we have $\Phi(G_L) = \Phi(\hat{Q})$; otherwise, $\Phi(G_L) \to \Phi(\hat{Q})$ as $L \to \infty$.*

Remark 3.4. Unlike in the classical setting where all read depths are identical, when heterogeneous read depths are incorporated, although $G \mapsto \Phi(G)$ is still a concave function, there is no theoretical guarantee about the uniqueness of $\hat{Q}$'s that maximize the objective function and whether the maximizer is unique or not is still open. This issue of computational uniqueness shall be compared to the parallel result in Theorem 3.2, which provides theoretical guarantee for the consistency of an arbitrary maximizer of the objective function as the sample size increases to infinity.

3.4 Simulation studies

This chapter is split to two parts. The first part aims to show that the two NPMLE-based tests presented in Chapter 3.2 cannot dominate each other. To this end, we fix $K = 2$ and consider three designs with several cases of population models that will be detailed in Chapter 3.4.1. The second part provides some preliminary discussions on the computation complexity of the proposed algorithm.

3.4.1 Finite-sample experiments

Designs.

(A) Balanced designs with all read depths set to be 1, $n_1 = n_2 = 10$, and $N_{jk} = 50$, 100, and 500 for each j, k.

(B) Balanced designs with read-depth effects with $n_1 = n_2 = 10$ and $N_{jk} = 50$, 100 and 500 for each j, k. In addition, in each round of the simulation, $\{r_{i1}^{(1)}, i \in [N_{11}]\}$ are i.i.d. generated from Uniform$(0.5, 1.5)$ and then let $r_{ij}^{(k)} = r_{i1}^{(1)}$ for each j, k.

(C) A particular unbalanced design motivated by the single-cell RNA-seq data in Chapter 3.5 ahead, with $n_1 = 10, n_2 = 13$ and N_{jk} be as in Table 3. For each round of the simulation, $\{r_{ij}^{(k)}, i \in [N_{jk}], j \in [n_k], k \in [K]\}$ are i.i.d. generated from Uniform$(0.5, 1.5)$.

Table 3: N_{jk} in the unbalanced design (Design (C))

$N_{1,1}$	$N_{2,1}$	$N_{3,1}$	$N_{4,1}$	$N_{5,1}$	$N_{6,1}$	$N_{7,1}$	$N_{8,1}$	$N_{9,1}$	$N_{10,1}$	$N_{1,2}$	$N_{2,2}$
388	1142	162	391	215	278	284	193	542	106	202	759
$N_{3,2}$	$N_{4,2}$	$N_{5,2}$	$N_{6,2}$	$N_{7,2}$	$N_{8,2}$	$N_{9,2}$	$N_{10,2}$	$N_{11,2}$	$N_{12,2}$	$N_{13,2}$	
415	69	327	431	414	451	275	733	422	65	362	

We then move on to specify the population model (3.3) used in our simulation studies. Hereafter, let Gam$(a, b; B)$ denote a truncated Gamma distribution with a shape parameter $a > 0$, a rate parameter $b > 0$, and with any realization larger than B shrunken to B. Let $\{\Delta_j^{(k)}, j \in [n_k], k \in [K]\}$ be i.i.d. generated from Uniform$(-1, 1)$.

Population models.

1. (a) $Q_j^{(k)} \sim$ Gam$(14 + \Delta_j^{(k)}, 7/4; 50)$ for each $j \in [n_k], k \in [2]$.
   (b) $Q_j^{(k)} \sim$ Gam$(14 + \Delta_j^{(k)}, 7; 50)$ for each $j \in [n_k], k \in [2]$.
   (c) $Q_j^{(k)} \sim$ Gam$(6 + \Delta_j^{(k)}, 1; 50)$ for each $j \in [n_k], k \in [2]$.
2. (a) $Q_j^{(1)} \sim$ Gam$(14 + \Delta_j^{(1)}, 7/4; 50)$ for $j \in [n_1]$ and $Q_j^{(2)} \sim$ Gam$(6 + \Delta_j^{(2)}, 3/4; 50)$ for $j \in [n_2]$.
   (b) $Q_j^{(1)} \sim$ Gam$(14 + \Delta_j^{(1)}, 7/3; 50)$ for $j \in [n_1]$ and $Q_j^{(2)} \sim$ Gam$(6 + \Delta_j^{(2)}, 1; 50)$ for $j \in [n_2]$.

(c) $Q_j^{(1)} \sim \text{Gam}(14+\Delta_j^{(1)}, 7/2; 50)$ for $j \in [n_1]$ and $Q_j^{(2)} \sim \text{Gam}(6+\Delta_j^{(2)}, 3/2; 50)$ for $j \in [n_2]$.

3. (a) $Q_j^{(1)} \sim \text{Gam}(4+\Delta_j^{(1)}, 1; 20)$ for $j \in [n_1]$ and $Q_j^{(2)} \sim \text{Gam}(5+\Delta_j^{(2)}, 1; 20)$ for $j \in [n_2]$.
   (b) $Q_j^{(1)} \sim \text{Gam}(5+\Delta_j^{(1)}, 1; 20)$ for $j \in [n_1]$ and $Q_j^{(2)} \sim \text{Gam}(6+\Delta_j^{(2)}, 1; 20)$ for $j \in [n_2]$.
   (c) $Q_j^{(1)} \sim \text{Gam}(6+\Delta_j^{(1)}, 1; 20)$ for $j \in [n_1]$ and $Q_j^{(2)} \sim \text{Gam}(7+\Delta_j^{(2)}, 1; 20)$ for $j \in [n_2]$.
4. (a) $Q_j^{(1)} \sim \text{Gam}(11+\Delta_j^{(1)}, 1; 50)$ for $j \in [n_1]$ and $Q_j^{(2)} \sim \text{Gam}(12+\Delta_j^{(2)}, 1; 50)$ for $j \in [n_2]$.
   (b) $Q_j^{(1)} \sim \text{Gam}(12+\Delta_j^{(1)}, 1; 50)$ for $j \in [n_1]$ and $Q_j^{(2)} \sim \text{Gam}(13+\Delta_j^{(2)}, 1; 50)$ for $j \in [n_2]$.
   (c) $Q_j^{(1)} \sim \text{Gam}(13+\Delta_j^{(1)}, 1; 50)$ for $j \in [n_1]$ and $Q_j^{(2)} \sim \text{Gam}(14+\Delta_j^{(2)}, 1; 50)$ for $j \in [n_2]$.

Our focus is on examining as well as comparing the empirical performance of the tests $\hat{T}_\alpha$ and $\hat{T}_{h,\alpha}$ with NPMLE calculated using the oracle B. Both of them are based on an exact critical value approximated by 1,000 Monte Carlo simulations. The underlying nominal significance level is 0.05. For each setting, $1,000$ rounds of simulations were performed. We use VEM to compute NPMLEs with a stop tolerance 0.01. Optimization in Step 1 and Step 2 in VEM is implemented by the default interior-point algorithm in Matlab; see the support page of function '*fmincon*' for further details.

Table 4 shows the empirical sizes and powers (rejection frequencies) of tests $\hat{T}_\alpha$ and $\hat{T}_{h,\alpha}$. In short, the results confirm our earlier theoretical claims on the sizes and powers of T_α and $\hat{T}_{h,\alpha}$ in the different models and balanced designs (Designs (A) and (B)). Moreover, even under the unbalanced design (Design (C)), T_α and $\hat{T}_{h,\alpha}$ still perform well in terms of their empirical sizes and powers.

Some more detailed comparisons between T_α and $\hat{T}_{h,\alpha}$ are in line. The following observations depend on the "signal strengths" D and D_h, defined as follows:

$$D := E\{W_1(Q_1^{(1)}, Q_1^{(2)})^2\} - \left(E\{W_1(Q_1^{(1)}, Q_2^{(1)})^2\} + E\{W_1(Q_1^{(2)}, Q_2^{(2)})^2\}\right)/2 \tag{3.10}$$

and

$$D_h := E\{W_1(h_{Q_1^{(1)}}, h_{Q_1^{(2)}})^2\} - \left(E\{W_1(h_{Q_1^{(1)}}, h_{Q_2^{(1)}})^2\} + E\{W_1(h_{Q_1^{(2)}}, h_{Q_2^{(2)}})^2\}\right)/2. \tag{3.11}$$

First, empirical results for Model 1 illustrates that under H_0, empirical powers are close to the nominal level $\alpha = 0.05$, confirming the size validity of $\hat{T}_\alpha$ and $\hat{T}_{h,\alpha}$. In addition, even under the unbalanced design (Design (C)), empirical powers are stable and close to the nominal level $\alpha = 0.05$, indicating the robustness of the studied tests.

Second, we compare the empirical powers using Models 2, 3, and 4. In Model 2, D is significantly larger than D_h and the corresponding empirical powers of $\hat{T}_\alpha$ are all larger than these of $\hat{T}_{h,\alpha}$ in all three considered designs (Designs (A), (B), and (C)). This phenomenon is not surprising to us as the difference between variation between groups and variation within groups in mixing distributions is much larger than that in mixture distributions. Therefore, $\hat{T}_\alpha$ is more powerful than $\hat{T}_{h,\alpha}$.

In Model 3, D is approximately equal to D_h and the empirical power of $\hat{T}_\alpha$ is smaller than the empirical power of $\hat{T}_{h,\alpha}$ when N is small (e.g., 50 and 100). However, the empirical powers of $\hat{T}_\alpha$ and $\hat{T}_{h,\alpha}$ are close when N is large. Similar observation applies to Model 4, where D is also approximately equal to D_h. However, compared to Model 3, the mixing distributions in Model 4 have larger B and thus the empirical powers of $\hat{T}_{h,\alpha}$ are higher than the empirical powers of $\hat{T}_\alpha$ even for $N = 500$, especially under Design (A). Some pilot studies to explain this phenomenon will be put in Chapter 3.6, where we analyze the finite-sample behavior of the NPMLE under an exploratory simplified setting where all read depths are fixed to be 1. There, the rate of convergence of NPMLE, at the worst case, is showed to be $O(\log\log N/\log N)$; in contrast, Lambert and Tierney (1984, Lemma 4.1 and Theorem 4.1) showed that the Poisson-smoothed NPMLE attains a near-root-n rate of convergence to the mixture distribution.

3.4.2 Time complexity and actual running time

This chapter provides some discussions on the algorithm implemented in Chapter 3.4.1, with B assumed to be bounded. This algorithm consists of two major parts: (1) implementing the VEM algorithm detailed in Chapter 3.3 for calculating estimates of the mixing distributions; (2) feeding the estimates to the permutation tests in Chapter 3.2.2. In the following we discussed the computation complexity of these two parts separately.

Time complexity of NPMLE

We start with an analysis of the VEM algorithm. Using the notation of Chapter 3.3, every iteration of VEM involves the following three steps:

1. find $\lambda_{\max} = \operatorname{argmax}_{\lambda\in[0,B]} \Phi'(G_L, \delta_\lambda)$;
2. find $\lambda_{\min} = \operatorname{argmin}_{\lambda\in\operatorname{supp}(G_L)} \Phi'(G_L, \delta_\lambda)$;
3. find $\alpha_{\max} = \operatorname{argmax}_{\alpha\in[0,1]} \Phi\{G_L \oplus [\alpha G_L(\lambda_{\min})(\delta_{\lambda_{\max}} \ominus \delta_{\lambda_{\min}})]\}$,

where it is reminded that

$$\Phi'(G_L, \delta_\lambda) = \frac{1}{N}\sum_{i\in[N]} \frac{e^{-\lambda r_i}(\lambda r_i)^{X_i}}{\sum_{m\in[M]} G_L(\lambda_m)\, e^{-\lambda_m r_i}(\lambda_m r_i)^{X_i}} - 1.$$

In Step 1, to obtain an ϵ-accuracy (with respect to the objective function) solution, we search over an ϵ-grid on $[0, B]$ (i.e., $[0, \epsilon, 2\epsilon, \ldots, B]$) and then use a gradient-descent algorithm with the best result over the grid as the initial value to obtain an accurate solution, an idea commonly used in literatures; see, e.g., Lindsay (1995, Chapter 6.1). Since the derivative of $\lambda \mapsto \Phi'(G_L, \delta_\lambda)$ is bounded on $[0, B]$, it is immediate that this

Table 4: Empirical sizes and powers of $\hat{T}_\alpha$ and $\hat{T}_{h,\alpha}$; here D and D_h are defined in (3.10) and (3.11)

Model	1(a)	1(b)	1(c)	2(a)	2(b)	2(c)	3(a)	3(b)	3(c)	4(a)	4(b)	4(c)
D	0	0	0	0.59	0.32	0.15	0.99	0.99	0.99	0.99	0.99	0.99
D_h	0	0	0	0.22	0.10	0.03	0.99	0.99	0.99	0.99	0.99	0.99
N	Empirical sizes/powers for $\hat{T}_\alpha$ under Design (A)											
50	0.054	0.050	0.045	0.644	0.595	0.356	0.811	0.772	0.698	0.538	0.501	0.502
100	0.043	0.055	0.053	0.901	0.835	0.583	0.870	0.872	0.843	0.723	0.680	0.668
500	0.049	0.049	0.060	0.996	0.999	0.965	0.952	0.958	0.941	0.850	0.831	0.829
N	Empirical sizes/powers for $\hat{T}_{h,\alpha}$ under Design (A)											
50	0.054	0.045	0.049	0.284	0.210	0.111	0.833	0.816	0.767	0.650	0.635	0.624
100	0.038	0.063	0.049	0.371	0.264	0.138	0.896	0.892	0.882	0.797	0.788	0.771
500	0.042	0.047	0.055	0.492	0.309	0.186	0.951	0.961	0.947	0.944	0.924	0.921
N	Empirical sizes/powers for $\hat{T}_\alpha$ under Design (B)											
50	0.044	0.048	0.058	0.644	0.508	0.338	0.796	0.763	0.729	0.559	0.522	0.520
100	0.053	0.050	0.062	0.863	0.779	0.518	0.878	0.862	0.846	0.714	0.735	0.679
500	0.036	0.052	0.054	1.000	0.998	0.972	0.958	0.952	0.939	0.922	0.920	0.913
N	Empirical sizes/powers for $\hat{T}_{h,\alpha}$ under Design (B)											
50	0.044	0.050	0.054	0.262	0.193	0.100	0.821	0.806	0.772	0.632	0.619	0.602
100	0.058	0.041	0.053	0.350	0.276	0.132	0.885	0.877	0.858	0.772	0.788	0.759
500	0.036	0.045	0.057	0.501	0.414	0.187	0.956	0.950	0.943	0.932	0.928	0.924
N	Empirical sizes/powers for $\hat{T}_\alpha$ under Design (C)											
Table 3	0.048	0.050	0.051	0.994	0.988	0.900	0.962	0.940	0.951	0.910	0.904	0.907
N	Empirical sizes/powers for $\hat{T}_{h,\alpha}$ under Design (C)											
Table 3	0.047	0.051	0.052	0.452	0.346	0.173	0.966	0.947	0.952	0.929	0.920	0.922

The mixing distribution Q	Gam(14,1;50)	Gam(10,1;50)	Gam(6,1;50)
Actual running time	6.31	5.74	4.99

Table 5: Actual running time (in seconds) for computing NPMLEs.

grid search method indeed leads to an ϵ-accuracy solution for sufficiently small ϵ; here the boundedness of the derivative of $\lambda \mapsto \Phi'(G_L, \delta_\lambda)$ follows from the derivative of $\lambda \mapsto e^{-\lambda r_i} (\lambda r_i)^{X_i}$, i.e.,

$$\frac{e^{-r_i\lambda}(r_i\lambda)^{X_i}(X_i - r_i\lambda)}{\lambda} = \begin{cases} -r_i e^{-r_i\lambda} & \text{if } X_i = 0, \\ r_i^{X_i} e^{-r_i\lambda}\lambda^{X_i-1}(X_i - r_i\lambda) & \text{if } X_i \geq 1, \end{cases}$$

which is bounded on $[0, B]$. The time complexity for the grid search method is $O(N^2/\epsilon)$, where N^2 comes from the sum of index $i \in [N]$ and $m \in [M]$ with $M \leq N$. For the time complexity of the gradient-descent algorithm, it follows from Nesterov's Theorem (Nesterov, 2003, Theorem 2.1.14) that one needs $O(N^2/\epsilon)$ iterations to obtain an ϵ-accuracy solution as long as the objective function has a Lipschitz continuous gradient, or equivalently, the boundedness of the second derivative of $\lambda \mapsto \Phi'(G_L, \delta_\lambda)$. This follows from the boundedness of the second derivative of $\lambda \mapsto e^{-\lambda r_i} (\lambda r_i)^{X_i}$, i.e.,

$$\frac{e^{-\lambda r_i}(\lambda r_i)^{X_i}\left(\lambda^2 r_i^2 - X_i(2\lambda r_i + 1) + X_i^2\right)}{\lambda^2} = \begin{cases} r_i^2 e^{-\lambda r_i} & \text{if } X_i = 0, \\ r_i^2 e^{-\lambda r_i}(\lambda r_i - 2) & \text{if } X_i = 1, \\ r_i^{X_i} e^{-\lambda r_i}\lambda^{X_i-2}[(\lambda r_i - X_i)^2 - X_i] & \text{if } X_i \geq 2. \end{cases}$$

on $[0, B]$.

In Step 2, since the support size of G_L is at most N and the summation over i is from 1 to N in $\Phi'(G_L, \delta_\lambda)$, a brutal force method requires at most $O(N^2)$ to find the $\lambda_{\min}$. The time complexity for Step 3 is the same as for Step 1 by analogous arguments, and hence the total time complexity in three steps is $O(N^2/\epsilon)$ to obtain an ϵ-accuracy solution in each iteration.

To determine the total time complexity, it remains to determine how many iterations (recalling that each iteration contains the above three steps) are needed. It follows from Böhning (1982, Assumption (iii) and its proof) or Equation (3.24) in the supplement that the increase of the objective function is strict and linear after each iteration. Accordingly, the number of iterations is $O(1/\epsilon)$ to obtain an ϵ-accuracy solution and the total time complexity for the VEM algorithm is $O(N^2/\epsilon^2)$.

Table 5 shows the actual running time (in seconds) for computing NPMLEs averaged over 100 simulations; recall that the tolerance level is set to be 0.01. Here we adopt the Design (B) with $N = 500$, $\{r_i : i \in [N]\} \overset{\text{i.i.d.}}{\sim}$Uniform(0.5,1.5), and consider three population models, Gam(14,1;50), Gam(10,1;50), and Gam(6,1;50). The simulation is conducted over a laptop with a 1.8 GHz Intel Core i5 processor and an 8 GB memory.

Time complexity of permutation tests

We then move on to examine the time complexity of the remaining parts. For computing W_1 distance between two Poisson mixture distributions, where the corresponding mixing distributions are both supported over at most M points, first note that the time complexity of computing the mixture density at any point is $O(M)$. Moreover, it follows from Poisson tail inequality (Lemma 3.5) that as long as the support mixing distributions is bounded by some positive constant, it suffices to compute the mixture densities over at most $O(\sqrt{\log(1/\epsilon)})$ points to obtain an ϵ-accuracy solution. As a consequence, the time complexity for computing the W_1 distance between two Poisson mixtures is $O(M\sqrt{\log(1/\epsilon)})$.

With a little abuse of notation, let's use N to denote $\max_{i,k}\{N_{jk}\}$. Since each NPMLE is supported over at most N points, the time complexity of computing the W_1 distance of estimated mixture and mixing distributions is $O(N\sqrt{\log(1/\epsilon)})$ and hence the time complexity of computing the W_1 distance matrix is $O(n_1 n_2 N\sqrt{\log(1/\epsilon)})$, where n_1 and n_2 are the number of subjects in each group. As a result, the total time complexity of performing permutation-based test is

$$O(n_1 n_2 N\sqrt{\log(1/\epsilon)} + n_1 n_2 \mathcal{M}),$$

where $\mathcal{M}$ represents the set number of permutations.

For an example of the actual running time, with $\mathcal{M} = 1,000$, the total time to perform permutation tests with mixture and mixing distribution estimates input is averagely 0.19 second on a laptop with a 1.8 GHz Intel Core i5 processor and a 8 GB memory, where the mixing distributions are NPMLEs estimated under Design (A) and Model 4(c) with $N = 500$.

3.5 Applications to single-cell genomics

This chapter applies the studied permutation tests to a scRNA-seq data. There has been a large literature studying fitting RNA-seq data using Poisson mixtures including, e.g., over-dispersed Poisson model (Robinson et al., 2010), Poisson-Gamma model (Love et al., 2014; Huang et al., 2018), Poisson-Beta model (Vu et al., 2016), Poisson-log normal model (Silva et al., 2019), Poisson mixture model with K-clusters (Rau et al., 2015), finite Poisson mixture models (Wu et al., 2013), zero-inflated mixture Poisson linear models (Liu et al., 2019), Poisson mixture models with unimodal mixing distributions (Lu, 2018). Compared to parametric Poisson mixture models, nonparametric Poisson mixture models haven't received much attention; some notable exceptions include Bi and Davuluri (2013), Dadaneh et al. (2018), Sarkar and Stephens (2021), the latter of which was closely followed by us.

3.5.1 Data set description

The scRNA-seq data used in this chapter is obtained from Velmeshev et al. (2019), which focused on autism spectrum disorder (ASD) and recorded gene expression of 23 subjects (13 ASD v.s. 10 control) and 18,041 genes for each subject from 17 different cell types and 2 different brain regions. Here we focus on the brain region prefrontal cortex, which is more relevant to autism disease etiology. Moreover, each subject has 7 covariates including age, sex, diagnosis, capbatch, seqbatch, post-mortem interval (PMI), and RNA integrity number (RIN).

We focus on a pre-selected subset including 100 genes (names of the genes put in Table 6) that were documented to be related to body height; for relation between ASD and body height, see, e.g., Fukumoto et al. (2011) and Chawarska et al. (2011). In addition to permutation testing with either estimated mixing distributions or mixture distributions, we also consider DESeq2 (Love et al., 2014) as a benchmark. In implementing the two considered permutation tests, we adopt a common strategy to incorporate four covariates: age, sex, seqbatch, and RIN. The other two covariates PMI and capbatch are not significantly associated with gene expression given the other covariates, since their p-value distributions across all genes are uniform. The corresponding tests were denoted as $\hat{T}_Z$ (with the original NPMLE) and $\hat{T}_{h,Z}$ (with the Poisson-smoothed NPMLE). Implementation details — including the choice of B, the choice of read depths, and an additional step of covariate adjustment based on the work of Zhang et al. (2022) — were put in the supplement Chapter 3.8.

3.5.2 Implementation results

Using $\hat{T}_Z$, 9 genes are significant under the threshold of false discovery rate (FDR) 0.05 after multiple testing correction by the Benjamini-Hochberg procedure. Replacing $\hat{T}_Z$ by $\hat{T}_{h,Z}$, 8 genes are significant under the same threshold of FDR and 7 genes are coincident with significant genes found by $\hat{T}_Z$. This shows some consistency between $\hat{T}_Z$ and $\hat{T}_{h,Z}$.

Furthermore, by DESeq2 (Love et al., 2014) there are 7 significant genes under the same threshold of FDR and all of them are coincident with significant genes found by $\hat{T}_Z$. In other words, among significant genes found by $\hat{T}_Z$, 78% significant genes are coincident with genes found by DESeq2 and 22% are new which means $\hat{T}_Z$ could enrich the set of significant genes found by the standard method DESeq2.

Similarly, 6 genes are coincident with significant genes found by $\hat{T}_{h,Z}$. In other words, among significant genes found by $\hat{T}_{h,Z}$, 75% significant genes are coincident with genes found by DESeq2 and 25% are new which means $\hat{T}_{h,Z}$ could enrich the set of significant genes found by the standard method DESeq2. In one word, both $\hat{T}_Z$ and $\hat{T}_{h,Z}$ could enrich the set of significant genes found by DESeq2. Further details are

summarized in Figure 9.

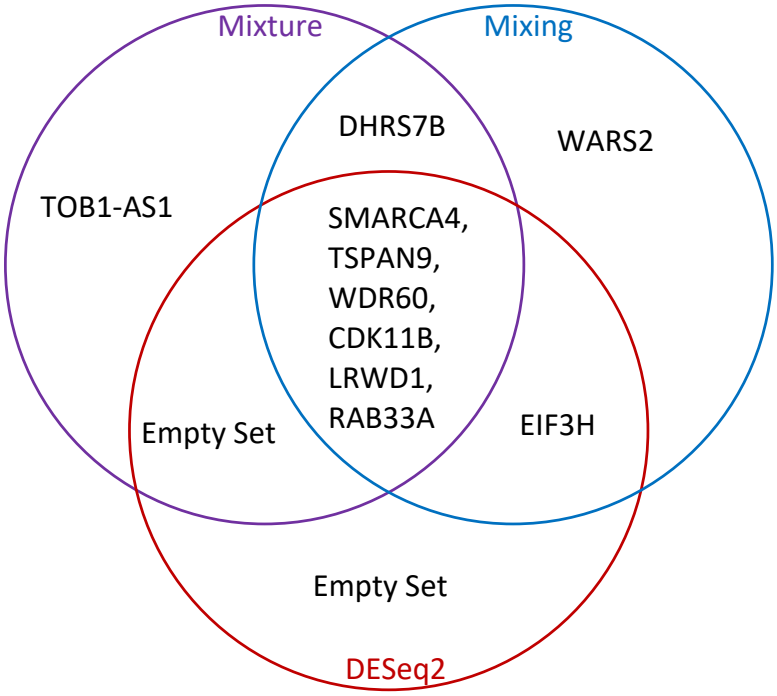

Figure 9: Significant genes selected using Mixing ($\hat{T}_Z$), Mixture ($\hat{T}_{h,Z}$), and DESeq2 methods.

Our results can also be justified by functions of significant genes. For example, fasting blood glucose measurement is not only one of functions of gene DHRS7B, but also related to ASD (Hoirisch-Clapauch and Nardi, 2019). More such results are summarized in Table 7.

3.6 Minimax optimality of the Poisson NPMLEs

This chapter provides additional theoretical support for the use of NPMLEs in forming up the tests $\hat{T}_\alpha$ and $\hat{T}_{h,\alpha}$ in Chapter 3.2. To this end, due to the technical challenges, focus is restricted to a simplified setting of (3.4), where the observations $\{X_i, i \in [N]\}$ independently follow a distribution of PMF

$$h_Q(x) = \int_0^B e^{-\lambda} \frac{\lambda^x}{x!} \mathrm{d}Q(\lambda), \quad x = 0, 1, 2, \ldots, \tag{3.12}$$

where Q is a deterministic measure supported on $[0, B]$ that cannot be characterized by a simple parametric model. This is exactly the classic nonparametric Poisson mixture setup, and we study the nonasymptotic behavior of the following NPMLE

$$\hat{Q} = \underset{Q \text{ of support } [0,B]}{\operatorname{argmax}} \sum_{i \in [N]} \log h_Q(X_i). \tag{3.13}$$

Table 6: All genes used in Chapter 3.5

DST	CHSY3	TSC2	EHD4	HERC1	KIF16B	DLGAP1	PIK3CG
ELL	ODF2L	FBXL5	LNX1	ERGIC3	CBFA2T2	FAM20A	STAT2
DAP	SSH2	WDR60	SAXO1	FOXP2	SAMD4A	TSPAN9	ARAP3
GHR	KCNK9	RGL1	SOCS5	ZNF76	ADAMTS2	DHRS7B	PNMA8C
KIZ	SHPRH	RBMS3	MFSD2B	NR4A3	CCDC171	RAB33A	WDR70
IL16	MTMR3	CDK10	ZNF628	CAPZB	ATXN7L3	PSKH1	FGFRL1
BST2	UMAD1	CPED1	ESYT2	LRRC43	SMARCA4	MYO18A	IL17RD
LHX2	FBP2	ZC3H13	SRRM2	NOTCH1	HSD17B3	SBNO1	EIF3H
RLF	LAYN	SUSD5	DOT1L	WARS2	RPS4XP13	PHF11	CDK11B
DAZL	CYFIP2	ST7L	CWC27	C9orf152	TOB1-AS1	HIF1AN	KLHL28
BCL9	LRWD1	LMO7	PTENP1	CEP112	LINC01572	PPP4R2	UBE2Z
NRK	GCLC	PPM1H	ITGA9	HIP1R	PPP1R16A	POLR3E	TANC2
ANKDD1A		ZNF710-AS1		ZRANB2-AS2		DNAJC27-AS1	

Table 7: Significant genes on ASD with some literature support. The first column includes names of genes, the second column includes functions potentially related to ASD, and the third column includes literature support

gene name	related functions	literatures
DHRS7B	fasting blood glucose	Hoirisch-Clapauch and Nardi (2019)
WDR60	abnormality of refraction	Ezegwui et al. (2014)
EIF3H	reaction time measurement	Baisch et al. (2017)
LRWD1	insomnia measurement	Hohn et al. (2019)
RAB33A	bipolar disorder	Joshi et al. (2012)
TSPAN9	creatinine measurement	Cameron et al. (2017)
WARS2, CDK11B	heel bone mineral density	Calarge and Schlechte (2017)
SMARC4, TOB1-AS1	cholesterol measurement	Benachenhou et al. (2019)

Note that the above NPMLE is the simplified version of (3.7) with all read depths there forced to be one.

There has been an enormous literature studying the NPMLE (3.13) under the nonparametric Poisson mixture model (3.12). Earlier results on the existence, discreteness (of the NPMLE support), and computation include, among many others, Simar (1976), Laird (1978), Jewell (1982), Lindsay (1983a), Lindsay (1983b), and Lindsay and Roeder (1993); see also Lindsay (1995) for a survey. Consistency of NPMLEs were established in, among many others, Kiefer and Wolfowitz (1956), Simar (1976), and Pfanzagl (1988); see also Chen (2017) for a survey.

Beyond these important results, there has been another track of substantial research that is focused on establishing the minimax rate in estimating the mixing distribution (mostly on the density function) of nonparametric Poisson mixtures. Notable results there include Zhang (1995), Loh and Zhang (1996), van de Geer (1996), Hengartner (1997), van de Geer (2003), Roueff and Rydén (2005), and Rebafka and Roueff (2015). However, to our knowledge, a study on the minimax optimality and the corresponding convergence rates for NPMLEs under a fully nonparametric Poisson mixture model is still absent from the literature.

Before presenting our main result in this chapter, we would love to highlight again that, due to the nature of nonasymptotic analysis, all the parameters in the model (3.12), including B, are allowed to change with N. This is a strict generalization of the "asymptotic" setting in Chapter 3.2, where, due to the additional hardness of handling the read depth as well as for simplifying notation and assumptions, we do not intend to establish similar nonasymptotic results.

Our first theorem concerns with the NPMLE's rate of convergence.

Theorem 3.6 (Upper bound of NPMLEs)**.**

(a) Suppose there exists a universal constant $c_0 > 0$ such that $B \le c_0 \log N$. Then there exists a positive constant $C = C(c_0)$ such that for all sufficiently large N ($> N_0(c_0)$) we have

$$\sup_{Q \text{ of support } [0,B]} E\Big\{W_1(\hat{Q}, Q)\Big\} \le C \frac{B}{\log N} \log\left(\frac{\log N}{B} \vee e\right).$$

(b) Suppose there exist universal strictly positive constants c_0, C_0 and $\epsilon_0 \in (0, 1/3)$ such that $B \in [c_0 \log N, C_0 N^{1/3-\epsilon_0}]$. Then there exists a strictly positive constant $C = C(\epsilon_0, c_0)$ such that for all sufficiently large N ($> N_0(c_0, C_0, \epsilon_0)$) we have

$$\sup_{Q \text{ of support } [0,B]} E\Big\{W_1(\hat{Q}, Q)\Big\} \le C\sqrt{\frac{B}{\log N}}.$$

Our second theorem concerns with minimax lower bounds in estimating mixing distributions in model (3.12). Combined with Theorem 3.6, it confirms the NPMLE's minimax optimality.

Theorem 3.7 (Minimax lower bound of mixing distribution estimation)**.**

(a) Supposing there exists $c_0 > 0$ *such that* $B \leq c_0 \log N$*,it follows that for any* $N \geq 3$*,*

$$\inf_{\tilde{Q}} \sup_{Q} E\{W_1(\tilde{Q}, Q)\} \geq \frac{B}{24e \log N} \log\Big(\frac{16 c_0 \log N}{B}\Big).$$

(b) Supposing there exists $c_0 > 0$ *such that* $B \geq c_0 \log N$*, it follows that for any* $N \geq 1$*,*

$$\inf_{\tilde{Q}} \sup_{Q} E\{W_1(\tilde{Q}, Q)\} \geq \frac{3}{40e^4} \sqrt{\frac{B}{c_0 \log N}}.$$

In the above, the infimum and supremum are understood to be taken over all estimators and all distributions of support $[0, B]$

Remark 3.5. Under fully nonparametric binomial mixture models, minimax optimal convergence rates for NPMLEs of mixing distributions were obtained by Vinayak et al. (2019, Section 3) in terms of the W_1 distance. Under fully nonparametric binomial and Gaussian mixture models, Tian et al. (2017, Theorem 1) and Wu and Yang (2020a, Page 1985) obtained optimal convergence rates for moment-based estimators in terms of W_1 distance; see also Polyanskiy and Wu (2020, Remark 2). Nguyen (2013, Theorems 1 and 2) upper bounded the Wasserstein distance between mixing distributions by the divergence between the corresponding mixture distributions under general mixture models, with normal mixture models as an example in Example 2. However, their results cannot be applied here since Theorem 1 restricts the mixing distribution being discrete and Theorem 2 is only for convolution mixture models.

3.7 Proofs

3.7.1 Proofs of theorems in Chapter 3.2

Proof of Theorem 3.1. To simplify notations, we temporarily drop the subject index j and the group index k in this proof. A restatement of this theorem is then as follows:

> *Suppose there exists an estimator* $\tilde{Q} = \tilde{Q}_N$ *on* $[0, B]$ *such that* $E\{W_1(\tilde{Q}, Q) \mid Q\} \to 0$ *as* $N = N_n \to \infty$ *for almost all* Q *with regard to the measure* $\mathcal{Q}$*. Then we have* $E\{W_1(h_{\tilde{Q}}, h_Q) \mid Q\} \to 0$ *as* $N = N_n \to \infty$ *for almost all* Q *with regard to the measure* $\mathcal{Q}$*.*

This proof consists of three steps. In the first step, we assume both Q and $\tilde{Q}$'s are ordinary distributions with no randomness and prove that $W_1(\tilde{Q}, Q) \to 0$ implies $W_1(h_{\tilde{Q}}, h_Q) \to 0$. In the second step, we temporarily forget the third-layer "population model" (3.3) and prove that $E\{W_1(\tilde{Q}, Q)\} \to 0$ implies $E\{W_1(h_{\tilde{Q}}, h_Q)\} \to 0$ where the expectation is with respect to randomness from the "measurement model" (3.1) and "expression

model" (3.2). In the third step, the third-layer "population model" (3.3) gets involved and we complete this proof.

Step 1. Suppose $\{\tilde{Q}\}$ is a sequence of ordinary distributions with no randomness. To prove that $W_1(\tilde{Q}, Q) \to 0$ implies $W_1(h_{\tilde{Q}}, h_Q) \to 0$, note that $W_1(\tilde{Q}, Q) \to 0$ is equivalent to $\tilde{Q} \xrightarrow{d} Q$ supplemented with $E\{|\tilde{\Lambda}|\} \to E\{|\Lambda|\}$ (Panaretos and Zemel, 2019, Section 2.3), where $\tilde{\Lambda}$ is a random variable with distribution $\tilde{Q}$ and Λ is a random variable with distribution Q. Moreover, it follows from Skorokhod's representation theorem that we can assume $\tilde{Q} \xrightarrow{a.s.} Q$. To prove $W_1(h_{\tilde{Q}}, h_Q) \to 0$, it suffices to prove that $h_{\tilde{Q}} \xrightarrow{d} h_Q$ and $E\{\tilde{X}\} \to E\{X\}$, where $\tilde{X}$ is a random variable with distribution $h_{\tilde{Q}}$ and X is a random variable with distribution h_Q. The second part follows immediately from $E\{\tilde{X}\} = E\{\tilde{\Lambda}\}$ and $E\{X\} = E\{\Lambda\}$. For the first part, it follows from $e^{-\lambda}\lambda^x \leq (x/e)^x$ for all $\lambda \in \mathbb{R}^+$ and the dominated convergence theorem that

$$h_{\tilde{Q}}(x) = \int_0^B e^{-\lambda}\frac{\lambda^x}{x!}d\tilde{Q}(\lambda) \to \int_0^B e^{-\lambda}\frac{\lambda^x}{x!}dQ(\lambda) = h_Q(x).$$

Step 2. Now suppose $\{\tilde{Q}\}$ is a sequence of estimators for Q with randomness from the "measurement model" (3.1) and "expression model" (3.2). Then it can be proved that $W_1(\tilde{Q}, Q) \xrightarrow{p} 0$ implies $W_1(h_{\tilde{Q}}, h_Q) \xrightarrow{p} 0$ based on the result in **Step 1** and the fact that a sequence converging in probability is equivalent to that its every subsequence has a further subsequence that converges almost surely. To prove $E\{W_1(\tilde{Q}, Q)\} \to 0$ implies $E\{W_1(h_{\tilde{Q}}, h_Q)\} \to 0$, it suffices to verify that $E\{W_1(h_{\tilde{Q}}, h_Q)^2\}$ is bounded which follows immediately from an upper bound of $W_1(h_{\tilde{Q}}, h_Q)$ based on the supports of $\tilde{Q}$ and Q in Lemma 3.1, or specifically, $W_1(h_{\tilde{Q}}, h_Q) \leq E\{\tilde{X}\} + E\{X\} = E\{\tilde{\Lambda}\} + E\{\Lambda\} \leq 2B$.

Step 3. Suppose $\mathbb{Q}_B$ is a set consisting of all distributions on $[0, B]$. For any $Q_0 \in \mathbb{Q}_B$ with $E\{W_1(\hat{Q}, Q) \mid Q = Q_0\} \to 0$, it follows from **Step 2** that

$$E\{W_1(h_{\tilde{Q}}, h_Q) \mid Q = Q_0\} = E\{W_1(h_{\tilde{Q}}, h_{Q_0}) \mid Q = Q_0\} = E\{W_1(h_{\tilde{Q}}, h_{Q_0})\} \to 0,$$

where the expectation in the last term $E\{W_1(h_{\tilde{Q}}, h_{Q_0})\}$ is with respect to the randomness from the "measurement model" (3.1) and "expression model" (3.2) only. Then we can complete this proof by noting that $P(Q \in \mathbb{Q}_B) = 1$. □

Proof of Theorem 3.2. To simplify notations, we temporarily drop the subject index j and the group index k in this proof. A restatement of this theorem is accordingly as follows:

> *Assume $N = N_n \to \infty$ as $n \to \infty$, $r_i = r_{i,n} \in [\gamma_0, \gamma_1]$ are uniformly upper and lower bounded by two positive universal constants γ_0, γ_1, and $\mathcal{Q}$ is supported on $[0, B]$. We then have $E\{W_1(\hat{Q}, Q) \mid Q\} \to 0$ as $n \to \infty$ for almost all Q with regard to the measure $\mathcal{Q}$.*

This proof consists of two steps.

In the first step, we temporarily drop further the third-layer "population model" (3.3) and prove that for each fixed distribution Q supported on $[0, B]$ we have $E\{W_1(\hat{Q}, Q)\} \to 0$ as $n \to \infty$, where the expectation is with respect to randomness from the "measurement model" (3.1) and "expression model" (3.2). Proof of this step is an analogue of the proof of Theorem 2.3 in Chen (2017) with, however, two notable differences: (a) replacing Kiefer-Wolfowitz distance with Wasserstein-1 distance; (b) including heteroscedasticity by the read depths $r_{i,n}$. The first one can be addressed by noting that Kiefer-Wolfowitz distance (Chen, 2017, page 51) and Wasserstein-1 distance induce the same topology on $\mathbb{Q}_B$ since

$$e^{-B}\int_0^B |F_{Q_1}(\lambda) - F_{Q_2}(\lambda)|\mathrm{d}\lambda \le \int_0^B |F_{Q_1}(\lambda) - F_{Q_2}(\lambda)|e^{-\lambda}\mathrm{d}\lambda \le \int_0^B |F_{Q_1}(\lambda) - F_{Q_2}(\lambda)|\mathrm{d}\lambda.$$

The second one can be addressed by imposing the uniform positive lower and upper bounds of the read depths $r_{i,n}$. In the second step, the third-layer "population model" (3.3) gets involved and this proof is then closed by the same arguments in Step 3 of the proof of Theorem 3.1.

A detailed proof is put below for easy reference.

Step 1. The first step consists of three substeps. In the first substep, we prove that the set containing all distributions which are at least $\delta > 0$ far from Q can be covered by finite open balls in W_1 distance. In the second substep, with the aid of finite balls, we prove that, with probability converging to 1, no distributions that are at least δ far away from Q can maximize the likelihood function and hence the W_1 distance between $\hat{Q}$ and Q is less than δ. Then the W_1 consistency of $\hat{Q}$ follows immediately from picking a arbitrarily small δ.

Step 1(a). Let $\mathbb{Q}_B$ be a metric space consisting of all distributions supported on $[0, B]$ with the W_1 distance. For any $\delta > 0$, define

$$\mathcal{B}_\delta(Q) := \left\{Q' \in \mathbb{Q}_B : W_1(Q', Q) < \delta\right\}$$

and its complement is denoted by $\mathcal{B}^c_\delta(Q)$.

In the sequel, fix ϵ to be a small positive number. Suppose Q_1, Q_2 are two distributions on $[0, B]$ and F_{Q_1}, F_{Q_2} are their distribution functions. It then follows from

$$e^{-B}\int_0^B |F_{Q_1} - F_{Q_2}| \le \int_0^B |F_{Q_1}(\lambda) - F_{Q_2}(\lambda)|e^{-\lambda}\mathrm{d}\lambda \le \int_0^B |F_{Q_1} - F_{Q_2}|$$

that Kiefer-Wolfowitz distance Chen (2017, page 51) and Wasserstein-1 distance induce the same topology on $\mathbb{Q}_B$. Hence it follows from Chen (2017, page 54) that there exists a finite number of distributions $Q_j \in \mathbb{Q}_B, j \in [J]$, such that

$$\mathcal{B}^c_\delta(Q) \subset \bigcup_{j\in[J]} \mathcal{B}_\epsilon(Q_j).$$

Without loss of generality, it is assumed that Q_j is neither a deterministic distribution at 0 (in other words,

degenerate distribution at 0) nor Q for each $j \in [J]$.

Step 1(b). Let $Y_{j,\epsilon}(r) := \log\left\{1 + u\left(h_{r,Q}(H_{r,Q}^{-1}(U))/h_{r,\mathcal{B}_\epsilon(Q_j)}(H_{r,Q}^{-1}(U)) - 1\right)\right\}$, where $r \in [\gamma_0, \gamma_1]$, $u \in (0,1)$, $h_{r,Q}(x) := \int_0^B e^{-r\lambda}\frac{(r\lambda)^x}{x!}dQ(\lambda)$, $h_{r,\mathcal{B}_\epsilon(Q_j)}(x) := \sup_{Q' \in \mathcal{B}_\epsilon(Q_j)} h_{r,Q'}(x)$, U is a uniform random variable on $[0,1]$, and $H_{r,Q}^{-1}(\cdot)$ is a function such that $H_{r,Q}^{-1}(U) \sim h_{r,Q}$.

(i) We first prove that there exist constants $\epsilon_j > 0$ and $c_j > 0$ such that for all $\epsilon \le \epsilon_j$ and $r \in [\gamma_0, \gamma_1]$ we have $E\{Y_{j,\epsilon}(r)\} \ge c_j > 0$.

Note that $Y_{j,\epsilon}(r) \ge \log(1-u)$ and $\lim_{\epsilon \to 0^+} Y_{j,\epsilon}(r) = Y_{j,0^+}(r)$ almost surely, where

$$Y_{j,0^+}(r) := \log\left\{1 + u\left(h_{r,Q}(H_{r,Q}^{-1}(U))/h_{r,Q_j}(H_{r,Q}^{-1}(U)) - 1\right)\right\}.$$

Since $Y_{j,\epsilon}(r)$ is monotonically decreasing with respect to ϵ, it follows from the monotone convergence theorem that $\lim_{\epsilon \to 0^+} E\{Y_{j,\epsilon}(r)\} = E\{Y_{j,0^+}(r)\}$ for each $r \in [\gamma_0, \gamma_1]$. Moreover, it follows from Chen (2017, Lemma 2.5) that $E\{Y_{j,0^+}(r)\} > 0$ for each $r \in [\gamma_0, \gamma_1]$. Since $E\{Y_{j,0^+}(r)\}$ is a continuous function with respect to r and $r \in [\gamma_0, \gamma_1]$, there exists a positive constant c_j such that $E\{Y_{j,0^+}(r)\} \ge 2c_j$ for all $r \in [\gamma_0, \gamma_1]$. Furthermore, since $E\{Y_{j,\epsilon}(r)\}$ is a monotonically decreasing function with respect to ϵ, then it follows from Dini's theorem that $E\{Y_{j,\epsilon}(r)\}$ uniformly converges to $E\{Y_{j,0^+}(r)\}$ as $\epsilon \to 0^+$ on $r \in [\gamma_0, \gamma_1]$ and hence there exists a ϵ_j which doesn't depend on r such that for all $\epsilon \le \epsilon_j$ and $r \in [\gamma_0, \gamma_1]$ we have $|E\{Y_{j,\epsilon}(r)\} - E\{Y_{j,0^+}(r)\}| \le c_j$. Hence $E\{Y_{j,\epsilon}(r)\} \ge c_j > 0$ for all $\epsilon \le \epsilon_j$ and $r \in [\gamma_0, \gamma_1]$. Replacing r by $r_{i,n}$ for all $\epsilon \le \epsilon_j$ we have

$$E\{Y_{j,\epsilon}(r_{i,n})\} \ge c_j > 0 \text{ for all } i \in [N_n].$$

In the following arguments, set $\epsilon = \min_{j \in [J]}\{\epsilon_j\}$ and let $\mathcal{B}_j := \mathcal{B}_\epsilon(Q_j)$ for simplicity.

(ii) We then prove that there exists a constant C_j such that $\mathrm{var}\{Y_{j,\epsilon}(r)\} \le C_j < \infty$ for all $r \in [\gamma_0, \gamma_1]$.

Since $h_{r,\mathcal{B}_j}(x) \ge h_{r,Q_j}(x)$ and $\log\{1 + u(h_{r,Q}(x)/h_{r,\mathcal{B}_j}(x) - 1)\} \ge \log(1-u)$, we have $E\{Y_{j,\epsilon}^2(r)\} < \infty$ or equivalently

$$E\{\left(\log\{1 + u(h_{r,Q}(X_r)/h_{r,\mathcal{B}_j}(X_r) - 1)\}\right)^2\} < \infty,$$

where $X_r := H_{r,Q}^{-1}(U) \sim h_{r,Q}$, as long as $E\{(h_{r,Q}(X_r)/h_{r,Q_j}(X_r) - 1)^2\} < \infty$, or simply,

$$E\{(h_{r,Q}(X_r)/h_{r,Q_j}(X_r))^2\} < \infty.$$

To prove it, note that Q_j is not a deterministic distribution at 0 and hence there exist $\lambda_j \in (0, B]$ such that $F_{Q_j}(\lambda_j) < 1$. Then we have

$$h_{r,Q}(x) \le e^{-B\gamma_1}\frac{(B\gamma_1)^x}{x!} \text{ and } h_{r,Q_j}(x) \ge \left(1 - F_{Q_j}(\lambda_j)\right)e^{-\gamma_0\lambda_j}\frac{(\gamma_0\lambda_j)^x}{x!}$$

for sufficiently large x, and hence

$$\frac{h_{r,Q}(x)}{h_{r,Q_j}(x)} \leq \frac{e^{\gamma_0\lambda_j - B\gamma_1}}{\big(1 - F_{Q_j}(\lambda_j)\big)} \left(\frac{B\gamma_1}{\gamma_0\lambda_j}\right)^x.$$

Then $E\{(h_{r,Q}(X_r)/h_{r,Q_j}(X_r))^2\} < \infty$ follows immediately from the existence of the moment generating function of Poisson distribution. Since $E\{Y^2_{j,\epsilon}(r)\}$ is a continuous function with respect to r and $r \in [\gamma_0, \gamma_1]$, then there exists a uniform constant C_j such that

$$E\{Y^2_{j,\epsilon}(r)\} \leq C_j$$

for all $r \in [\gamma_0, \gamma_1]$. Replacing r by $r_{i,n}$ it follows that $E\{Y^2_{j,\epsilon}(r_{i,n})\} \leq C_j$ for all $i \in [N_n]$.

(iii) Suppose $X_{i,n}, i \in [N_n]$ is a sequence of independent random variables with $X_{i,n} \sim h_{r_{i,n},Q}$. Define $Z_{ij,n} := \log\Big\{1 + u\left(h_{r_{i,n},Q}(X_{i,n})/h_{r_{i,n},\mathcal{B}_\epsilon(Q_j)}(X_{i,n}) - 1\right)\Big\}$. Note that $Z_{ij,n} \stackrel{d}{=} Y_{j,\epsilon}(r_{i,n})$. Built on (i) and (ii), we have $E\{Z_{ij,n}\} \geq c_j > 0$ and $\mathrm{var}\{Z_{ij,n}\} \leq C_j < \infty$ for $i \in [N_n]$ and $j \in [J]$. Therefore,

$$\mathrm{var}\left\{\sum_{i\in[N_n]} Z_{ij,n}\right\} \leq N_n C_j$$

and hence $\mathrm{var}\left\{\sum_{i\in[N_n]} Z_{ij,n}\right\}/N_n^2 \to 0$ as $n \to \infty$. Then it follows from Markov's inequality that

$$\frac{1}{N_n}\sum_{i\in[N_n]} (Z_{ij,n} - E\{Z_{ij,n}\}) \stackrel{p}{\to} 0$$

for each $j \in [J]$. In other words, for any positive number ξ and events

$$A_{j,n} := \left|\frac{1}{N_n}\sum_{i\in[N_n]} (Z_{ij,n} - E\{Z_{ij,n}\})\right| \leq \xi,$$

we have $\lim_{n\to\infty} P(A_{j,n}) = 1$. Combined with $E\{Z_{ij,n}\} \geq c_j$, we have, under $A_{j,n}$ with $\xi \leq c_j/2$,

$$\frac{1}{N_n}\sum_{i\in[N_n]} \log\{1 + u\left(h_{r_{i,n},Q}(X_{i,n})/h_{r_{i,n},\mathcal{B}_j}(X_{i,n}) - 1\right)\} > c_j/2,$$

and hence

$$\begin{aligned}
0 &< \sum_{i\in[N_n]} \log\{1 + u\left(h_{r_{i,n},Q}(X_{i,n})/h_{r_{i,n},\mathcal{B}_j}(X_{i,n}) - 1\right)\} \\
&= \sum_{i\in[N_n]} \inf_{Q'\in\mathcal{B}_j} \log\left\{1 + u\left(h_{r_{i,n},Q}(X_{i,n})/h_{r_{i,n},Q'}(X_{i,n}) - 1\right)\right\} \\
&\leq \inf_{Q'\in\mathcal{B}_j} \sum_{i\in[N_n]} \log\{1 + u\left(h_{r_{i,n},Q}(X_{i,n})/h_{r_{i,n},Q'}(X_{i,n}) - 1\right)\}
\end{aligned}$$

for each $j \in [J]$. Here the equality follows from the definition of $h_{r_{i,n},\mathcal{B}_j}(X_{i,n}) = \sup_{Q'\in\mathcal{B}_j} h_{r_{i,n},Q'}(x)$ and

hence $1/h_{r_{i,n},\mathcal{B}_j}(X_{i,n}) = \inf_{Q'\in\mathcal{B}_j} 1/h_{r_{i,n},Q'}(x)$, and the last inequality follows from the fact that the sum of infima is smaller than the infimum of the sum. Noting that $\mathcal{B}_\delta^c(Q) \subset \bigcup_{j=1}^J \mathcal{B}_j$, the last display implies that under events $A_n := \bigcap_{j\in[J]} A_{j,n}$ with $\xi \le \min_{j\in[J]} c_j/2$ we have

$$0 < \inf_{Q'\notin\mathcal{B}_\delta(Q)} \sum_{i\in[N_n]} \log\{1 + u\left(h_{r_{i,n},Q}(X_{i,n})/h_{r_{i,n},Q'}(X_{i,n}) - 1\right)\},$$

or equivalently,

$$l_n(uQ + (1-u)Q') > l_n(Q')$$

for all $Q' \in \mathcal{B}_\delta^c(Q)$, where for each $Q' \in \mathbb{Q}_B$

$$l_n(Q') := \sum_{i\in[N_n]} \log \int_0^B e^{-r_{i,n}\lambda} \frac{(r_{i,n}\lambda)^{X_{i,n}}}{X_{i,n}!} dQ'(\lambda).$$

Therefore, under events A_n with $\xi \le \min_{j\in[J]} c_j/2$, the maximum likelihood estimator $\hat{Q}$ must belong to $\mathcal{B}_\delta(Q)$ and hence $W_1(\hat{Q}, Q) \le \delta$. Since $P(A_n) \to 1$, we have $P(W_1(\hat{Q}, Q) \le \delta) \to 1$, or equivalently, $W_1(\hat{Q}, Q) \xrightarrow{p} 0$ as $n \to \infty$. It further follows from $W_1(\hat{Q}, Q) \le B$ that $E\{W_1(\hat{Q}, Q)\} \to 0$ as $n \to \infty$.

Step 2. For any $Q_0 \in \mathbb{Q}_B$, it follows from **Step 1** that

$$E\{W_1(\hat{Q}, Q) \mid Q = Q_0\} = E\{W_1(\hat{Q}, Q_0) \mid Q = Q_0\} = E\{W_1(\hat{Q}, Q_0)\} \to 0,$$

where the expectation in the last term $E\{W_1(\hat{Q}, Q_0)\}$ is with respect to the randomness from the "measurement model" (3.1) and "expression model" (3.2) only. Then it follows from $P(Q \in \mathbb{Q}_B) = 1$ that $E\{W_1(\hat{Q}, Q) \mid Q\} \to 0$ for almost all Q with regard to the measure $\mathcal{Q}$. □

Proof of Theorem 3.3. By the construction of the population model (3.3), under the H_0 in (3.5) $Q_j^{(k)}$'s are independent and identically distributed. Furthermore, since the sets $\{r_{ij}^{(k)}, i \in [N]\}$ are invariant with respect to $j \in [n_k]$ and $k \in [K]$ for each $i \in [N]$, the random vectors $(X_{1j}^{(k)}, \ldots, X_{Nj}^{(k)})^\top$'s are independent and identically distributed. Therefore, $\tilde{Q}_j^{(k)}$'s are independent and identically distributed. As a consequence, the distribution of $\tilde{F}$ is the empirical distribution of

$$\tilde{\mathcal{F}}^\pi := \left\{\tilde{F}^\pi : \pi \in \text{ all permutations of } [n] \to [n]\right\},$$

and hence $P(\tilde{F} > \text{ the } 1-\alpha \text{ quantile of } \tilde{\mathcal{F}}^\pi \mid H_0) \le \alpha$. Since the event $\tilde{F} > \text{ the } 1-\alpha$ quantile of $\tilde{\mathcal{F}}^\pi$ is identical to the event $P(\tilde{F}^\pi < \tilde{F} \mid \tilde{Q}_j^{(k)}\text{'s}) \ge 1-\alpha$, where probability here is with respect to π, we have $P(\tilde{T}_\alpha = 1|H_0) \le \alpha$. The proof of $P(\tilde{T}_{h,\alpha} = 1|H_0) \le \alpha$ is analogous and hence omitted. □

Proof of Theorem 3.4(Preparation). Let

$$F := \Big(SS_T - \sum_{k\in[K]} SS_k\Big) \Big/ \sum_{k\in[K]} SS_k, \tag{3.14}$$

where

$$SS_T := \frac{1}{n} \sum_{k_1,k_2\in[K]} \sum_{j_1\in[n_{k_1}],j_2\in[n_{k_2}]} W_1(Q_{j_1}^{(k_1)}, Q_{j_2}^{(k_2)})^2 \quad \text{and} \quad SS_k := \frac{1}{n_k} \sum_{j_1,j_2\in[n_k]} W_1(Q_{j_1}^{(k)}, Q_{j_2}^{(k)})^2.$$

For any permutation $\pi : [n] \to [n]$, let

$$F^\pi := \Big(SS_T - \sum_{k\in[K]} SS_k^\pi\Big) \Big/ \sum_{k\in[K]} SS_k^\pi, \tag{3.15}$$

where

$$SS_k^\pi := \frac{1}{n_k} \sum_{j_1,j_2\in[n_k]} W_1\Big(Q_{\Pi_1^{j_1,k}}^{(\Pi_2^{j_1,k})}, Q_{\Pi_1^{j_2,k}}^{(\Pi_2^{j_2,k})}\Big)^2 \text{ for each } k \in [K].$$

These notations will be used in the sequel. □

Proof of Theorem 3.4(a). Throughout this proof, unless conditioning on certain events, the probability refers to randomness from all three layers as well as the permutation. Without loss of generality, it is assumed that $\mathcal{Q}_k$ is non-degenerate for each $k \in [K]$. Otherwise, the proof is analogous and omitted.

This proof consists of five steps. In the first step, we prove that if the following Equation (3.16) is true,

$$\lim_{n\to\infty} P\Big(\tilde{F} > \tilde{F}^\pi | H_1\Big) = 1, \tag{3.16}$$

then $\lim_{n\to\infty} P(\tilde{T}_\alpha = 1 \mid H_1) = 1$ for any $\alpha \in (0, 1)$. The rest four steps are devoted to proving Equation (3.16). Note that $\tilde{F} - \tilde{F}^\pi = (\tilde{F} - F) + (F - F^\pi) + (F^\pi - \tilde{F}^\pi)$, where F is defined in (3.14) and F^π is defined in (3.15). The second step proves that $\tilde{F} - F \xrightarrow{p} 0$ as $n \to \infty$. The third step proves that $F^\pi - \tilde{F}^\pi \xrightarrow{p} 0$ as $n \to \infty$. The fourth step proves that $F - F^\pi \xrightarrow{p}$ some strictly positive constant. In the fifth step, we combine results in Steps 2-4 to prove (3.16) and hence finish the proof of Theorem 3.4(a).

In the following the notion H_1 in the probability is abandoned as long as no confusion is possible.

Step 1. Note that

$$P(\tilde{T}_\alpha = 1) = P\Big\{P(\tilde{F}^\pi < \tilde{F} \mid \tilde{Q}_j^{(k)}\text{'s}) \geq 1 - \alpha\Big\} = 1 - P\Big\{P(\tilde{F}^\pi < \tilde{F} \mid \tilde{Q}_j^{(k)}\text{'s}) < 1 - \alpha\Big\},$$

$$P\Big\{P(\tilde{F}^\pi < \tilde{F} \mid \tilde{Q}_j^{(k)}\text{'s}) < 1 - \alpha\Big\} = P\Big\{P(\tilde{F}^\pi \geq \tilde{F} \mid \tilde{Q}_j^{(k)}\text{'s}) > \alpha)\Big\} \leq \frac{E\Big\{P(\tilde{F}^\pi \geq \tilde{F} \mid \tilde{Q}_j^{(k)}\text{'s})\Big\}}{\alpha},$$

and

$$E\Big\{P(\tilde{F}^{\pi} \geq \tilde{F} \mid \tilde{Q}_j^{(k)}\text{'s})\Big\} = P(\tilde{F}^{\pi} \geq \tilde{F}) = 1 - P(\tilde{F}^{\pi} < \tilde{F}).$$

Therefore, we have $\lim_{n\to\infty} P\{P(\tilde{F}^{\pi} < \tilde{F} \mid \tilde{Q}_j^{(k)}\text{'s}) < 1-\alpha\} = 0$ and hence $\lim_{n\to\infty} P(\tilde{T}_\alpha = 1) = 1$ as long as (3.16) holds.

Step 2. In this step, we prove that $\tilde{F} - F \xrightarrow{p} 0$ as $n \to \infty$, where the probability here refers to randomness from all three layers.

Note that

$$\left|\tilde{F} - F\right| = \left|\frac{SS_T}{\sum_{k\in[K]} SS_k} - \frac{\tilde{SS}_T}{\sum_{k\in[K]} SS_k} + \frac{\tilde{SS}_T}{\sum_{k\in[K]} SS_k} - \frac{\tilde{SS}_T}{\sum_{k\in[K]} \tilde{SS}_k}\right|,$$

where SS_k and SS_T are defined in (3.14). It then follows from the triangle inequality that

$$\left|\tilde{F} - F\right| \leq \left|\frac{SS_T - \tilde{SS}_T}{\sum_{k\in[K]} SS_k}\right| + \left|\frac{B^2}{\frac{1}{n-1}\sum_{k\in[K]} \tilde{SS}_k}\right| \left|\frac{\sum_{k\in[K]} (\tilde{SS}_k - SS_k)}{\sum_{k\in[K]} SS_k}\right|, \tag{3.17}$$

where $\frac{1}{n-1}\tilde{SS}_T \leq B^2$ follows from $W_1(\tilde{Q}_{j_1}^{(k_1)}, \tilde{Q}_{j_2}^{(k_2)}) \leq B^2$ for each j_1, j_2, k_1, k_2.

Step 2(a). We first prove that $\left|\frac{\tilde{SS}_k}{n_k-1} - \frac{SS_k}{n_k-1}\right| \xrightarrow{p} 0$ and $\frac{1}{n-1}\left|\tilde{SS}_T - SS_T\right| \xrightarrow{p} 0$.

It follows from the triangle inequality that for each $k \in [K]$

$$\left|\frac{\tilde{SS}_k}{n_k - 1} - \frac{SS_k}{n_k - 1}\right| \leq \frac{4B}{n_k} \sum_{j\in[n_k]} W_1(\tilde{Q}_j^{(k)}, Q_j^{(k)})$$

as well as

$$\left|\frac{\tilde{SS}_T}{n-1} - \frac{SS_T}{n-1}\right| \leq \frac{4B}{n} \sum_{k\in[K]} \sum_{j\in[n_k]} W_1(\tilde{Q}_j^{(k)}, Q_j^{(k)}).$$

Therefore,

$$E\left\{\left|\frac{\tilde{SS}_k}{n_k - 1} - \frac{SS_k}{n_k - 1}\right|\right\} \leq 4B \cdot \frac{1}{n_k} \sum_{j\in[n_k]} E\left\{W_1(\tilde{Q}_j^{(k)}, Q_j^{(k)})\right\} = 4B \cdot E\left\{W_1(\tilde{Q}_1^{(k)}, Q_1^{(k)})\right\},$$

where the last equality follows from Assumption 3.1. Combining $W_1(\tilde{Q}_1^{(k)}, Q_1^{(k)}) \leq B$ with

$$E\Big\{W_1(\tilde{Q}_1^{(k)}, Q_1^{(k)}) \mid Q_1^{(k)}\Big\} \xrightarrow{a.s.} 0,$$

it follows that $E\left\{W_1(\tilde{Q}_1^{(k)}, Q_1^{(k)})\right\} \to 0$ as $n \to \infty$. Analogously, we have $E\left\{\frac{|\tilde{SS}_T - SS_T|}{n-1}\right\} \to 0$ as $n \to \infty$. Therefore, in (3.17) we have $\frac{|\tilde{SS}_T - SS_T|}{n-1} \xrightarrow{p} 0$ and $\sum_{k\in[K]} \frac{|\tilde{SS}_k - SS_k|}{n_k - 1} \xrightarrow{p} 0$.

Step 2(b). We then prove that $\frac{1}{n_k-1} SS_k \xrightarrow{p}$ some strictly positive constant.

It follows from $|W_1(Q_{j_1}^{(k_1)}, Q_{j_2}^{(k_2)})| \leq B$ and the strong law of large numbers for U-statistics (Serfling, 1980,

Chapter 5.4) that

$$\frac{1}{n_k-1}SS_k \quad \overset{a.s.}{\to} \quad E\{W_1(Q_1^{(k)},Q_2^{(k)})^2\}. \tag{3.18}$$

If $E\{W_1(Q_1^{(k)},Q_2^{(k)})^2\}=0$, then $W_1(Q_1^{(k)},Q_2^{(k)})=0$ almost surely and hence $Q_1^{(k)}=Q_2^{(k)}$ almost surely, which implies that $\mathcal{Q}_k$ is degenerate.

Step 2(c). Built on Step 2(a) and Step 2(b), we have

$$\frac{1}{n_k-1}\tilde{SS}_k=\frac{1}{n_k-1}(\tilde{SS}_k-SS_k)+\frac{1}{n_k-1}SS_k \overset{p}{\to} E\Big\{W_1(Q_1^{(k)},Q_2^{(k)})^2\Big\}.$$

Therefore, Slutsky's theorem guarantees $|\tilde{F}-F| \overset{p}{\to} 0$, where the probability here refers to randomness from all three layers.

Step 3. In this step, we prove that $F^\pi-\tilde{F}^\pi \overset{p}{\to} 0$ as $n\to\infty$, where the probability here refers to randomness from all three layers as well as the permutation.

To prove it, it suffices to show that $\frac{1}{n_k-1}\left|\tilde{SS}_k^\pi-SS_k^\pi\right| \overset{p}{\to} 0$ and $\frac{1}{n_k-1}SS_k^\pi \overset{p}{\to}$ some strictly positive constant since

$$\left|\tilde{F}^\pi-F^\pi\right| \le \left|\frac{SS_T-\tilde{SS}_T}{\sum_{k\in[K]}SS_k^\pi}\right| + \left|\frac{B^2}{\frac{1}{n-1}\sum_{k\in[K]}\tilde{SS}_k^\pi}\right| \left|\frac{\sum_{k\in[K]}(\tilde{SS}_k^\pi-SS_k^\pi)}{\sum_{k\in[K]}SS_k^\pi}\right|, \tag{3.19}$$

where SS_k^π and SS_T^π are defined in (3.15).

Step 3(a). We first prove $\frac{1}{n_k-1}\left|\tilde{SS}_k^\pi-SS_k^\pi\right| \overset{p}{\to} 0$.

To prove it, note that with similar arguments in Step 2(a) we have

$$E\left\{\left|\frac{\tilde{SS}_k^\pi}{n_k-1}-\frac{SS_k^\pi}{n_k-1}\right|\right\} \quad \le \quad 4B\cdot\frac{1}{n_k}\sum_{j\in[n_k]}E\left\{W_1(\tilde{Q}^{(\Pi_2^{j,k})}_{\Pi_1^{j,k}},Q^{(\Pi_2^{j,k})}_{\Pi_1^{j,k}})\right\}.$$

Let $n_{k,k'}^\pi$ represent the number of indices exchanged between group k and k' after the specific permutation π. Note that $n_{k,k'}^\pi=n_{k',k}^\pi$ and $\sum_{k'}n_{k,k'}^\pi=n_k$. Then, with the aid of the notations $\{n_{k,k'}^\pi\}$, it follows that

$$\sum_{j\in[n_k]}E\left\{W_1(\tilde{Q}^{(\Pi_2^{j,k})}_{\Pi_1^{j,k}},Q^{(\Pi_2^{j,k})}_{\Pi_1^{j,k}})\Big|\pi\right\} \quad = \quad \sum_{k'\in[K]}n_{k,k'}^\pi E\left\{W_1(\tilde{Q}_1^{(k')},Q_1^{(k')})\right\}$$

and hence

$$\sum_{j\in[n_k]}E\left\{\frac{W_1(\tilde{Q}^{(\Pi_2^{j,k})}_{\Pi_1^{j,k}},Q^{(\Pi_2^{j,k})}_{\Pi_1^{j,k}})}{n_k}\right\}=\sum_{k'\in[K]}E\left\{\frac{W_1(\tilde{Q}_1^{(k')},Q_1^{(k')})}{K}\right\}.$$

Therefore, it follows from $E\left\{W_1(\tilde{Q}_1^{(k')},Q_1^{(k')})\right\}\to 0$ for each $k\in[K]$ that $E\left\{\left|\frac{\tilde{SS}_k^\pi}{n_k-1}-\frac{SS_k^\pi}{n_k-1}\right|\right\}\to 0$.

Step 3(b). We then prove $\frac{1}{n_k-1}SS_k^\pi \overset{p}{\to}$ some strictly positive constant.

To prove it, let $X \overset{d}{=} Y$ denote that the two random variables X, Y are identically distributed and note that

$$\frac{SS_k^\pi}{n_k - 1} \overset{d}{=} \underbrace{\sum_{k' \in [K]} \sum_{j_1, j_2 \in [n_{k,k'}^\pi]} \frac{W_1(Q_{j_1}^{(k')}, Q_{j_2}^{(k')})^2}{n_k(n_k-1)}}_{(a)} + \sum_{k_1' \neq k_2' \in [K]} \sum_{j_1 \in [n_{k,k_1'}^\pi], j_2 \in [n_{k,k_2'}^\pi]} \frac{W_1(Q_{j_1}^{(k_1')}, Q_{j_2}^{(k_2')})^2}{n_k(n_k-1)}. \tag{3.20}$$

(i) We first prove that the variance of (a) converges to 0.

Note that the variance of (a) equals to

$$\operatorname{var}\left\{ E\left\{ \sum_{j_1, j_2 \in [n_{k,k'}^\pi]} \frac{W_1(Q_{j_1}^{(k')}, Q_{j_2}^{(k')})^2}{n_k(n_k-1)} \Big| \pi \right\} \right\} + E\left\{ \operatorname{var}\left\{ \sum_{j_1, j_2 \in [n_{k,k'}^\pi]} \frac{W_1(Q_{j_1}^{(k')}, Q_{j_2}^{(k')})^2}{n_k(n_k-1)} \Big| \pi \right\} \right\},$$

where the first term equals to

$$\operatorname{var}\left\{ \frac{n_{k,k'}^\pi(n_{k,k'}^\pi - 1)}{n_k(n_k-1)} E\{W_1(Q_1^{(k')}, Q_2^{(k')})^2\} \right\} \leq B^4 \cdot \operatorname{var}\left\{ \frac{n_{k,k'}^\pi(n_{k,k'}^\pi - 1)}{n_k(n_k-1)} \right\}.$$

For the second term, note that

$$\operatorname{var}\left\{ \sum_{j_1 \neq j_2 \in [n_{k,k'}^\pi]} W_1(Q_{j_1}^{(k')}, Q_{j_2}^{(k')})^2 \;\Big|\; \pi \right\} \leq B^4 \cdot 2 n_{k,k'}^\pi (n_{k,k'}^\pi - 1)^2.$$

Therefore, the variance of (a) is upper bounded by

$$\frac{2B^4}{\left(n_k(n_k-1)\right)^2} \left[\operatorname{var}\{n_{k,k'}^\pi(n_{k,k'}^\pi - 1)\} + E\left\{(n_{k,k'}^\pi)^3\right\} \right] \to 0,$$

where the convergence follows from $n_{k,k'}^\pi / n_k \xrightarrow{p} 1/K$, $n_{k,k'}^\pi / n_k \leq 1$, the dominated convergence theorem such that

$$E\Big\{(n_{k,k'}^\pi)^3\Big\} \Big/ \left(n_k(n_k-1)\right)^2 \to E\{0\} = 0,$$

and

$$\frac{\operatorname{var}\{n_{k,k'}^\pi(n_{k,k'}^\pi - 1)\}}{\left(n_k(n_k-1)\right)^2} \to \frac{1}{K^4} - \frac{1}{K^4} = 0.$$

To prove $n_{k,k'}^\pi / n_k \xrightarrow{p} 1/K$, note that $E\{n_{k,k'}^\pi / n_k\} = 1/K$ and $\operatorname{var}\{n_{k,k'}^\pi / n_k\} = \operatorname{var}\{n_{1,1}^\pi / n_1\} = \frac{n_1^2(n-n_1)^2}{n_1^2 n^2 (n-1)} \to 0$ (cf. Chapuy (2007, page 460)).

(ii) We then prove that the expectation of (a) converges to $E\left\{W_1(Q_1^{(k')}, Q_2^{(k')})^2\right\}/K^2$. For this, we have

$$E\{(a)\} = E\left\{ \frac{n_{k,k'}^\pi(n_{k,k'}^\pi - 1)}{n_k(n_k-1)} E\left\{W_1(Q_1^{(k')}, Q_2^{(k')})^2\right\} \right\},$$

which converges to $E\left\{W_1(Q_1^{(k')}, Q_2^{(k')})^2\right\}/K^2$ by the dominated convergence theorem.

(iii) Built on (i) and (ii), it follows from Markov's inequality that $(a) \xrightarrow{p} E\left\{W_1(Q_1^{(k')}, Q_2^{(k')})^2\right\}/K^2$. Analo-

gously, we can prove that the second term in (3.20) converges to a constant in probability, i.e.,

$$\sum_{k_1' \neq k_2' \in [K]} \frac{1}{n_k(n_k-1)} \sum_{j_1 \in [n^\pi_{k,k_1'}], j_2 \in [n^\pi_{k,k_2'}]} W_1(Q_{j_1}^{(k_1')}, Q_{j_2}^{(k_2')})^2 \xrightarrow{p} \sum_{k_1' \neq k_2' \in [K]} \frac{E\{W_1(Q_1^{(k_1')}, Q_1^{(k_2')})^2\}}{K^2}.$$

As a result, we have

$$\frac{1}{n_k - 1} SS_k^\pi \xrightarrow{p} \sum_{k' \in [K]} \frac{E\left\{W_1(Q_1^{(k')}, Q_2^{(k')})^2\right\}}{K^2} + \sum_{k_1' \neq k_2' \in [K]} \frac{E\{W_1(Q_1^{(k_1')}, Q_1^{(k_2')})^2\}}{K^2} > 0. \tag{3.21}$$

Step 4. In this step we prove that $F - F^\pi \xrightarrow{p}$ some strictly positive constant, where the probability here refers to randomness from all three layers and permutations.

To prove it, note that

$$F - F^\pi = \frac{1}{n-1} SS_T \left(\frac{1}{\frac{1}{n-1} \sum_{k \in [K]} SS_k} - \frac{1}{\frac{1}{n-1} \sum_{k \in [K]} SS_k^\pi} \right).$$

Step 4(a). We first prove that $\frac{1}{n-1} SS_T \xrightarrow{p}$ some strictly positive constant. Note that

$$\frac{1}{n-1} SS_T \xrightarrow{p} \frac{1}{K^2} \sum_{k_1, k_2 \in [K]} E\left\{W_1(Q_1^{(k_1)}, Q_2^{(k_2)})^2\right\} > 0,$$

which follows from the strong law of large numbers for U-statistics (Serfling, 1980, Chapter 5.4).

Step 4(b). Built on Step 3(a), (3.18), and (3.21), it suffices to prove that $\sum_{k \in [K]} \frac{1}{n-1}(SS_k - SS_k^\pi) \xrightarrow{p}$ some strictly negative constant. It follows from (3.18) and (3.21) that

$$\begin{aligned} &\sum_{k \in [K]} \frac{1}{n_k - 1}(SS_k - SS_k^\pi) \\ &\quad \xrightarrow{p} (K-1) \left(\frac{1}{K} \sum_{k \in [K]} E\{W_1(Q_1^{(k)}, Q_2^{(k)})^2\} - \sum_{k_1' \neq k_2' \in [K]} \frac{E\{W_1(Q_1^{(k_1')}, Q_1^{(k_2')})^2\}}{K(K-1)} \right) < 0, \end{aligned}$$

where the last inequality follows from H_1 and hence

$$\sum_{k \in [K]} \frac{1}{n-1}(SS_k - SS_k^\pi) \xrightarrow{p} \frac{K-1}{K} \left(\frac{1}{K} \sum_{k \in [K]} E\{W_1(Q_1^{(k)}, Q_2^{(k)})^2\} - \sum_{k_1' \neq k_2' \in [K]} \frac{E\{W_1(Q_1^{(k_1')}, Q_1^{(k_2')})^2\}}{K(K-1)} \right),$$

which is a strictly negative constant.

Step 5. Building on the previous three steps, we have established that $\tilde{F} - \tilde{F}^\pi \xrightarrow{p} C$, where C is a strictly positive constant. Accordingly, we have $\lim_{n \to \infty} P(\tilde{F} > \tilde{F}^\pi \mid H_1) = 1$. □

Proof of Theorem 3.4(b). Noting Theorem 3.1 and Lemma 3.1, this is analogous to the proof of Theorem 3.4(a) and is hence omitted. □

3.7.2 Proof of theorems in Chapter 3.3

Proof of Theorem 3.5. The proof of the existence of $\hat{Q}$ is an analog of Simar (1976, Section 3.1). For the convergence of VDM, we refer to Böhning (1982) and arguments for VEM and ISDM are analogous. Although our model includes heteroscedasticity induced by read depths, their arguments are still valid after careful examination.

In the following we illustrate details using the VDM algorithm as an example. After understanding the proof of VDM, arguments for VEM and ISDM are straightforward and hence omitted. This proof consists of four steps. In the first step, we prove the existence of $\hat{Q}$. In the second step, we prove an important property (3.22) for proving $\Phi(G_L) \to \Phi(\hat{Q})$ as $L \to \infty$ if this algorithm doesn't stop. In the third step, we complete the proof in the case that this algorithm doesn't stop. In the fourth step, we complete the proof in the case that algorithm does stop at some L.

Step 1. This step gives a proof of the existence of $\hat{Q}$, which is an analogue of Simar (1976, Section 3.1).

Let $\bar{\mathbb{Q}}_B$ be the set of all sub-distributions (total mass less or equal to 1) on $[0, B]$ and let $\bar{\Gamma}_N := \{\boldsymbol{\mu}(\bar{G}) | \bar{G} \in \bar{\mathbb{Q}}_B\}$, where $\bar{G} \mapsto \boldsymbol{\mu}(\bar{G}) := \big(\mu_1(\bar{G}), \ldots, \mu_N(\bar{G})\big)$ and

$$\bar{G} \mapsto \mu_i(\bar{G}) := \int_0^B \exp(-\lambda r_i)(\lambda r_i)^{X_i} \mathrm{d}\bar{G}(\lambda) \text{ for } i \in [N] \text{ and } \bar{G} \in \bar{\mathbb{Q}}_B.$$

We claim that $\bar{\Gamma}_N$ is convex and compact. Convexity is obvious. Compactness follows from the weak compactness of $\bar{\mathbb{Q}}_B$, boundedness and continuity of $\lambda \mapsto \exp(-\lambda r_i)(\lambda r_i)^{X_i}$ on $[0, B]$, and Helly–Bray theorem, see Simar's arguments for further details. It further follows from the concavity of $(\mu_1, \ldots, \mu_N) \mapsto \Psi(\mu_1, \ldots, \mu_N) := \frac{1}{N}\sum_{i=1}^N \log \mu_i$ on $\bar{\Gamma}_N$ that there exists a unique maximizer $(\hat{\mu}_1, \ldots, \hat{\mu}_N)$ of Ψ on $\bar{\Gamma}_N$. By the construction of $\bar{\Gamma}_N$, there exists a sub-distribution $\bar{G}_{max} \in \bar{\mathbb{Q}}_B$ such that $(\hat{\mu}_1, \ldots, \hat{\mu}_N) = (\mu_1(\bar{G}_{max}), \ldots, \mu_N(\bar{G}_{max}))$. The proof of that $\bar{G}_{max}$ is actually a distribution follows from exactly same arguments by Simar (1976, Page 1202). Now we complete the proof of the existence of $\hat{Q}$.

Step 2. Let $\mathbb{Q}_B$ be the set of all distributions on $[0, B]$ and let δ_λ be the deterministic distribution at $\lambda \in [0, B]$. Since we have $\Phi(G) > -\infty$ for each $G \in \mathbb{Q}_B \backslash \{\delta_0\}$, we can define the following directional directive

$$\Phi'(G, \delta_\lambda) := \lim_{\epsilon \to 0^+} \epsilon^{-1}\Big\{\Phi\{(1-\epsilon)G \oplus \epsilon\delta_\lambda\} - \Phi(G)\Big\} = \frac{1}{N}\sum_{i \in [N]} \frac{e^{-\lambda r_i}(\lambda r_i)^{X_i}}{\mu_i(G)} - 1$$

for $G \in \mathbb{Q}_B \backslash \{\delta_0\}$ and $\lambda \in [0, B]$.

In the second step, we prove that for all $\nu > 0, \alpha \in \mathbb{R}$ there exists $\epsilon_0 = \epsilon_0(\nu, \alpha) \in (0, 1)$ such that

$$\Phi'(G, \delta_\lambda) \geq \nu \text{ implies } \Phi\{(1-\epsilon)G \oplus \epsilon\delta_\lambda\} - \Phi(G) \geq \epsilon\nu/2 \tag{3.22}$$

for all $\epsilon \in [0, \epsilon_0(\nu, \alpha)]$, all $G \in \Delta_\alpha := \{G \in \mathbb{Q}_B | \Phi(G) \geq \alpha\}$, and all $\lambda \in [0, B]$.

$\Psi, \boldsymbol{\mu}, \bar{\mathbb{Q}}_B$ and $\bar{\Gamma}_N$ are defined in Step 1. Since Ψ is continuously differentiable on $\bar{\Gamma}_N \backslash \{0\}$, it follows from the mean value theorem that

$$\begin{aligned}\Phi\{(1-\epsilon)G \oplus \epsilon\delta_\lambda\} - \Phi(G) &= \Psi\{(1-\epsilon)\boldsymbol{\mu}(G) + \epsilon\boldsymbol{\mu}(\delta_\lambda)\} - \Psi(\boldsymbol{\mu}(G)) \\ &= \epsilon\nabla\Psi\left\{(1-\xi\epsilon)\boldsymbol{\mu}(G) + \xi\epsilon\boldsymbol{\mu}(\delta_\lambda)\right\}^T \boldsymbol{\mu}(\delta_\lambda) - 1,\end{aligned}$$

where $\nabla\Psi$ denotes the gradient of Ψ, for some $\xi \in [0,1]$. Therefore,

$$\Phi\{(1-\epsilon)G \oplus \epsilon\delta_\lambda\} - \Phi(G) - \epsilon\Phi'(G,\delta_\lambda) = \epsilon\Big\{\nabla\Psi\left\{(1-\xi\epsilon)\boldsymbol{\mu}(G) + \xi\epsilon\boldsymbol{\mu}(\delta_\lambda)\right\} - \nabla\Psi(\boldsymbol{\mu}(G))\Big\}^T \boldsymbol{\mu}(\delta_\lambda).$$

Define $\mathcal{L}_{\alpha'} := \{\boldsymbol{\mu} \in \bar{\Gamma}_N : \Psi \geq \alpha'\}$ for $\alpha' \in \mathbb{R}$. Note that $\mathcal{L}_{\alpha'}$ is a compact set, on which $\nabla\Psi$ is uniformly continuous, for $\alpha' = \alpha - 1$. Since $\boldsymbol{\mu}(G) \in \mathcal{L}_\alpha$, we can find a sufficiently small $\epsilon_0 = \epsilon_0(\alpha, \nu)$ such that for all $\epsilon \in [0, \epsilon_0]$ we have $(1-\xi\epsilon)\boldsymbol{\mu}(G) + \xi\epsilon\boldsymbol{\mu}(\delta_\lambda) \in \mathcal{L}_{\alpha-1}$ and

$$\|\nabla\Psi\left\{(1-\xi\epsilon)\boldsymbol{\mu}(G) + \xi\epsilon\boldsymbol{\mu}(\delta_\lambda)\right\} - \nabla\Psi(\boldsymbol{\mu}(G))\| \leq \nu/(2S),$$

where $\|\cdot\|$ denotes the Euclidean norm and $S := \sup_{\boldsymbol{\mu}\in\bar{\Gamma}_N} \|\boldsymbol{\mu}\|$. Therefore we have

$$|\Phi\{(1-\epsilon)G \oplus \epsilon\delta_\lambda\} - \Phi(G) - \epsilon\Phi'(G,\delta_\lambda)| \leq \epsilon\nu/(2S) \cdot S = \epsilon\nu/2. \tag{3.23}$$

If the claim doesn't hold, i.e. $\Phi\{(1-\epsilon)G \oplus \epsilon\delta_\lambda\} - \Phi(G) < \epsilon\nu/2$, it follows from $-\Phi'(G,\delta_\lambda) \leq -\nu$ that

$$\Phi\{(1-\epsilon)G \oplus \epsilon\delta_\lambda\} - \Phi(G) - \epsilon\Phi'(G,\delta_\lambda) < \epsilon\nu/2 - \epsilon\nu = -\epsilon\nu/2,$$

which contradicts (3.23).

Step 3. In this step, we assume that VDM doesn't stop and we have $\Phi(G_L) \to \Phi(\hat{Q})$ as $L \to \infty$.

Note that $\Phi(G_L)$ is monotonically increasing and suppose $\lim_{L\to\infty}\Phi(G_L) = \Phi^+$. If $\Phi^+ < \Phi(\hat{Q})$, then we have

$$\Phi'(G_L, \delta_{\lambda_{\max}}) = \max_{\lambda\in[0,B]} \Phi'(G_L,\delta_\lambda) \geq \Phi'(G_L,\hat{Q}) \geq \Phi(\hat{Q}) - \Phi(G_L) \geq \Phi(\hat{Q}) - \Phi^+ \geq \nu > 0,$$

for some $\nu > 0$, where the first inequality follows from Simar (1976, Page 1204) and the second inequality follows from the concavity of $\epsilon \mapsto \Phi((1-\epsilon)G_L + \epsilon\hat{Q})$ with $\epsilon \in [0,1]$. Then it follows from the claim in Step 2 that

$$\Phi(G_{L+1}) - \Phi(G_L) \geq \Phi\{(1-\epsilon_0)G_L \oplus \epsilon_0\delta_{\lambda_{\max}}\} - \Phi(G_L) \geq \nu\epsilon_0/2 > 0, \tag{3.24}$$

which contradicts $\lim_{L\to\infty}\Phi(G_L) = \Phi^+$.

Step 4. In this step, we prove that if VDM stops at some L, then $\Phi(G_L) = \Phi(\hat{Q})$.

If $\Phi(G_L) < \Phi(\hat{Q})$, we then have

$$\max_{\lambda\in[0,B]} \Phi'(G_L, \delta_\lambda) \geq \Phi(\hat{Q}) - \Phi(G_L) > 0,$$

which contradicts the criterion for stopping this algorithm. □

3.7.3 Proof of theorems in Chapter 3.6

Proof of Theorem 3.6. The proof of this theorem is built on the techniques developed in (Vinayak et al., 2019, Section 4). Loosely speaking, $W_1(Q, \hat{Q})$ is upper bounded by three parts, i.e.,

$$W_1(Q, \hat{Q}) \leq \sup_{\ell\in \mathsf{Lip}(1)} \Bigg(2 \left\| \ell - \hat{\ell} \right\|_\infty + \sum_{x=0}^{\infty} b_x \left(h_Q(x) - h_Q^{obs}(x) \right) + \sum_{x=0}^{\infty} b_x \left(h_Q^{obs}(x) - h_{\hat{Q}}(x) \right) \Bigg), \tag{3.25}$$

where $h_Q^{obs}(x) := \sum_{i=1}^N I(x = X_i)/N$, $I(\cdot)$ is an indicator function, $\|\ell - \hat{\ell}\|_\infty := \sup_{\lambda\in[0,B]} |\ell(\lambda) - \hat{\ell}(\lambda)|$ and $\lambda \mapsto \hat{\ell}(\lambda) := \sum_{x=0}^{\infty} b_x \frac{\lambda^x e^{-\lambda}}{x!}$ with $b_x \in \mathbb{R}$. These three parts can be further upper bounded separately with the aid of Lemma 3.2 and Lemma 3.3. Before diving into the detailed proof, we first add two remarks.

(i) When applying Lemma 3.3(a) for the proof of Part (a) of this theorem, we use $k = k(N, B) = \frac{B}{\sqrt{e}} \exp\left(W\left(\frac{\sqrt{e}\log N}{8B} \right) \right)$, where $W(\cdot)$ is the Lambert W function. (ii) Since the supremum over 1-Lipschitz functions ℓ in $W_1(Q, \hat{Q})$ depends on $\hat{Q}$, the choice of ℓ depends on the data $X_1, \ldots, X_N$. It further follows from the dependence between ℓ and $\hat{\ell}$ that the coefficients b_x of $\hat{\ell}$ depend on the data $X_1, \ldots, X_N$, i.e., $b_x = b_x(X_1, \ldots, X_N)$.

Proof of Theorem 3.6(a). This proof consists of two steps, similar to Section 4 in Vinayak et al. (2019). In the first step, we prove that $W_1(Q, \hat{Q})$ can be upper bounded by three parts, see (3.25). In the second step, we upper bound these three parts separately with the aid of Lemma 3.2 as well as Lemma 3.3 and complete this proof.

Step 1. For $x = 0, 1, \ldots$, let $x \mapsto h_Q^{obs}(x)$ denote the sample proportion, i.e. $h_Q^{obs}(x) := \sum_{i=1}^N I(x = X_i)/N$, where $I(\cdot)$ is an indicator function. Recall that

$$W_1(Q, \hat{Q}) = \sup_{\ell\in \mathrm{Lip}_1} \int_0^B \ell \mathsf{d}(Q - \hat{Q}),$$

where Lip_1 represents all 1-Lipschitz functions on $[0, B]$ and ℓ is one of those 1-Lipschitz functions. Without loss of generality, it is assumed that $\ell(0) = 0$. The idea is to use the following function

$$\lambda \mapsto \hat{\ell}(\lambda) := \sum_{x=0}^{\infty} b_x \frac{\lambda^x e^{-\lambda}}{x!}, \text{ where } b_x \in \mathbb{R} \text{ and } \lambda \in [0, B],$$

to approximate the 1-Lipschitz function $\lambda \mapsto \ell(\lambda)$ and upper bound $W_1(Q, \hat{Q})$ by three parts. It follows from a straight-forward algebra that

$$\int_0^B \ell(\lambda)\mathsf{d}\left(Q(\lambda) - \hat{Q}(\lambda)\right) = \int_0^B \left(\ell(\lambda) - \hat{\ell}(\lambda)\right)\mathsf{d}\left(Q(\lambda) - \hat{Q}(\lambda)\right) + \int_0^B \sum_{x=0}^{\infty} b_x \frac{\lambda^x e^{-\lambda}}{x!}\mathsf{d}\left(Q(\lambda) - \hat{Q}(\lambda)\right)$$
$$\leq 2\left\|\ell - \hat{\ell}\right\|_\infty + \sum_{x=0}^{\infty} b_x \left(h_Q(x) - h_Q^{obs}(x)\right) + \sum_{x=0}^{\infty} b_x \left(h_Q^{obs}(x) - h_{\hat{Q}}(x)\right),$$

where $\|\ell - \hat{\ell}\|_\infty := \sup_{\lambda\in[0,B]} |\ell(\lambda) - \hat{\ell}(\lambda)|$, and hence

$$W_1(Q, \hat{Q}) \leq \sup_{\ell\in \mathsf{Lip}(1)} \left(2\left\|\ell - \hat{\ell}\right\|_\infty + \sum_{x=0}^{\infty} b_x \left(h_Q(x) - h_Q^{obs}(x)\right) + \sum_{x=0}^{\infty} b_x \left(h_Q^{obs}(x) - h_{\hat{Q}}(x)\right)\right).$$

Step 2. It follows from upper bounds in Lemma 3.2 that for an arbitrary $\delta \in (0, 1/2)$ and an arbitrary $\epsilon \in (0, 1)$ there exists constants $N(\epsilon)$ and $C(\epsilon)$ depending only on ϵ such that the sum of the last two terms in (3.25) is upper bounded by $C(\epsilon)\sup_{x\geq 0}|b_x|\sqrt{\frac{B\vee 1}{N^{1-\epsilon}\delta^{1+\epsilon}}}$ for all $N \geq N(\epsilon)$ with probability at least $1 - 2\delta$.

Step 2(a). Suppose $c_0 \leq 0.001$. It follows from an approximation of $\ell(\cdot)$ in Proposition 3.3(a) that $\ell(\lambda)$ can be approximated by $\hat{\ell}(\lambda) = \sum_{x=0}^{k} b_x \frac{\lambda^x e^{-\lambda}}{x!}$ with an uniform approximation error of $C_1 B/k$ with $\max_x |b_x| \leq C_1 (\sqrt{e}k/B)^k$ for $k \geq 4(B \vee 1)$, where $C_1 > 1$ is a universal constant. Hence we have

$$W_1(Q, \hat{Q}) \leq 2C_1\frac{B}{k} + C_1 C(\epsilon)\left(\frac{\sqrt{e}k}{B}\right)^k \sqrt{\frac{B\vee 1}{N^{1-\epsilon}}\frac{1}{\delta^{1+\epsilon}}},$$

for $N \geq N(\epsilon)$ and $k \geq 4(B \vee 1)$ with probability at least $1 - 2\delta$. Taking $k = k(N, B)$ satisfying $(\sqrt{e}k/B)^k = N^{1/8}$, it follows that

$$W_1(Q, \hat{Q}) \leq 2C_1 B/k + C_1 C(\epsilon) N^{-3/8+\epsilon/2}\sqrt{(B\vee 1)/\delta^{1+\epsilon}}. \tag{3.26}$$

To verify $k(N, B) \geq 4(B \vee 1)$, note that $(\sqrt{e}k/B)^k = N^{1/8}$ is equivalent to

$$\log\left(\sqrt{e}k/B\right)\exp\{\log\left(\sqrt{e}k/B\right)\} = (\sqrt{e}/8 \cdot \log N)/B.$$

It further follows from $(\sqrt{e}c\log N)/B > 0$ that $k(N, B)$, as the solution of $(\sqrt{e}k/B)^k = N^{1/8}$, can be written using the Lambert W function, i.e. $k(N, B) = \frac{B}{\sqrt{e}}\exp\left(W\left(\frac{\sqrt{e}/8\cdot\log N}{B}\right)\right)$, where W is the Lambert W function. It follows from Hoorfar and Hassani (2008, Theorem 3.1) that $\exp(W(x)) \geq x/\log x$ for $x \geq e$. Noting that $\frac{\sqrt{e}/8\cdot\log N}{B} \geq \frac{\sqrt{e}/8\cdot\log N}{c_0\cdot\log N} \geq \frac{\sqrt{e}/8\cdot\log N}{0.001\cdot\log N} > e$, it follows immediately that

$$k(N, B) = \frac{B}{\sqrt{e}}\exp\left(W\left(\frac{\sqrt{e}/8\cdot\log N}{B}\right)\right) \geq \frac{\log N}{8\log\left(\frac{\log N}{B}\right)}. \tag{3.27}$$

Therefore,

$$\frac{k(N, B)}{B} \geq \frac{\log N}{8B}/\log\frac{\log N}{B} \geq \frac{1}{8}\frac{1000}{\log 1000} \geq 4,$$

since $B \le c_0 \log N$ with $c_0 \le 0.001$. If $\frac{1}{8}\frac{\log N}{\log \frac{\log N}{B}} \ge 4$ doesn't hold, then $E\{W_1(Q,\hat{Q})\} \le B \le \frac{32B}{\log N}\log\left(\frac{\log N}{B} \vee e\right)$ and hence Theorem 3.6(a) trivially holds. Therefore without loss of generality we assume that $\frac{1}{8}\frac{\log N}{\log \frac{\log N}{B}} \ge 4$ and hence $k(N,B) \ge 4$. As a consequence, we finally prove $k(N,B) \ge 4(B \vee 1)$.

Combining (3.26) with (3.27) and letting $\epsilon = 1/4$, we have

$$W_1(Q,\hat{Q}) \le 16C_1\frac{B\log\frac{\log N}{B}}{\log N} + C_1C(\epsilon)|_{\epsilon=1/4}\cdot N^{-1/4}\sqrt{\frac{B\vee 1}{\delta^{1+\epsilon}}},$$

where $C(\epsilon)|_{\epsilon=1/4}$ means the value of the function $\epsilon \mapsto C(\epsilon)$ at $1/4$. Therefore, for an arbitrary $\delta \in (0,1/2)$, there exists a universal constant C_3 such that for sufficiently large N we have

$$W_1(Q,\hat{Q}) \le C_3\frac{B}{\log N}\left(\log\frac{\log N}{B}\right)\frac{1}{\delta^{5/8}},$$

with probability at least $1-2\delta$. Therefore, for sufficiently large N we have

$$E\{W_1(Q,\hat{Q})\} \le 5C_3\frac{B}{\log N}\log\frac{\log N}{B} \le 5C_3\frac{B}{\log N}\log\left(\frac{\log N}{B}\vee e\right).$$

Step 2(b). Suppose $c_0 > 0.001$. Then for $B \in [0.001\cdot\log N, c_0\log N]$, it follows from Theorem 3.6(b) that $E\{W_1(Q,\hat{Q})\} \le C_4\sqrt{B/\log N} \le C_4\sqrt{c_0}$, where C_4 is a universal constant. On the other hand, in this case $\frac{B}{\log N}\log\left(\frac{\log N}{B}\vee e\right) \ge 0.001$ and hence

$$E\{W_1(Q,\hat{Q})\} \le \max\left\{5C_3, \frac{C_4\sqrt{c_0}}{0.001}\right\}\frac{B}{\log N}\log\left(\frac{\log N}{B}\vee e\right)$$

holds for all $B \le c_0\log N$. □

Proof of Theorem 3.6(b). Since $B \ge c_0\log N$, we have $B \ge 1$ for sufficiently large N. It follows from Step 1 in the proof of Theorem 3.6(a) and Lemma 3.2 that for an arbitrary $\delta \in (0,1/2)$ and an arbitrary $\epsilon \in (0,1)$ there exist constants $N(\epsilon)$ and $C(\epsilon)$ depending only on ϵ such that the sum of the last two terms in (3.25) is upper bounded by $C(\epsilon)\max_{x\ge 0}|b_x|\sqrt{\frac{B}{N^{1-\epsilon}\delta^{1+\epsilon}}}$ for all $N \ge N(\epsilon)$ with probability at least $1-2\delta$.

If $c_0 \ge 100$, it follows further from an approximation of $\ell(\cdot)$ in Lemma 3.3(b) that for sufficiently small ϵ there exists a constant $C_1 = C_1(\epsilon)$ such that

$$W_1(Q,\hat{Q}) \quad \le \quad C_1\left(\sqrt{\frac{B}{\log N}} + B^{3/2}N^{-1/2+2\epsilon}\sqrt{\frac{1}{\delta^{1+\epsilon}}}\right),$$

with probability at least $1-2\delta$. Since $B^3 \le C_0^3 N^{1-3\epsilon_0}$, then it follows from choosing $\epsilon = (\epsilon_0/2)\wedge 0.01$ that there exists a constant $C_2 = C_2(\epsilon_0)$ such that

$$W_1(Q,\hat{Q}) \le C_2\sqrt{\frac{B}{\log N}\frac{1}{\delta^{1+\epsilon}}} \text{ and hence } E\{W_1(Q,\hat{Q})\} \le C_3\sqrt{\frac{B}{\log N}},$$

where $C_3 = C_3(\epsilon_0)$ is a constant.

If $c_0 < 100$, a 1-Lipschitz function on $[0, B]$ can also be viewed as a Lipschitz function on $[0, 100 \log N]$ and hence it follows from letting $B = 100 \log N$ in Lemma 3.3(b) that for sufficiently small ϵ there exists a constant $C_4 = C_4(\epsilon)$ such that with probability $1 - 2\delta$

$$W_1(Q, \hat{Q}) \le C_4 \left(1 + \sqrt{B} N^{-1/2+2\epsilon} \log N \cdot \sqrt{\frac{1}{\delta^{1+\epsilon}}}\right) \le C_4 \left(1 + \sqrt{C_0} N^{-1/3+2\epsilon} \log N \cdot \sqrt{\frac{1}{\delta^{1+\epsilon}}}\right).$$

Therefore it follows from letting $\epsilon = 0.01$ that for sufficiently large N

$$E\{W_1(Q, \hat{Q})\} \le 2C_4 \le \frac{2C_4}{\sqrt{c_0}} \sqrt{\frac{B}{\log N}},$$

where the last inequality follows from $B \ge c_0 \log N$. □

Combining the above two sections complete the proof. □

Proof of Theorem 3.7. The proof of this theorem is built on the Le Cam minimax lower bound framework and it is analogous to the proof of Theorem 3.3 in Vinayak et al. (2019).

First, it follows from Proposition 4.3 in Vinayak et al. (2019) that there exist two random variables Λ_1 and Λ_2 supported on $[a - M, a + M]$ such that $E\{\Lambda_1^j\} = E\{\Lambda_2^j\}$, $0 \le j \le L$, where $a \ge M \ge 0$ are constants and L is a positive integer. Let Q_1 and Q_2 be the distributions of Λ_1 and Λ_2, respectively. Combining a lower bound of $W_1(Q_1, Q_2)$ in Vinayak et al. (2019, Proposition 4.3) with a upper bound of $\mathrm{TV}(h_{Q_1}, h_{Q_2})$ in Proposition 3.2, it comes to the end with properly choosing the value of a, M and L. For Part (a), let $a = C_1 B$, $M = B$ and $L = (Be^2/C_1) \cdot \exp\{W[4C_1 \log(N)/(e^2 B)]\} - 1$, where $C_1 \ge 1$ is a constant depending on c_0 and $W(\cdot)$ is the Lambert W function. For Part (b), let $a = B/(4e^4 c_0)$, $M = \sqrt{B \log N}/(4e^4 \sqrt{c_0})$ and $L = \log N$.

A proof to include the details is put below for easy reference.

Proof of Theorem 3.7(a). Suppose $a \ge M \ge 0$ are constants, and Λ_P and Λ_Q are two random variables supported on $[a - M, a + M]$ with distributions P and Q. Furthermore, assume $E\{\Lambda_P^j\} = E\{\Lambda_Q^j\}$, $0 \le j \le L$. Existence of P and Q is guaranteed by Proposition 4.3 in Vinayak et al. (2019).

For $0 < B \le c_0 \log N$, setting $a = C_1 B$, $M = B$ and

$$(L + 1)/2 = Be^2/(2C_1) \cdot \exp(W(4C_1 \log(N)/(e^2 B))),$$

where $W(\cdot)$ is the Lambert W function, $C_1 = \max\{1, 4e^2 c_0, C_2 c_0 e^2/4, C_3 c_0 e^2/4\}$ and C_2, C_3 are universal positive constants specified later. Since $W(x) = \log x - \log \log x + o(1)$ as $x \to \infty$, there exists a universal

constant C_2 such that for $x \geq C_2$ we have $W(x) \geq \frac{1}{2}\log x$. Therefore, it follows from

$$4C_1\frac{\log N}{e^2 B} \geq 4C_2c_0\frac{e^2}{4}\frac{\log N}{e^2 B} \geq 4C_2c_0\frac{e^2}{4}\frac{\log N}{e^2c_0\log N} = C_2$$

that

$$L+1 = \frac{Be^2}{C_1}\cdot\exp\{W(4C_1\log(N)/(e^2B))\} \geq \frac{Be^2}{C_1}\sqrt{4C_1\frac{\log N}{e^2B}} \geq \frac{Be^2}{C_1}\sqrt{4C_1\frac{\log N}{e^2c_0\log N}} \geq \frac{4Be^2}{C_1}.$$

By $(2eM)^2/a = (2eB)^2/(C_1B) = 4e^2B/C_1$, it follows that $L+1 \geq (2eM)^2/a$. Hence it follows from Proposition 3.2 that

$$\mathsf{TV}(P,Q) \quad \leq \quad 2\left(\frac{eB}{\sqrt{C_1B(L+1)}}\right)^{L+1} = 2\left(\frac{e^2B}{2C_1(L+1)/2}\right)^{\frac{L+1}{2}} = 2N^{-2},$$

where the last equality follows from the definition of the Lambert W function (see the proof of Theorem 3.6(a) for details). It follows from the LeCam minimax lower bound that for $N \geq 3$

$$\inf_{\tilde{Q}}\sup_{Q} E\{W_1(Q,\tilde{Q})\} \geq \frac{1}{2}W_1(P,Q)(1-\mathsf{TV}(P_N,Q_N)) \geq \frac{1}{2}W_1(P,Q)\left(1-2NN^{-2}\right) \geq \frac{1}{6}W_1(P,Q).$$

On the other hand, it follows immediately from Proposition 4.3 in Vinayak et al. (2019) that $W_1(P,Q) \geq 2M/(2L) = B/L$. Since $W(x) = \log x - \log\log x + o(1)$ as $x \to \infty$, there exists a universal constant C_3 such that for $x \geq C_3$ we have $W(x) \leq 1 + \log x - \log\log x$. Therefore, it follows from

$$4C_1\frac{\log N}{e^2 B} \geq 4C_3c_0\frac{e^2}{4}\frac{\log N}{e^2 B} \geq 4C_3c_0\frac{e^2}{4}\frac{\log N}{e^2c_0\log N} = C_3$$

that

$$L \leq \frac{Be^2}{C_1}\cdot\exp\{W(4C_1\log(N)/(e^2B))\} \leq \frac{Be^3}{C_1}\frac{4C_1\log N}{e^2B}/\log\frac{4C_1\log N}{e^2B} = 4e\log N/\log\frac{4C_1\log N}{e^2B}.$$

Therefore,

$$\inf_{\tilde{Q}}\sup_{Q} E\{W_1(Q,\tilde{Q})\} \geq \frac{1}{6}W_1(P,Q) \geq \frac{1}{6}\frac{B}{4e\log N}\log\frac{4C_1\log N}{e^2B} \geq \frac{B}{24e\log N}\log\frac{16c_0\log N}{B}.$$

This completes the proof. □

Proof of Theorem 3.7(b). Suppose $a \geq M \geq 0$ are constants, and Λ_P and Λ_Q are two random variables supported on $[a-M, a+M]$ with distributions P and Q. Furthermore, assume $E\{\Lambda_P^j\} = E\{\Lambda_Q^j\}$, $0 \leq j \leq L$. Existence of P and Q is guaranteed by Proposition 4.3 in Vinayak et al. (2019).

For $B \geq c_0\log N$, setting $a = c_1B/\sqrt{c_0}$, $L = \log N$ and $M = c_1\sqrt{B\log N}$ with $c_1 = 1/(4e^4\sqrt{c_0})$. Note that

$$\frac{a}{M} = \frac{c_1B/\sqrt{c_0}}{c_1\sqrt{B\log N}} = \sqrt{\frac{B}{c_0\log N}} \geq 1$$

and

$$\frac{(2eM)^2}{a} = \frac{4e^2c_1^2B\log N}{c_1B/\sqrt{c_0}} = \sqrt{c_0}4e^2c_1\log N = \frac{\sqrt{c_0}4e^2}{4e^4\sqrt{c_0}}\log N = \frac{1}{e^2}\log N \le 1+\log N = L+1.$$

Therefore, it follows from the LeCam minimax lower bound and an upper bound of total variation in Proposition 3.2 that

$$\begin{aligned}
\inf_{\tilde{Q}}\sup_{Q} E\{W_1(Q,\tilde{Q})\} &\ge \frac{1}{2}W_1(P,Q)(1-\mathsf{TV}(P_N,Q_N)) \\
&\ge \frac{1}{2}W_1(P,Q)\left(1-2N\left(\frac{ec_1\sqrt{B\log N}}{\sqrt{c_1B(1+\log N)/\sqrt{c_0}}}\right)^{1+\log N}\right) \\
&\ge \frac{1}{2}W_1(P,Q)\left(1-\frac{1}{e}N^{-\log 2}\right) \ge \frac{3}{10}W_1(P,Q).
\end{aligned}$$

On the other hand, it follows from Proposition 4.3 in Vinayak et al. (2019) that

$$W_1(P,Q) \ge \frac{2M}{2L} = \frac{c_1\sqrt{B\log N}}{\log N} = c_1\sqrt{\frac{B}{\log N}}.$$

Hence

$$\inf_{\tilde{Q}}\sup_{Q} E\{W_1(Q,\tilde{Q})\} \ge \frac{3c_1}{10}\sqrt{\frac{B}{\log N}} \ge \frac{3}{40e^4\sqrt{c_0}}\sqrt{\frac{B}{\log N}}.$$

This completes the proof. □

Combining the above two sections complete the proof. □

Lemma 3.1. *Suppose P and Q are two distributions supported on $[0,\infty)$. Then $W_1(P,Q) \le E\{\Lambda_P\} + E\{\Lambda_Q\}$ and $W_1(P,Q) \ge |E\{\Lambda_P\} - E\{\Lambda_Q\}|$, where Λ_P and Λ_Q are random variables with distributions P and Q respectively.*

Proof. Denote distribution functions of P and Q by F_P and F_Q respectively. Then it follows from the triangle inequality that $W_1(P,Q) = \int_0^\infty |(1-F_P)-(1-F_Q)| \le \int_0^\infty (1-F_P) + \int_0^\infty (1-F_Q) = E\{\Lambda_P\} + E\{\Lambda_Q\}$ and $W_1(P,Q) = \int_0^\infty |(1-F_P)-(1-F_Q)| \ge |E\{\Lambda_P\} - E\{\Lambda_Q\}|$. □

Lemma 3.2. *Suppose Q is a distribution on $[0,B]$ and $\{X_i, i\in[N]\}$ are N observations generated from h_Q defined in (3.12). For an arbitrary $\delta\in(0,1)$ and an arbitrary $\epsilon\in(0,1)$, there exist constants $N(\epsilon)>0$ and $C=C(\epsilon)>0$ such that for all $N\ge N(\epsilon)$, both*

$$\left|\sum_{x=0}^{\infty} b_x\left(h_Q^{obs}(x) - h_Q(x)\right)\right| \le C\cdot\sup_{x\ge 0}|b_x|\cdot\sqrt{\frac{B\vee 1}{N^{1-\epsilon}\delta^{1+\epsilon}}}$$

and

$$\left|\sum_{x=0}^{\infty} b_x \left(h_Q^{obs}(x) - h_{\hat{Q}}(x)\right)\right| \quad \leq \quad C \sup_{x\geq 0}|b_x| \sqrt{\frac{B \vee 1}{N^{1-\epsilon}\delta^{1+\epsilon}}}$$

hold with probability at least $1-\delta$.

Proof of Lemma 3.2. It follows from the triangle inequality and Pinsker's inequality in Proposition 3.1 that

$$\left|\sum_{x=0}^{\infty} b_x \left(h_Q^{obs}(x) - h_Q(x)\right)\right| \leq \sup_{x\geq 0}|b_x| \cdot \|h_Q^{obs} - h_Q\|_1 \leq \sup_{x\geq 0}|b_x| \cdot \sqrt{2 \cdot \mathrm{KL}(h_Q^{obs}, h_Q)},$$

where $\|h_Q^{obs} - h_Q\|_1 = \sum_{x\geq 0} |h_Q^{obs}(x) - h_Q(x)|$ and $\mathrm{KL}(h_Q^{obs}, h_Q) = \sum_{x\geq 0} h_Q^{obs}(x) \log[h_Q^{obs}(x)/h_Q(x)]$ is the Kullback–Leibler divergence between h_Q^{obs} and h_Q. On the other hand, analogous arguments imply

$$\left|\sum_{x=0}^{\infty} b_x \left(h_Q^{obs}(x) - h_{\hat{Q}}(x)\right)\right| \leq \sup_{x\geq 0}|b_x| \cdot \sqrt{2 \cdot \mathrm{KL}(h_Q^{obs}, h_{\hat{Q}})} \leq \sup_{x\geq 0}|b_x| \cdot \sqrt{2 \cdot \mathrm{KL}(h_Q^{obs}, h_Q)},$$

where the last inequality follows from that $\hat{Q}$ is the Kullback–Leibler divergence minimizer. Therefore, it suffices to upper bounding $\mathrm{KL}(h_Q^{obs}, h_Q)$.

Let $B_0 = B + 3\log(N/\delta) + \sqrt{3B\log(N/\delta)}$ with a constant $\delta \in (0,1)$. It follows from the union bound and Poisson tail inequality in Lemma 3.5 that

$$P\Big(\bigcup_{i\in[N]} \{X_i > B_0\}\Big) \leq N \cdot P(X_1 > B_0) \leq \delta.$$

In other words, with probability at least $1-\delta$ we have $h_Q^{obs}(x) = 0$ for all $x > B_0$ and

$$\sum_{x > B_0} h_Q^{obs}(x) \log[h_Q^{obs}(x)/h_Q(x)] = 0.$$

As a result, it suffices to upper bounding

$$\sum_{0\leq x\leq B_0} h_Q^{obs}(x) \log \frac{h_Q^{obs}(x)}{h_Q(x)} = \sum_{0\leq x\leq B_0} h_Q^{obs}(x) \log \frac{h_Q^{obs}(x)}{h_Q(x)/P(X\leq B_0)} + \log \frac{1}{P(X\leq B_0)} \tag{3.28}$$

conditional on $X_i \leq B_0$ for all $i \in [N]$, where $X \sim h_Q$. Under this condition, we have $X_1, \ldots, X_N \overset{i.i.d.}{\sim} X|X \leq B_0$, and then the first term in (3.28) can be dealt with using Mardia et al. (2019), i.e.,

$$\sum_{0\leq x\leq B_0} h_Q^{obs}(x) \log \frac{h_Q^{obs}(x)}{h_Q(x)/P(X\leq B_0)} \leq \frac{B_0}{2N} \log \frac{4N}{B_0} + \frac{1}{N} \log \frac{3e}{\delta}.$$

The second term in (3.28) can be upper bounded by Poisson tail inequality (Lemma 3.5), i.e.,

$$-\log P(X \leq B_0) \leq -\log(1-\delta/N) \leq 2/N$$

for $N \geq 2$. The proof is thus complete. □

Proposition 3.1. *(Pinsker's Inequality, see Cover and Thomas (2006).) For discrete distributions P and Q, it follows that*

$$\|P - Q\|_1 \leq \sqrt{2 \cdot KL(P, Q)},$$

where $\|P-Q\|_1$ is the total variation norm of the signed measure $P-Q$, and KL(P, Q) is the Kullback–Leibler divergence between P and Q.

Lemma 3.3.

(a) For any positive integer $k \geq 4(B \vee 1)$ and any 1-Lipschitz function $\lambda \mapsto \ell(\lambda)$ on $[0, B]$ with $\ell(0) = 0$, there exists an approximation $\hat{\ell}(\lambda) = \sum_{x=0}^{k} b_x \frac{\lambda^x e^{-\lambda}}{x!}$ such that

$$\sup_{\lambda \in [0,B]} |\hat{\ell}(\lambda) - \ell(\lambda)| \leq CB/k$$

and $\max_x |b_x| \leq C\left(\sqrt{e}k/B\right)^k$, where $C > 1$ is a universal constant.

(b) Suppose $B > 0$, $N \in \mathbb{N}^+$ and there exists constants $c_0, C_0 > 0$ such that $B \in [c_0 \log N, C_0 N]$. Then, for any fixed $c_0 \geq 100$ and any small $\epsilon \in (0, 0.02)$ there exist constants $C(\epsilon) > 0$ and $N(\epsilon) > 1$ and a sequence of coefficients $\{b_x\}_{x=0}^{\infty}$ such that for $N \geq N(\epsilon)$ any 1-Lipschitz function $\ell(\lambda)$ on $[0, B]$ with $\ell(0) = 0$ can be approximated by $\hat{\ell}(\lambda) = \sum_{x=0}^{\infty} b_x \frac{\lambda^x e^{-\lambda}}{x!}$ with an uniform approximation error of $C(\epsilon)\sqrt{\frac{B}{\log N}}$ with $\max_x |b_x| \leq C(\epsilon) B N^{\epsilon}$.

Remark 3.6. Part (a) in Lemma 3.3 can be obtained as a special case of the local step in (b); see Han and Shiragur (2021, proof of Theorem 3.4). However, here we give a different proof, which seems to be short and elegant.

Proof of Lemma 3.3 (a). The following two facts are used in our proof.

Fact 3.8 (Chapter 2.6 Equation 9 in Timan (2014))**.** *Suppose k is a non-negative integer and $\lambda \mapsto p_k(\lambda)$ is a polynomial function with coefficients $c_0, \ldots, c_k$, i.e. $p_k(\lambda) := \sum_{x=0}^{k} c_x \lambda^x$. Then it follows that coefficients $\{c_x\}_{x=0}^{k}$ satisfy*

$$|c_x| \leq \frac{k^x}{x!} \max_{|\lambda| \leq 1} |p_k(\lambda)|.$$

Fact 3.9 (Approximating e^{λ} with Taylor expansion)**.** *Let $\lambda \in [0, B]$. For any $k \geq 2B$, it follows that*

$$e^{\lambda} - \sum_{x=0}^{k} \frac{\lambda^x}{x!} = \sum_{x=k+1}^{\infty} \frac{\lambda^x}{x!} = \frac{\lambda^k}{k!} \sum_{x=1}^{\infty} \left(\frac{\lambda}{k+x} \cdots \frac{\lambda}{k+1} \right) \leq \frac{\lambda^k}{k!} \sum_{x=1}^{\infty} \frac{1}{2^x} = \frac{\lambda^k}{k!}$$

and hence

$$|e^{\lambda} - \sum_{x=0}^{k} \frac{\lambda^x}{x!}| / e^{\lambda} \leq \frac{\lambda^k}{k! e^{\lambda}} \leq \frac{B^k}{k! e^B}.$$

Applying Fact 3.9, it holds that for any $k \geq 2B$, there exists a polynomial $q_k(\lambda) = \sum_{x=0}^{k} \lambda^x/x!$ of degree k such that $|1 - q_k(\lambda)e^{-\lambda}| \leq B^k/(k!e^B)$ for all $\lambda \in [0, B]$. It is well known through Jackson's theorem (see Lemma 3.4) that for any 1-Lipschitz function $\ell(\cdot)$ on $[0, B]$, there exists a polynomial $p_k(\lambda)$ of degree k such that $\sup_{\lambda\in[-B,B]} |\ell(\lambda) - p_k(\lambda)| \leq C_1 B/k$, where $\ell(\lambda) := -\ell(-\lambda)$ for $\lambda < 0$ and $C_1 > 0$ is a universal constant independent of k and ℓ. Combining $p_k(\lambda)$, $q_k(\lambda)$ and the fact that $|p_k(\lambda)| \leq B + C_1 B/k \leq (1 + C_1)B$, it follows that for $\lambda \in [0, B]$

$$\begin{aligned}|p_k(\lambda)q_k(\lambda)e^{-\lambda} - \ell(\lambda)| &\leq |p_k(\lambda)(q_k(\lambda)e^{-\lambda} - 1)| + |p_k(\lambda) - \ell(\lambda)| \\ &\leq (1 + C_1)\frac{B}{k}\left(\frac{kB^k e^k}{\sqrt{k}k^k e^B} + 1\right),\end{aligned}$$

where the last inequality follows from $k! \geq \sqrt{k}\,(k/e)^k$ for $k \geq 2$ by Stirling's approximation. It further follows from the increasing monotonicity of $B \mapsto B^k/e^B$ for $B \leq k/2$ that

$$\sqrt{k}B^k e^k/(k^k e^B) \leq \sqrt{k}(k/2)^k e^k/(k^k e^{k/2}) = \sqrt{k}\left(\sqrt{e}/2\right)^k < 1,$$

where the last inequality holds for all $k \geq 2$, and hence

$$|p_k(\lambda)q_k(\lambda)e^{-\lambda} - \ell(\lambda)| \leq 2(1 + C_1)B/k$$

for $k \geq 2(B \vee 1)$. Therefore, we have shown that for any $k \geq 2(B \vee 1)$, there exists a function

$$\hat{\ell}(\lambda) = p_k(\lambda)q_k(\lambda)e^{-\lambda} = \sum_{x=0}^{2k} b_x \frac{\lambda^x e^{-\lambda}}{x!}$$

such that $|\hat{\ell}(\lambda) - \ell(\lambda)| \leq 2(1 + C_1)B/k$. For the bound on the coefficients b_x, let us first define the polynomial $r(\lambda) := p_k(B \cdot \lambda)q_k(B \cdot \lambda) = \sum_{x=0}^{2k} b'_x \frac{\lambda^x}{x!}$. Note that $b_x = b'_x/B^x$,

$$|r(\lambda)| \leq \left(B + 2(1 + C_1)B/k\right)e^B \leq 2(1 + C_1)Be^B$$

for $\lambda \in [0, 1]$ and

$$|r(\lambda)| \leq |q_k(B \cdot \lambda)|(1 + C_1)B \leq \left(1 + \frac{B^{k+1}}{(k + 1)!}\right)(1 + C_1)B \leq \left(1 + e/2e^B\right)(1 + C_1)B \leq 3(1 + C_1)Be^B$$

for $\lambda \in [-1, 0)$. Then we can apply Fact 3.8 for the polynomial $r(\lambda)$, which implies that

$$\frac{|b'_x|}{x!} \leq \frac{(2k)^x}{x!} \max_{|\lambda|\leq 1} |r(\lambda)| \leq \frac{(2k)^x}{x!} 3(1 + C_1)Be^B,$$

and hence

$$\max_x |b_x| = \max_x \frac{|b'_x|}{B^x} \leq \max_x \left(\frac{2k}{B}\right)^x 3(1 + C_1)Be^B = 3(1 + C_1)\left(\frac{2k}{B}\right)^{2k} Be^B \leq 3(1 + C_1)\left(\frac{2\sqrt{e}k}{B}\right)^{2k},$$

where the last inequality follows from $B \le k/2 \le \exp(k/2)$ and $e^B \le \exp(k/2)$. □

Proof of Lemma 3.3 (b). Since $B \ge c_0 \log N$, we have $B \ge 1$ for sufficiently large N. Note that $\lambda \mapsto \frac{1}{B}\ell(B\lambda)$ is a Lipschitz-1 function on $[0,1]$. By Proposition 3.3, it follows that there exists a sequence of coefficients $\{b_x\}_{x=0}^{\infty}$ such that

$$\left|\frac{1}{B}\ell(B\lambda) - \sum_{x=0}^{\infty} b_x P(\mathsf{Poi}(B\lambda) = x)\right| \le C(\epsilon)\sqrt{\frac{1}{B\log N}}, \text{ for any } \lambda \in [0,1],$$

where $b_x = 0$ for $x > 4B$, and

$$\left|b_x - \frac{1}{B}\ell\left(B \cdot \frac{x}{B}\right)\right| \le \frac{C(\epsilon)(1+x^{1/2})N^{\epsilon}}{B}, \text{ for } x \le 4B.$$

Defining $b_x^* = Bb_x$ and replacing $B\lambda$ by λ, it follows that

$$|\ell(\lambda) - \sum_{x=0}^{\infty} b_x^* P(\mathsf{Poi}(\lambda) = x)| \le C(\epsilon)\sqrt{\frac{B}{\log N}}, \text{ for any } \lambda \in [0,B],$$

where $b_x^* = 0$ for $x > 4B$, and

$$|b_x^* - \ell(x)| \le C(\epsilon)(1+x^{1/2})N^{\epsilon}, \text{ for } x \le 4B.$$

Moreover, It follows from the triangle inequality that $|b_x^*| \le 4B + C(\epsilon)(1+2B^{1/2})N^{\epsilon} = O(BN^{\epsilon})$. □

The following proposition is an extension of Wu and Yang (2016, Lemma 3); see also Wu and Yang (2020b, Section 3.3) for a nice survey.

Proposition 3.2 (Lemma 32 in Jiao et al. (2018))**.** *Suppose U_0, U_1 are two random variables supported on $[a-M, a+M]$, where $a \ge M \ge 0$ are constants. Suppose $E\{U_0^j\} = E\{U_1^j\}, 0 \le j \le L$. Denote the marginal distribution of X where $X|\lambda \sim Poi(\lambda)$, $\lambda \sim U_i$ as F_i, where $i = 0, 1$. If $L+1 \ge (2eM)^2/a$, then TV$(F_0, F_1) \le 2(eM/\sqrt{a(L+1)})^{L+1}$*

The following proposition is essentially Theorem 3.4 in Han and Shiragur (2021).

Proposition 3.3 (Theorem 3.4 of Han and Shiragur (2021))**.** *Suppose $B > 0$ and $N \in \mathbb{N}^+$ and there exists constants $c_0, C_0 > 0$ such that $B \in [c_0 \log N, C_0 N]$. Let $\ell(\cdot)$ be any 1-Lipschitz function on $\mathbb{R}$ with $\ell(0) = 0$. Then, for any fixed $c_0 \ge 96$ and any small $\epsilon \in (0, 0.02)$ there exist positive constants $C(\epsilon) > 0$ and $N(\epsilon) > 1$ depending on ϵ and a sequence of coefficients $\{b_x\}_{x=0}^{\infty}$ such that the following inequality holds for $N \ge N(\epsilon)$, i.e.*

$$|\ell(\lambda) - \sum_{x=0}^{\infty} b_x P(\mathit{Poi}(B\lambda) = x)| \le C(\epsilon)\sqrt{\frac{1}{B\log N}}, \lambda \in [0,1] \tag{3.29}$$

where $b_x = 0$ *for* $x > 4B$*, and*

$$\left| b_x - \ell\left(\frac{x}{B}\right) \right| \le C(\epsilon)(1 + x^{1/2})\frac{N^\epsilon}{B}, \text{ for any } x \le 4B. \tag{3.30}$$

Lemma 3.4 (Jackson's Theorem, Lemma 10 of Han and Shiragur (2021))**.** *Let* $k > 0$ *be any integer, and* $[a, b] \subseteq \mathbb{R}$ *be any bounded interval. For any Lipschitz-*1 *function* $\ell(\cdot)$ *on* $[a, b]$*, there exists a universal constant* C *independent of* k, ℓ *such that there exists a polynomial* $p_k(\cdot)$ *of degree at most* k *such that*

$$|\ell(\lambda) - p_k(\lambda)| \le C\sqrt{(b-a)(\lambda - a)}/k, \ \forall \lambda \in [a, b]. \tag{3.31}$$

In particular, the following norm bound holds:

$$\sup_{\lambda \in [a,b]} |\ell(\lambda) - p_k(\lambda)| \le C(b-a)/k. \tag{3.32}$$

Lemma 3.5 (Poisson tail inequality, Lemma 12 of Han and Shiragur (2021))**.** *For* $X \sim$ *Poi*(λ) *and any* $\delta > 0$*, we have*

$$P(X \ge (1+\delta)\lambda) \le \exp\left(-(\delta^2 \wedge \delta)\lambda/3\right) \quad \textit{and} \quad P(X \le (1-\delta)\lambda) \le \exp\left(-\delta^2\lambda/3\right).$$

3.8 Implementation details in Section 3.5

Let $n_1 = 13$ and $n_2 = 10$ denote the number of subjects in ASD and control groups respectively and $n = 23$ represent the number of total subjects. Since 99 percent of $\{X_{ij}^{(k)}/r_{ij}^{(k)}, i \in [N_{jk}], j \in [n_k], k \in [K]\}$ for 100 genes are smaller than 15.09, we choose $B = 20$ and $r_{ij}^{(k)}$'s as the summation of the particular cell's read count across all genes divided by 10,000, which is a conventional way to compute the read depth value of $r_{ij}^{(k)}$'s (Sarkar and Stephens, 2021). We use VEM to compute NPMLEs with a stop tolerance 0.01. The tests are conducted using the R package "*ideas*" by Sun and Zhang with 10^5 Monte Carlo simulations (Zhang et al., 2022).

To account for covariates, the pseudo-F statistics described in Chapter 3.2 has to be changed a little bit. Let $\mathbf{D}_n$ be the n by n distance matrix corresponding to the mixing distributions, with each entry equal to the squared W_1 distance between the two corresponding NPMLEs, and let

$$\mathbf{G}_n := \left(\mathbf{I}_n - \frac{1}{n}\mathbf{1}_n\mathbf{1}_n^\top\right)\mathbf{A}_n\left(\mathbf{I}_n - \frac{1}{n}\mathbf{1}_n\mathbf{1}_n^\top\right),$$

be the Grower's center matrix of $\mathbf{A}_n$, where $\mathbf{A}_n := -(1/2)\mathbf{D}_n$, $\mathbf{1}_n := (\underbrace{1, 1, \ldots, 1}_{n})^\top$, and $\mathbf{I}_n$ stands for the n-dimensional identity matrix. Note that $\mathbf{G}_n$ may have some negative eigenvalues, and we set those negative eigenvalues to 0. Let $\mathbf{Z}$ be an n by 5 matrix consisting of diagnostics (1 as ASD and 0 as control), age, sex,

seqbatch, and RIN. Let $\mathbf{H}_Z$ be the hat matrix $\mathbf{H}_Z := \mathbf{Z}(\mathbf{Z}^\top\mathbf{Z})^{-1}\mathbf{Z}^\top$. Then the new F-statistic accounting for covariates is

$$\hat{F}_Z := \frac{\mathrm{tr}(\mathbf{H}_Z\mathbf{G}\mathbf{H}_Z)}{\mathrm{tr}((\mathbf{I}-\mathbf{H}_Z)\mathbf{G}(\mathbf{I}-\mathbf{H}_Z))}, \tag{3.33}$$

where $\mathrm{tr}(\cdot)$ denotes the trace of a matrix. To implement the permutation test, we permute the variable "diagnostics" with all the rest covariates fixed and accordingly generate a new data matrix Z^π. The corresponding F-statistic is denoted by $\hat{F}_Z^\pi$ and the p-value is

$$\frac{\text{the number of permutations } \pi \text{ such that } \hat{F}_Z^\pi \geq \hat{F}_Z}{\text{the number of all possible permutations } \pi}. \tag{3.34}$$

When replacing the distance matrix $\mathbf{D}_n$ by the corresponding Poisson-smoothed version, the corresponding p-value is

$$\frac{\text{the number of permutations } \pi \text{ such that } \hat{F}_{h,Z}^\pi \geq \hat{F}_{h,Z}}{\text{the number of all possible permutations } \pi}, \tag{3.35}$$

where $\hat{F}_{h,Z}$ and $\hat{F}_{h,Z}^\pi$ are the Poisson-smoothed versions of $\hat{F}_Z$ and $\hat{F}_Z^\pi$.

4 Bayesian Ising mixture models

4.1 Introduction

Loglinear models have been widely used in the analysis of multivariate categorical data in many scientific fields, such as biological sciences, natural language processing, data mining (Bishop et al., 1975; Christensen, 1997) due to their ability to capture first, second and higher order interactions among the observed variables. They originated from testing for the absence of interactions in $2 \times 2 \times 2$ contingency tables (Bartlett, 1935), and were later generalized to multidimensional contingency tables (Roy and Kastenbaum, 1956; Darroch, 1962; Good, 1963; Goodman, 1963). In this paper we focus on Ising models (Kindermann and Snell, 1980) which can be viewed as graphical loglinear models for binary variables that include only first order interaction terms (Lauritzen, 1996). Ising models have particular relevance in network analysis of binary data (van Borkulo et al., 2014).

Various frequentist approaches for estimation of loglinear models have been proposed in the literature. For example, the existence and uniqueness of maximum likelihood estimators (MLEs) have been studied for different types of tables, e.g., three-way contingency tables (Birch, 1963), general contingency tables (Haberman, 1974; Aickin, 1979; Verbeek, 1992), and sparse contingency tables (Fienberg and Rinaldo, 2012). The computation of the MLEs can be performed using closed-form expressions (Bishop et al., 1975, Chapter 3.4), via iterative proportional fitting based on matrix inversion techniques (Goodman, 1964) or Newton-Raphson techniques (Haberman, 1974). Fienberg (2000) provides a comprehensive review. Bayesian approaches for loglinear modelling have also received a lot of interest, with a particular focus on the development of suitable prior distributions. Key examples include the multivariate normal prior (Knuiman and Speed, 1988; Dellaportas and Forster, 1999; Brooks and King, 2001; Dobra et al., 2006), the spike-and-slab prior (Ročková, 2018), the hyper Dirichlet conjugate prior (Dawid and Lauritzen, 1993) and its generalization, as well as the Diaconis–Ylvisaker (DY) conjugate prior (Massam et al., 2009). The use of the DY prior for model selection has been studied in Dobra and Massam (2010); Letac and Massam (2012).

Sparse contingency tables raise key issues related to estimation and fit for Ising models as well as for the larger family of loglinear models. Scalability of Ising models to discrete datasets with many variables has been solved in various ways – see, among others, Ravikumar et al. (2010). Sparse contingency tables are also characterized by the imbalance of their cell counts (Dobra and Lenkoski, 2011). Ising models together with the richer class of loglinear models tend to oversmooth the fitted cell probabilities which makes them unable to capture the magnitude of the larger cell counts. To this end, several classes of mixture models have been proposed as alternatives. Multivariate Bernoulli mixture models (Carreira-Perpinán and Renals,

2000; Allman et al., 2009) have shown promising results in applications (Juan and Vidal, 2002, 2004). Other classes of mixture models for categorical data include parallel factor analysis (PARAFAC) (Bro, 1997), simplex factor models (Bhattacharya and Dunson, 2012), sparse PARAFAC (Zhou et al., 2015), the Tucker decomposition (De Lathauwer et al., 2000), and the collapsed Tucker decomposition (Johndrow et al., 2017). While these mixture models perform well with respect to fit for sparse contingency tables, they are not easily interpretable especially when it comes to inferring relevant patterns of multivariate associations among discrete variables. We note two studies exploring the connections between the two modeling paradigms. Papathomas and Richardson (2016) leverages mixture models to reduce the number of parameters in a loglinear model, while Johndrow et al. (2017) delves into the relationship between dimension reduction in mixture models and loglinear models.

A common issue in finite mixture models is their identifiability (Teicher, 1967; Yakowitz and Spragins, 1968; Titterington et al., 1985). Identifiability has been studied for various types of mixture models such as uniform mixtures and binomial mixtures (Teicher, 1961), normal mixtures, exponential mixtures, Gamma mixtures (Teicher, 1963), Poisson mixtures (Teicher, 1960), and negative binomial mixtures (Titterington et al., 1985). For general mixture models, Teicher (1963) suggests using moment generating functions to prove identifiability, and Teicher (1967) presents sufficient conditions for the mixture of product densities. The most recent identifiability results are for multivariate Bernoulli mixture models through the conditional independence assumption (Allman et al., 2009; Xu, 2017).

The novel contributions of our work are as follows. In Chapter 4.2, we propose a novel Bayesian method for fitting finite mixtures of Ising models with the modeling goal of inferring associations between binary variables. In Chapters 4.3 and 4.4, our novel framework is illustrated through simulation experiments and real data applications. We provide sufficient and necessary conditions for identifiability of the model in Chapter 4.5 with proofs in Chapter 4.7. Finally, in Chapter 4.6, we discuss our results together with several potential extensions.

4.2 Ising Mixture Models

4.2.1 Notation

Let $\boldsymbol{X} := (X_1, \ldots, X_d)^T$ be a vector of $d \in \mathbb{N}^+$ binary random variables, each taking values of 0 or 1. The set of possible values for $\boldsymbol{X}$ is denoted as $I := \{0,1\}^d$ with elements $\boldsymbol{i} = (i_1, \ldots, i_d) \in I$ assumed to be order lexicographically. The vector of cell probabilities of $\boldsymbol{X}$ are $(P(\boldsymbol{X} = \boldsymbol{i}) : \boldsymbol{i} \in I)^T$.

Let the main effect of X_v be denoted by $\theta_v \in \mathbb{R}$, where $v \in [d] = \{1, 2, \ldots, d\}$. The interaction effect between $X_{v'}$ and X_v is denoted by $\theta_{v'v} \in \mathbb{R}$, where $v' < v$ and both $v', v \in [d]$. We say that $\boldsymbol{X}$ follows an

Ising model if the logarithm of the probability associated with cell $\boldsymbol{i} \in I$ is proportional to a linear combination of main effects $(\theta_v : v \in [d])^T$ and interaction effects $(\theta_{v'v} : v' < v)^T$:

$$\log p_{\boldsymbol{i}}(\boldsymbol{\theta}) = \sum_{v=1}^{d} \theta_v i_v + \sum_{v'=1}^{d-1} \sum_{v=v'+1}^{d} \theta_{v'v} i_{v'} i_v + C(\boldsymbol{\theta}),$$

where $C(\boldsymbol{\theta})$ is the logarithm of the normalizing constant, $\boldsymbol{\theta}$ represents the union of main effects and interaction effects, i.e., $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_d, \theta_{12}, \ldots, \theta_{(d-1)d})$ and $p_{\boldsymbol{i}}(\boldsymbol{\theta}) = P(\boldsymbol{X} = \boldsymbol{i} \mid \boldsymbol{\theta})$. If $\theta_{v'v} = 0$, variables $X_{v'}$ and X_v are conditionally independent given the rest. The main effects and the interaction terms can be interpreted as conditional log odds and log odds ratios (Agresti, 2002).

The Ising model can be expressed as

$$\boldsymbol{p}(\boldsymbol{\theta}) = (p_{\boldsymbol{i}}(\boldsymbol{\theta}) : \boldsymbol{i} \in I)^T = \exp(A^T \boldsymbol{\theta}) / [\mathbf{1}_{|I|}^T \exp(A^T \boldsymbol{\theta})], \tag{4.1}$$

where $\mathbf{1}_{|I|}$ is a $|I|$-dimensional constant vector of all ones, $A \in \mathbb{R}^{[d(d+1)/2] \times |I|}$ is a conventionally defined constant design matrix (Wang et al., 2019). Here applying the exponential function $\exp(\cdot)$ to a vector means applying it element-wise to obtain a vector. We illustrate the definition of A through an example.

Example 4.1. Suppose we have two binary random variables $\boldsymbol{X} = (X_1, X_2)^T$ that follow an Ising model with main effects θ_1, θ_2 and interaction effect θ_{12}. The following linear combination

$$\log p_{\boldsymbol{i}} = \log p_{(i_1, i_2)} = \theta_1 i_1 + \theta_2 i_2 + \theta_{12} i_1 i_2 + C(\boldsymbol{\theta}),$$

for $\boldsymbol{i} \in I = \{(0,0), (1,0), (0,1), (1,1)\}$, is equivalent to

$$\begin{pmatrix} \log p_{(0,0)} \\ \log p_{(1,0)} \\ \log p_{(0,1)} \\ \log p_{(1,1)} \end{pmatrix} - C(\boldsymbol{\theta}) = \begin{pmatrix} \theta_1 \cdot 0 + \theta_2 \cdot 0 + \theta_{12} \cdot 0 \cdot 0 \\ \theta_1 \cdot 1 + \theta_2 \cdot 0 + \theta_{12} \cdot 1 \cdot 0 \\ \theta_1 \cdot 0 + \theta_2 \cdot 1 + \theta_{12} \cdot 0 \cdot 1 \\ \theta_1 \cdot 1 + \theta_2 \cdot 1 + \theta_{12} \cdot 1 \cdot 1 \end{pmatrix} = \begin{pmatrix} 0 \\ \theta_1 \\ \theta_2 \\ \theta_1 + \theta_2 + \theta_{12} \end{pmatrix} = A^T \begin{pmatrix} \theta_1 \\ \theta_2 \\ \theta_{12} \end{pmatrix},$$

where $A^T = [[0,0,0], [1,0,0], [0,1,0], [1,1,1]]$.

We say X follows an Ising mixture model if its vector of cell probabilities is expressed as a finite mixture of Ising models, i.e.

$$(P(\boldsymbol{X} = \boldsymbol{i}) : \boldsymbol{i} \in I)^T = \boldsymbol{p}_{\text{mix}}(\boldsymbol{w}, \boldsymbol{\Theta}) := (p_{\text{mix},\boldsymbol{i}}(\boldsymbol{w}, \boldsymbol{\Theta}), \boldsymbol{i} \in I)^T := \sum_{k=1}^{K} w^{(k)} \boldsymbol{p}(\boldsymbol{\theta}^{(k)}), \tag{4.2}$$

with

$$\boldsymbol{p}(\boldsymbol{\theta}^{(k)}) = \exp(A^T \boldsymbol{\theta}^{(k)}) / [\mathbf{1}_{|I|}^T \exp(A^T \boldsymbol{\theta}^{(k)})].$$

Here $K \in \mathbb{N}^+$ is the number of components, $\boldsymbol{w} = (w_k : k \in [K])^T \in (0,1)^K$ represents the weights of the K

components with $\sum_{k\in[K]} w_k = 1$, $\boldsymbol{\theta}^{(k)} := (\theta_1^{(k)}, \ldots, \theta_d^{(k)}, \theta_{12}^{(k)}, \ldots, \theta_{(d-1)d}^{(k)}) \in \mathbb{R}^{d(d+1)/2}$ is the vector of main effects and interaction effects for component k. We define the vector of parameters of the Ising mixture model as $\boldsymbol{\Theta} := (\boldsymbol{\theta}^{(k)}, k \in [K]) \in \mathbb{R}^{Kd(d+1)/2}$.

The Ising mixture model says that the binary random vector $\boldsymbol{X}$ is drawn from K subpopulations with probabilities $\boldsymbol{w}$. Given the k-th subpopulation, $k \in [K]$, $\boldsymbol{X}$ follows an Ising model with parameters $\boldsymbol{\theta}^{(k)}$.

We assume that the observed data consist of N i.i.d. observations of $\boldsymbol{X}$ under the simple multinomial sampling theme (Cochran, 1952). It then follows that the resulting cell counts $\boldsymbol{n} := (n_{\boldsymbol{i}} : \boldsymbol{i} \in I)^T$ follow a Multinomial$(N, \boldsymbol{p})$ distribution, where $N = \sum_{\boldsymbol{i}\in I} n_{\boldsymbol{i}}$. Let $\|\cdot\|_2$ be the Euclidean norm of a vector.

4.2.2 Prior specification

In our proposed Bayesian framework, we assume that the mixture weights $\boldsymbol{w}$ follow a Dirichlet distribution $\text{Dirichlet}(\boldsymbol{\alpha})$ with parameters $\boldsymbol{\alpha} := (\alpha^{(k)}, k \in [K])$ with $\alpha^{(k)} > 0$. This is a common choice for nonnegative parameters that sum to 1 (Olkin and Rubin, 1964). For each component k, we assume that the main effects $(\theta_v^{(k)}, v \in [d])^T$ independently and identically follow a normal distribution with mean 0 and variance $\sigma_1^2 > 0$, denoted by $N(0, \sigma_1^2)$. We also assume that the interaction effects $(\theta_{v'v}^{(k)}, v' < v)^T$ independently and identically follow a continuous spike-and-slab prior with spike variance σ_0^2 and slab variance σ_1^2 where $0 < \sigma_0 < \sigma_1$. Specifically,

$$\theta_{v'v}^{(k)} | \gamma_{v'v}^{(k)} \sim (1 - \gamma_{v'v}^{(k)}) N(0, \sigma_0^2) + \gamma_{v'v}^{(k)} N(0, \sigma_1^2), \tag{4.3}$$

where $\gamma_{v'v}^{(k)}$ is the indicator of the association between variables $X_{v'}$ and X_v in the k-th component, and it is assumed to follow a Bernoulli distribution with known parameter $\beta \in (0, 1)$. More precisely, $\gamma_{v'v}^{(k)} = 0$ indicates that the interaction effect between variables v' and v in the kth component is more likely to be close to 0, while a value of $\gamma_{v'v}^{(k)} = 1$ implies that this interaction effect is more likely to be non-zero. This is particularly clear when the spike variance, σ_0^2, is set to zero, resulting in a point-mass mixture of a point mass at 0 and a normal distribution for the interaction effects. Denote $\boldsymbol{\gamma}^{(k)} := (\gamma_{v'v}^{(k)}, v' < v)^T$ and let $\boldsymbol{\Gamma} := (\boldsymbol{\gamma}^{(k)}, k \in [K]) \in \{0, 1\}^{Kd(d-1)/2}$. The collection of binary random variables $\boldsymbol{\Gamma}$ is of key interest since it represents the presence of non-zero interaction effects between variables in each component.

The Normal distributions in the spike-and-slab prior can be changed to other distributions such as the Laplace distribution. The continuous spike-and-slab prior, which serves as the predecessor to the point-mass mixture, has been gaining renewed attention in recent years (Ročková and George, 2014, 2018). The continuity of this prior allows for a more fluid exploration of posteriors using both MCMC and optimization techniques due to its ability to decrease the spike variance to zero and explore the entire path of posteriors as it approaches the point mass mixture. In addition to computational benefits, continuous mixture priors

can also exhibit optimal posterior behavior, such as the oracle property of the posterior mean (Ishwaran and Rao, 2005; Ročková, 2018).

4.2.3 Posterior distribution

To obtain the joint posterior distribution of $\boldsymbol{w}$, $\boldsymbol{\Theta}$ and $\boldsymbol{\Gamma}$, we begin by discussing the Ising model (4.1), and then consider the Ising mixture model (4.2) with $K \geq 2$ components.

The Ising model

Under Multinomial sampling, we have $\boldsymbol{n} \mid \boldsymbol{\theta} \sim \text{Multinomial}(N, \boldsymbol{p}(\boldsymbol{\theta}))$. It then follows that the probability mass function of $\boldsymbol{n}$ given $\boldsymbol{\theta}$ is

$$\pi(\boldsymbol{n} \mid \boldsymbol{\theta}) = \frac{N!}{\prod_{\boldsymbol{i} \in I} n_{\boldsymbol{i}}!} [\mathbf{1}_{|I|} \cdot \exp(A^T \boldsymbol{\theta})]^{-N} \exp(\boldsymbol{n}^T A^T \boldsymbol{\theta}) = \frac{N!}{\prod_{\boldsymbol{i} \in I} n_{\boldsymbol{i}}!} \exp[N \ell(\boldsymbol{\theta} | \boldsymbol{n})],$$

where $\ell(\boldsymbol{\theta} \mid \boldsymbol{n}) := -\log[\mathbf{1}_{|I|} \cdot \exp(A^T \boldsymbol{\theta})] + \boldsymbol{n}^T A^T \boldsymbol{\theta} / N$ is the log-likelihood function of the Ising model. It further follows from $\pi(\boldsymbol{n}, \boldsymbol{\theta}, \boldsymbol{\gamma}) = \pi(\boldsymbol{n}|\boldsymbol{\theta}) \pi(\boldsymbol{\theta}|\boldsymbol{\gamma}) \pi(\boldsymbol{\gamma})$ that

$$\begin{aligned}
\pi(\boldsymbol{n}, \boldsymbol{\theta}, \boldsymbol{\gamma}) = \frac{N!}{\prod_{\boldsymbol{i} \in I} n_{\boldsymbol{i}}!} \exp[N \ell(\boldsymbol{\theta})] \cdot \prod_{v \in [d]} \frac{1}{\sigma_1 \sqrt{2\pi}} \exp\Big(- \frac{\theta_v^2}{2\sigma_1^2} \Big) \\
\cdot \prod_{v' < v} \Big(\frac{1}{\sigma_0 \sqrt{2\pi}} \exp\Big(- \frac{\theta_{v'v}^2}{2\sigma_0^2} \Big) \Big)^{1 - \gamma_{v'v}} \Big(\frac{1}{\sigma_1 \sqrt{2\pi}} \exp\Big(- \frac{\theta_{v'v}^2}{2\sigma_1^2} \Big) \Big)^{\gamma_{v'v}} \\
\cdot \prod_{v' < v} \beta^{\gamma_{v'v}} (1 - \beta)^{1 - \gamma_{v'v}}.
\end{aligned}$$

After some algebra we obtain

$$\pi(\boldsymbol{\gamma} \mid \boldsymbol{n}) \propto \int \exp\left[N \ell(\boldsymbol{\theta} \mid \boldsymbol{n}) - \frac{\sum_{v \in [d]} \theta_v^2}{2\sigma_1^2} \right] \cdot \prod_{v' < v} \left[\exp\left(\frac{-\theta_{v'v}^2}{2\sigma_0^2} \right) \right]^{1 - \gamma_{v'v}} \left[\frac{\beta \sigma_0}{(1 - \beta)\sigma_1} \exp\left(\frac{-\theta_{v'v}^2}{2\sigma_1^2} \right) \right]^{\gamma_{v'v}} \mathrm{d}\boldsymbol{\theta}. \tag{4.4}$$

We note that the posterior distribution of $\boldsymbol{\gamma}$ is a mixture of Bernoulli distributions. Specifically, given $\boldsymbol{\theta}$ the posterior distribution of $\boldsymbol{\gamma}$ is

$$\gamma_{v'v} | \boldsymbol{\theta} \sim \text{Bernoulli}\left(\frac{1}{1 + (1 - \beta)\sigma_1 / (\beta \sigma_0) \cdot e^{\theta_{v'v}^2 (1/\sigma_1^2 - 1/\sigma_0^2)/2}} \right), v' < v, \text{ independently}.$$

As a consequence, the posterior mean of $\boldsymbol{\gamma}$ can be written as $E_{\theta_{v'v} \sim \pi(\boldsymbol{\theta} | \boldsymbol{n})}[r(\theta_{v'v})], v' < v$, where

$$\theta \mapsto r(\theta) := \frac{1}{1 + (1 - \beta)\sigma_1 / (\beta \sigma_0) \cdot e^{\theta^2 (1/\sigma_1^2 - 1/\sigma_0^2)/2}}, \tag{4.5}$$

For any fixed $\sigma_1 > 0$, we have $r(\theta) \to I(\theta \neq 0)$ as $\sigma_0 \to 0$ for any $\theta \in \mathbb{R}$. This means that $r(\theta)$ can be interpreted as a smoothed measurement of whether the interaction effect θ is zero. For a small σ_0, $r(\theta)$ will be still close to 1 even if $|\theta|$ is small. This is interpreted that the sensitivity of measuring $I(\theta \neq 0)$

increases as σ_0 decreases. It is worth noting that the function $r(\theta)$ has a lower bound that is always positive due to the continuous spike-and-slab prior. Specifically, the lower bound is given by $r(\theta) \geq r(0) = 1/[1+(1-\beta)\sigma_1/(\beta\sigma_0)]$, where $r(0)$ is a function that monotonically increases as σ_0 increases. The posterior density $\pi(\boldsymbol{\theta} \mid \boldsymbol{n})$ is proportional to

$$h_1(\boldsymbol{\theta}) := \exp\left(-N\log[\mathbf{1}_{|I|}^T \exp(A^T\boldsymbol{\theta})] + \boldsymbol{n}^T\boldsymbol{A}^T\boldsymbol{\theta} - \frac{\boldsymbol{\theta}^T\boldsymbol{\theta}}{2\sigma_1^2}\right) \cdot \prod_{v'<v}\left[\frac{(1-\beta)\sigma_1}{\beta\sigma_0}\exp\left(\frac{\theta_{v'v}^2}{2\sigma_1^2} - \frac{\theta_{v'v}^2}{2\sigma_0^2}\right) + 1\right].$$

In other words, $E(\boldsymbol{\gamma} \mid \boldsymbol{n}) = E_{\boldsymbol{\theta}\sim\pi(\boldsymbol{\theta}|\boldsymbol{n})}[r(\boldsymbol{\theta})]$, where applying $r(\cdot)$ to a vector means applying it element-wisely, i.e., $r(\boldsymbol{\theta}) = (r(\theta_{v'v}) : v' < v)^T$.

The Ising mixture model

The posterior density of $\boldsymbol{w}$, $\boldsymbol{\Theta}$ and $\boldsymbol{\Gamma}$ can be written as

$$\pi(\boldsymbol{w}, \boldsymbol{\Theta}, \boldsymbol{\Gamma} \mid \boldsymbol{n}) = \prod_{k, v'<v} \pi(\gamma_{v'v}^{(k)} \mid \theta_{v'v}^{(k)}) \cdot \pi(\boldsymbol{w}, \boldsymbol{\Theta} \mid \boldsymbol{n}),$$

where the posterior distribution of $\gamma_{v'v}^{(k)}$ given $\theta_{v'v}^{(k)}$ follows a $\mathrm{Bernoulli}\left(r(\theta_{v'v}^{(k)})\right)$ distribution and $\pi(\boldsymbol{w}, \boldsymbol{\Theta} \mid \boldsymbol{n})$ is the posterior density of $\boldsymbol{w}, \boldsymbol{\Theta}$ which is proportional to

$$h_4(\boldsymbol{w}, \boldsymbol{\Theta} \mid \boldsymbol{n}) := \exp[N\tilde{\ell}(\boldsymbol{w}, \boldsymbol{\Theta} \mid \boldsymbol{n})] \cdot \prod_{v'<v,k}\Big[\frac{(1-\beta)\sigma_1}{\beta\sigma_0}\exp\Big(\frac{[\theta_{v'v}^{(k)}]^2}{2\sigma_1^2} - \frac{[\theta_{v'v}^{(k)}]^2}{2\sigma_0^2}\Big) + 1\Big].$$

Here

$$\tilde{\ell}(\boldsymbol{w}, \boldsymbol{\Theta} \mid \boldsymbol{n}) = \ell(\boldsymbol{w}, \boldsymbol{\Theta} \mid \boldsymbol{n}) + \sum_{k\in[K]}(\alpha_k - 1)\log(w_k)/N - \|\boldsymbol{\Theta}\|_2^2/(2N\sigma_1^2),$$

where $\ell(\boldsymbol{w}, \boldsymbol{\Theta} \mid \boldsymbol{n})$ is the log-likelihood function of the Ising mixture model

$$\ell(\boldsymbol{w}, \boldsymbol{\Theta} \mid \boldsymbol{n}) = \frac{\boldsymbol{n}^T}{N}\log\Big(\sum_{k\in[K]} w^{(k)}\exp[A^T\boldsymbol{\theta}^{(k)} - \log(\mathbf{1}_{|I|}^T\exp(A^T\boldsymbol{\theta}^{(k)}))]\Big).$$

Given the posterior distribution of $\boldsymbol{\Theta}$, the posterior distribution of $\boldsymbol{\gamma}$ does not depend on the data $\boldsymbol{n}$. The posterior distribution of $\gamma_{v'v}^{(k)}$ is a mixture of Bernoulli distributions with $E(\gamma_{v'v}^{(k)} \mid \boldsymbol{n}) = E_{\theta_{v'v}^{(k)}\sim\pi(\boldsymbol{w},\boldsymbol{\Theta}|\boldsymbol{n})}[r(\theta_{v'v}^{(k)})]$. The posterior mean of $\boldsymbol{w}$ is $E(\boldsymbol{w} \mid \boldsymbol{n}) = E_{\boldsymbol{w}\sim\pi(\boldsymbol{w},\boldsymbol{\Theta}|\boldsymbol{n})}\boldsymbol{w}$. We are specifically interested in the posterior mean of $\gamma_{v'v}^{(k)}$ because its value between 0 and 1 reflects the magnitude of $|\theta_{v'v}^{(k)}|$ through the function $|\theta| \mapsto r(|\theta|)$. As the true value of $|\theta|$ increases, the posterior mean of $\boldsymbol{\Gamma}$ also approaches 1, allowing us to identify the most significant non-zero interaction effects.

4.2.4 Computing posterior means

We present importance sampling algorithms for computing the posterior means of $\mathbf{\Gamma}$ in the Ising mixture model. In the case of a single component $K = 1$, we use the normal approximation as the sampling distribution. In the case of multiple components $K \geq 2$, we use the normal mixture approximation instead.

The Ising model

Recall that $E(\boldsymbol{\gamma} \mid \boldsymbol{n}) = E_{\boldsymbol{\theta} \sim \pi(\boldsymbol{\theta}|\boldsymbol{n})}[r(\boldsymbol{\theta})]$, where $r(\cdot)$ is defined in (4.5) and

$$\pi(\boldsymbol{\theta} \mid \boldsymbol{n}) \propto h_1(\boldsymbol{\theta}) = \exp(N\tilde{\ell}(\boldsymbol{\theta})) \cdot \prod_{v'<v} \left[\frac{(1-\beta)\sigma_1}{\beta\sigma_0} \exp\left(\frac{\theta_{v'v}^2}{2\sigma_1^2} - \frac{\theta_{v'v}^2}{2\sigma_0^2} \right) + 1 \right]$$

with $\tilde{\ell}(\boldsymbol{\theta}) := \ell(\boldsymbol{\theta}) - \boldsymbol{\theta}^T\boldsymbol{\theta}/[2N\sigma_1^2]$ and $\ell(\boldsymbol{\theta}) := -\log[\mathbf{1}_{|I|}^T \exp(A^T\boldsymbol{\theta})] + \boldsymbol{n}^T\boldsymbol{A}^T\boldsymbol{\theta}/N$. Note that $\ell(\boldsymbol{\theta})$ is the log-likelihood function of Ising models and $\tilde{\ell}(\boldsymbol{\theta})$ is its regularized version. Because the function

$$(y_1, \ldots, y_n) \mapsto \log[\exp(y_1) + \ldots + \exp(y_n)],$$

is convex, the function $\boldsymbol{\theta} \mapsto \tilde{\ell}(\boldsymbol{\theta})$ is strictly concave, thus it has a unique maximum at the point $\tilde{\boldsymbol{\theta}} := \operatorname{argmax}_{\boldsymbol{\theta}} \tilde{\ell}(\boldsymbol{\theta})$. It follows from the Taylor series of the regularized log-likelihood $\tilde{\ell}$ that

$$\tilde{\ell}(\boldsymbol{\theta}) \approx \tilde{\ell}(\tilde{\boldsymbol{\theta}}) - \frac{1}{2}(\boldsymbol{\theta} - \tilde{\boldsymbol{\theta}})^T \widetilde{\Sigma}^{-1} (\boldsymbol{\theta} - \tilde{\boldsymbol{\theta}}),$$

where $\widetilde{\Sigma}$ is the inverse of Hessian matrix of $\boldsymbol{\theta} \mapsto -\tilde{\ell}(\boldsymbol{\theta})$ at $\tilde{\boldsymbol{\theta}}$. Thus $\exp(N\tilde{\ell}(\boldsymbol{\theta}))$ can be approximated by a Normal density with mean $\tilde{\boldsymbol{\theta}}$ and covariance $\widetilde{\Sigma}/N$ (up to multipling a constant). After some algebra it follows that

$$\widetilde{\Sigma} = \left[A \left\{ \frac{\operatorname{diag}(\exp(A^T\boldsymbol{\theta}))}{\mathbf{1}_{|I|}^T \exp(A^T\boldsymbol{\theta})} - \frac{\exp(A^T\boldsymbol{\theta}) \exp(A^T\boldsymbol{\theta})^T}{[\mathbf{1}_{|I|}^T \exp(A^T\boldsymbol{\theta})]^2} \right\} A^T + \frac{\boldsymbol{I}_{d(d+1)/2}}{N\sigma_1^2} \right]^{-1},$$

where $\operatorname{diag}(\exp(\mathrm{A}^{\mathrm{T}}\boldsymbol{\theta}))$ represents a diagonal matrix with diagonal $\exp(A^T\boldsymbol{\theta})$. Let $h_2(\boldsymbol{\theta})$ be the density function of $N(\tilde{\boldsymbol{\theta}}, \widetilde{\Sigma}/N)$. We obtain that the ratio $h_1(\boldsymbol{\theta})/h_2(\boldsymbol{\theta})$ is proportional to

$$h_3(\boldsymbol{\theta}) := \exp\left(N\tilde{\ell}(\boldsymbol{\theta}) - N\tilde{\ell}(\tilde{\boldsymbol{\theta}}) + \frac{N}{2}(\boldsymbol{\theta} - \tilde{\boldsymbol{\theta}})^T \widetilde{\Sigma}^{-1} (\boldsymbol{\theta} - \tilde{\boldsymbol{\theta}}) \right) \cdot \prod_{v'<v} \left[\frac{(1-\beta)\sigma_1}{\beta\sigma_0} \exp\left(\frac{\theta_{v'v}^2}{2\sigma_1^2} - \frac{\theta_{v'v}^2}{2\sigma_0^2} \right) + 1 \right].$$

The importance sampling estimate of the posterior mean of $\boldsymbol{\gamma}$ is

$$E(\boldsymbol{\gamma} \mid \boldsymbol{n}) = \frac{E_{\boldsymbol{\theta} \sim h_2} r(\boldsymbol{\theta}) h_3(\boldsymbol{\theta})}{E_{\boldsymbol{\theta} \sim h_2} h_3(\boldsymbol{\theta})} \approx \frac{\frac{1}{M}\sum_{m \in [M]} r(\boldsymbol{\theta}_m) h_3(\boldsymbol{\theta}_m)}{\frac{1}{M}\sum_{m \in [M]} h_3(\boldsymbol{\theta}_m)} = \sum_{m \in [M]} r(\boldsymbol{\theta}_m) \frac{h_3(\boldsymbol{\theta}_m)}{\sum_{m \in [M]} h_3(\boldsymbol{\theta}_m)},$$

where $(\boldsymbol{\theta}_m, m \in [M])^T$ are i.i.d. sampled from $N(\tilde{\boldsymbol{\theta}}, \widetilde{\Sigma}/N)$.

The Ising mixture model

In this case, we use the normal mixture approximation (Gamerman and Lopes, 2006, page 85-86)

instead of the normal approximation because the log-likelihood function $\boldsymbol{w}, \boldsymbol{\Theta} \mapsto \ell(\boldsymbol{w}, \boldsymbol{\Theta} \mid \boldsymbol{n})$ may have multiple modes. The normal mixture sampling distribution denoted by $h_5(\boldsymbol{w}, \boldsymbol{\Theta})$ can be constructed as follows.

1. Initialize a random start value of $\boldsymbol{w}, \boldsymbol{\Theta}$ and find a local optimal point $\tilde{\boldsymbol{w}}, \tilde{\boldsymbol{\Theta}}$ that minimizes $\boldsymbol{w}, \boldsymbol{\Theta} \mapsto \tilde{\ell}(\boldsymbol{w}, \boldsymbol{\Theta} \mid \boldsymbol{n})$ based on this start value.
2. Repeat the last step J times. These local optimal points are denoted by $\{\tilde{\boldsymbol{w}}_j, \tilde{\boldsymbol{\Theta}}_j\}_{j=1}^J$ and the corresponding optimal value is $\tilde{\ell}_j := \tilde{\ell}(\tilde{\boldsymbol{w}}_j, \tilde{\boldsymbol{\Theta}}_j \mid \boldsymbol{n})$. Let $\tilde{\Sigma}_j$ be the inverse of Hessian matrix of $\boldsymbol{\Theta} \mapsto -\tilde{\ell}(\tilde{\boldsymbol{w}}_j, \boldsymbol{\Theta})$ at $\tilde{\boldsymbol{\Theta}}_j$.
3. Let $f(\boldsymbol{\Theta} \mid \tilde{\boldsymbol{\Theta}}_j, \tilde{\Sigma}_j/N)$ be the density function of $N(\tilde{\boldsymbol{\Theta}}_j, \tilde{\Sigma}_j/N)$ and let $f(\boldsymbol{w} \mid N\tilde{\boldsymbol{w}}_j + \mathbf{1}_K)$ be the density function of Dirichlet distribution with parameters $N\tilde{\boldsymbol{w}}_j + \mathbf{1}_K$.
4. Then let $h_5(\boldsymbol{w}, \boldsymbol{\Theta})$ be $\sum_{j \in [J]} \frac{\exp(\tilde{\ell}_j)}{\sum_{j \in [J]} \exp(\tilde{\ell}_j)} f(\boldsymbol{w} \mid N\tilde{\boldsymbol{w}}_j + \mathbf{1}_K) f(\boldsymbol{\Theta} \mid \tilde{\boldsymbol{\Theta}}_j, \tilde{\Sigma}_j/N)$.

The number of components, J, in the normal mixture sampling distribution can be chosen arbitrarily, but using $J = 5$ or 10 is typically sufficient to capture the majority of important modes.

The first two steps identify main local maximum points and preparing for normal approximations for each of them. The sampling distributions are constructed in the third step. We choose to use the normal approximation for the main effects and interaction effects $\boldsymbol{\Theta}$. We note that the mode of the Dirichlet sampling distribution of the weights $\boldsymbol{w} \in (0, 1)^K$ corresponds to the local maximum points. Its parameters are then scaled by N to account for the sample size, which can be justified by equating the second derivative of the objective function with that of the sampling density function in the specific scenario where the number of components K is 2. These sampling distributions are combined into the full sampling distribution in which their weights are given by the corresponding values of the likelihood. The components with higher likelihood receive higher weights in the full sampling distribution.

The posterior mean of $\boldsymbol{\Theta}$ and $\boldsymbol{w}$ is given by

$$
\begin{aligned}
E(\boldsymbol{\Gamma} \mid \boldsymbol{n}) &= E_{\boldsymbol{\Theta} \sim \pi(\boldsymbol{w}, \boldsymbol{\Theta} \mid \boldsymbol{n})} r(\boldsymbol{\Theta}) = \int \pi(\boldsymbol{w}, \boldsymbol{\Theta} \mid \boldsymbol{n}) r(\boldsymbol{\Theta}) \mathrm{d}\boldsymbol{\Theta} \mathrm{d}\boldsymbol{w} = \int \frac{\pi(\boldsymbol{w}, \boldsymbol{\Theta} \mid \boldsymbol{n})}{h_5(\boldsymbol{w}, \boldsymbol{\Theta})} h_5(\boldsymbol{w}, \boldsymbol{\Theta}) r(\boldsymbol{\Theta}) \mathrm{d}\boldsymbol{\Theta} \mathrm{d}\boldsymbol{w} \\
&= E_{\boldsymbol{w}, \boldsymbol{\Theta} \sim h_5(\boldsymbol{w}, \boldsymbol{\Theta})} \frac{\pi(\boldsymbol{w}, \boldsymbol{\Theta} \mid \boldsymbol{n})}{h_5(\boldsymbol{w}, \boldsymbol{\Theta})} r(\boldsymbol{\Theta}) = \frac{E_{\boldsymbol{w}, \boldsymbol{\Theta} \sim h_5(\boldsymbol{w}, \boldsymbol{\Theta})} \frac{h_4(\boldsymbol{w}, \boldsymbol{\Theta} \mid \boldsymbol{n})}{h_5(\boldsymbol{w}, \boldsymbol{\Theta})} r(\boldsymbol{\Theta})}{E_{\boldsymbol{\Theta}, \boldsymbol{w} \sim h_5(\boldsymbol{w}, \boldsymbol{\Theta})} \frac{h_4(\boldsymbol{w}, \boldsymbol{\Theta} \mid \boldsymbol{n})}{h_5(\boldsymbol{w}, \boldsymbol{\Theta})}},
\end{aligned}
$$

where $\pi(\boldsymbol{\Theta}, \boldsymbol{w} \mid \boldsymbol{n}) \propto h_4(\boldsymbol{\Theta}, \boldsymbol{w} \mid \boldsymbol{n})$.

As the sample size N increases, the regularized log-likelihood $\tilde{\ell}$ gets closer to the log-likelihood func-

tion ℓ. Its maximum optimal point $\tilde{\boldsymbol{\theta}} = \operatorname{argmax}_{\boldsymbol{\theta}} \tilde{\ell}(\boldsymbol{\theta})$ gets closer to that of the log-likelihood function, $\operatorname{argmax}_{\boldsymbol{\theta}} \ell(\boldsymbol{\theta})$. If the Ising mixture model is identifiable, the mean of this sampling distribution converges to the true values of $\boldsymbol{\Theta}$ as N goes to infinity. Additionally, as N increases, the covariance matrix of the sampling distribution, $\tilde{\Sigma}/N$, converges to the zero matrix, indicating that the sampling distribution is becoming more concentrated at the true values of interaction effects. It is worth noting that the classical MLE usually plays an important role in the sampling algorithm in Bayesian analysis, as demonstrated in various studies (Dobra and Massam, 2010; Fienberg and Rinaldo, 2012). For Ising mixture models the regularized log-likelihood function $\tilde{\ell}$ can also be replaced by the log-likelihood function ℓ in the sampling algorithm.

Since the weights $\boldsymbol{w}$ are between 0 and 1, it is not appropriate to sample them from a Normal distribution. Instead, we sample them from a Dirichlet distribution with mode $\hat{\boldsymbol{w}}$. The parameter of this Dirichlet distribution is set to $N\hat{\boldsymbol{w}} + 1$ in order to reflect the increased concentration around the mode as the sample size N increases.

The posterior mean of $\boldsymbol{\Gamma}$ remains a meaningful method for inferring associations between variables, even if the density function $\pi(\boldsymbol{n} \mid \boldsymbol{\Gamma})$ is non-identifiable. This is due to the fact that the posterior distribution of $\boldsymbol{\Gamma}$ is proportional to the product of the likelihood function $\pi(\boldsymbol{\Gamma} \mid \boldsymbol{n})$ and the prior distribution of $\boldsymbol{\Gamma}$, $\pi(\boldsymbol{\Gamma})$. If $\pi(\boldsymbol{n} \mid \boldsymbol{\Gamma})$ is non-identifiable, meaning that multiple values of $\boldsymbol{\Gamma}$ produce the same likelihood, the posterior mean of $\boldsymbol{\Gamma}$ is the weighted average of all such values as the sample size $N \to \infty$. In particular, if the prior parameter for each element $\gamma_{v'v}^{(k)}$ is set to $\beta = 0.5$, the posterior mean can be interpreted as a majority vote, with a value greater than 0.5 indicating that the majority of values of $\gamma_{v'v}^{(k)}$ that produce the same likelihood are 1. Additionally, decreasing the prior parameter, e.g., $\beta < 0.5$, can favor sparser association structures.

The identifiability of the density function $\pi(\boldsymbol{n} \mid \boldsymbol{\Gamma})$ is implied by the identifiability of $\pi(\boldsymbol{n} \mid \boldsymbol{\Theta}, \boldsymbol{w})$. We discuss necessary and sufficient identifiability conditions for $\pi(\boldsymbol{n} \mid \boldsymbol{\Theta}, \boldsymbol{w})$ in Chapter 4.5. On the other hand, the identifiability of $\pi(\boldsymbol{n} \mid \boldsymbol{\Gamma})$ can be partially solved by considering the rank of the observed information matrix evaluated at the MLEs for the mixture parameters, as discussed in Frühwirth-Schnatter (2006, Chapter 9.5.2). This is a consequence of the equivalence between local identifiability and the rank of the information matrix, as established in (Rothenberg, 1971; Catchpole and Morgan, 1997).

We determine the Fisher information matrix of an Ising mixture model with parameters $\boldsymbol{w}$, $\boldsymbol{\Theta}$ from the log-likelihood function $\ell(\boldsymbol{w}, \boldsymbol{\Theta} \mid \boldsymbol{X}) = \log p_{\text{mix},\boldsymbol{X}}(\boldsymbol{w}, \boldsymbol{\Theta})$:

$$\mathcal{I}(\boldsymbol{w}, \boldsymbol{\Theta}) := -E\left[\frac{\partial^2 \log p_{\text{mix},\boldsymbol{X}}(\boldsymbol{w}, \boldsymbol{\Theta})}{\partial(\boldsymbol{w}, \boldsymbol{\Theta})^2}\right] = -\sum_{\boldsymbol{i} \in I} p_{\text{mix},\boldsymbol{i}}(\boldsymbol{w}, \boldsymbol{\Theta}) \frac{\partial^2 \log p_{\text{mix},\boldsymbol{i}}(\boldsymbol{w}, \boldsymbol{\Theta})}{\partial(\boldsymbol{w}, \boldsymbol{\Theta})^2}.$$

The Fisher information matrix provides a justification for the local identifiability of an Ising mixture model– see Chapter 4.5.

4.3 Simulation experiments

We evaluate the empirical performance of our proposed Bayesian framework for assessing the strength of association in Ising mixture models. The number of binary variables is fixed at $d = 6$.

4.3.1 The Ising model

We set the number of variables to $d = 6$, the sample size to $N = 10000$, and the main effects to

$$(\theta_1, \theta_2, \theta_3, \theta_4, \theta_5, \theta_6) = (1, -1, 1, -1, 1, -1).$$

For the interaction effects $(\theta_{v'v} : v' < v)$ we used two designs. In design A, $(\theta_{12}, \theta_{13}, \theta_{14}, \theta_{23}) = (1, -1, 1, -1)$ and others are 0. In design B, $(\theta_{12}, \theta_{13}, \theta_{14}, \theta_{23}) = (1, -0.5, 0.2, -.1)$ and others are 0.

We chose the following combinations for the hyperparameters of the prior distributions. In Setting 1, $\sigma_0 = 0.1$, $\sigma_1 = 1$, $\beta = 0.5$. In Setting 2: $\sigma_0 = 0.01$, $\sigma_1 = 1$, $\beta = 0.5$. The two settings illustrate the sensitivity of the results with respect to the ratio of the two variances in the spike-and-slab prior (4.3). The sampling size M in the importance sampling algorithm is 10^5.

The data consist of the six-way contingency table with counts given by $N \cdot \boldsymbol{p}(\boldsymbol{\theta})$, where $p(\boldsymbol{\theta})$ is determined as in Equation (4.1). Keeping the data fixed as opposed to sampling it from $\text{Multinomial}(N, \boldsymbol{p}(\boldsymbol{\theta}))$ allows us to evaluate he sampling error caused by the importance sampling procedure and the performance of the proposed Bayesian method as the sample size N approaches infinity. The posterior mean of the association indicators, $\boldsymbol{\gamma}$, is reported for each combination of designs and settings in Table 8. These results are an average of 100 independent replicates of the importance sampling algorithm with $M = 10^5$.

The results in Table 8 provide evidence for the effectiveness of the proposed Bayesian framework in inferring associations between variables. Under Design A, all four non-zero interaction terms have an absolute value of 1 which makes them clearly distinguishable from the other interaction terms that are set to zero. In both prior settings, the estimated posterior means of the indicators corresponding to non-zero interaction effects is 1, while the estimated posterior means of the zero interaction effects is 0.1 or less.

Under Design B, three of the four non-zero interaction effects have an absolute value of 0.5 or less. Their smaller size makes them less distinguishable from the remaining interaction terms that are set to zero. In Setting 1 ($\sigma_0 = .1$), the estimated posterior means of the association indicators for larger interaction effects is close to 1. The estimated posterior means for smaller interaction effects is less than 0.5, suggesting that these associations may be harder to identify. In Setting 2 ($\sigma_0 = .01$), the estimated posterior means of the association indicators for all four non-zero interaction effects is greater than 0.5, demonstrating the increased effectiveness of a smaller value of σ_0 in detecting small interaction effects.

The estimated posterior mean of γ is smaller in Setting 1 than in Setting 2 for the interaction effects that are set to zero. This should not be surprising, since the expectation $E(\gamma \mid \boldsymbol{n}) = E_{\theta\sim\pi(\boldsymbol{\theta}|\boldsymbol{n})}[r(\theta)]$, where $r(\theta)$ is defined in (4.5) is monotonically increasing with respect to σ_0 for small values of $|\theta|$. As a result, when σ_0 approaches 0, the lower bound of the posterior mean of γ, which is $r(0) = 1/(1 + (1 - \beta)\sigma_1/\beta\sigma_0)$, becomes smaller. On the other hand, the sampling error is larger in Setting 2 ($\sigma_0 = 0.01$) compared to Setting 1 ($\sigma_0 = 0.1$). Under both designs the importance sampling standard errors for the posterior mean estimates of association indicators are approximately 0.0001 in Setting 1 and 0.1 in Setting 2. The reason relates to the sampling density function $h_2(\boldsymbol{\theta})$ being closer to the objective density function $\pi(\boldsymbol{\theta} \mid \boldsymbol{n})$ in Setting 1, resulting in a lower variance of the importance sampling method. As such, selecting the value of σ_0 involves balancing the stability of the importance sampling algorithm with the ability to detect distinguish non-zero interaction effects.

γ_{12}	γ_{13}	γ_{14}	γ_{15}	γ_{16}	γ_{23}	γ_{24}	γ_{25}	γ_{26}	γ_{34}	γ_{35}	γ_{36}	γ_{45}	γ_{46}	γ_{56}
Posterior mean under Design A and Setting 1														
1.0	**1.0**	**1.0**	.10	.10	**1.0**	.10	.10	.10	.10	.10	.10	.10	.10	.10
Posterior mean under Design B and Setting 1														
1.0	**1.0**	.34	.10	.10	.14	.10	.10	.10	.10	.10	.10	.10	.10	.10
Posterior mean under Design A and Setting 2														
1.0	**1.0**	**1.0**	.08	.10	**1.0**	.08	.07	.08	.06	.09	.07	.07	.06	.09
Posterior mean under Design B and Setting 2														
1.0	**1.0**	**1.0**	.10	.10	**.68**	.07	.08	.07	.08	.08	.09	.08	.10	.08

Table 8: Estimated posterior mean of γ in Ising models under Design A, B, and Setting 1 and 2. Under both designs, the importance sampling standard errors are approximately 0.0001 for Setting 1 and 0.1 for Setting 2.

4.3.2 The Ising mixture model with two components

We set the number of variables to $d = 6$, the sample size $N = 10000$, the weight for the first component to $w^{(1)} = 0.4$, and the main effects to $\boldsymbol{\theta}^{(1)} = \boldsymbol{\theta}^{(2)} = (1, -1, 1, -1, 1, -1)$. The spike-and-slab prior parameters are set to $\sigma_0 = 0.1$, $\sigma_1 = 1$, $\beta = .5$. This is Setting 1 in the previous simulation experiment. The sampling size M in the importance sampling algorithm is 10^5. The number of components J in the normal mixture sampling distribution for Bayesian Ising mixture models is 5.

The data consist of the six-way contingency table with counts given by $\boldsymbol{n} = N \cdot \boldsymbol{p}_{\text{mix}}(\boldsymbol{w}, \boldsymbol{\Theta})$ where $p_{\text{mix}}(\boldsymbol{w}, \boldsymbol{\Theta})$ is defined in Equation (4.2). We used two designs for the interaction effects. In design C, $(\theta^{(1)}_{12}, \theta^{(1)}_{13}, \theta^{(2)}_{46}, \theta^{(2)}_{56}) = (1, -1, 1, -1)$ and others are 0. In design D, $(\theta^{(1)}_{12}, \theta^{(1)}_{13}, \theta^{(1)}_{23}, \theta^{(2)}_{14}, \theta^{(2)}_{15}) = (1, -1, 1, 1, -1)$ and others are 0. Under both designs, we fit an Ising model as well as an Ising mixture model with two components. Table 9 presents the estimated posterior means of the association indicators for both models.

These results are an average of 100 independent replicates of the importance sampling algorithm with $M = 10^5$. The importance sampling standard error is about 0.05 for both designs which is an indication of the stability of the importance sampling algorithm. As an illustration of computation time on average, the Bayesian Ising model required only 0.37 seconds, while the Bayesian two-component Ising mixture model took 5.14 minutes to complete the importance sampling algorithm under Design D and Setting 1. Both experiments were conducted on a laptop with a 1.8 GHz Intel Core i5 processor and 8 GB of memory.

Under both designs, the Ising model identifies the non-zero interaction effects from both components of the mixture. However, under Design D, it incorrectly identifies two additional interaction effects that are actually zero in both mixture components. On the other hand, the Ising mixture model with two components identifies all non-zero and all zero interaction effects in both components based on a cutoff of 0.5 under both designs. The estimated posterior mean of the weight $w^{(1)}$ is $.40$ under Design C and 0.41 under Design D, both estimates very close to the true value of 0.4.

This simulation setting shows that, if an Ising mixture model with one component is fit when the data corresponds with a mixture model with two components, the inferred association structure might include pairwise effects that do not exist. This result is in a sense not surprising given that the first component has non-zero interaction effects between variables 1 and 2, variables 1 3, and variables 2 and 3, while the second component has non-zero interaction effects between variables 1 4, and variables 1 and 5. Due to this configuration, the Ising model might show non-zero interaction effects between variables 2 and 4, and variables 2 and 5. Nevertheless, the estimated posterior mean of the association indicators for these variables is lower compared to the truly non-zero associations.

Model	γ_{12}	γ_{13}	γ_{14}	γ_{15}	γ_{16}	γ_{23}	γ_{24}	γ_{25}	γ_{26}	γ_{34}	γ_{35}	γ_{36}	γ_{45}	γ_{46}	γ_{56}
Posterior mean under Design C															
Ising model	**.99**	**.98**	.10	.10	.11	.30	.11	.11	.14	.11	.11	.14	.11	**1.0**	**1.0**
Ising mixture, Component 1	**1.0**	**1.0**	.17	.16	.20	.19	.16	.15	.17	.18	.17	.18	.16	.19	.19
Ising mixture, Component 2	.17	.14	.14	.13	.13	.16	.17	.16	.15	.14	.13	.13	.15	**1.0**	**1.0**
Posterior mean under Design D															
Ising model	**.98**	**.94**	**1.0**	**1.0**	.10	**.88**	**.69**	**.67**	.10	.13	.12	.13	.22	.22	.22
Ising mixture, Component 1	**.99**	**.99**	.19	.17	.14	**.99**	.17	.15	.15	.17	.16	.14	.17	.18	.15
Ising mixture, Component 2	.19	.14	**.99**	**.99**	.12	.17	.15	.17	.15	.12	.12	.12	.12	.12	.12

Table 9: Estimated posterior mean of γ in the Ising model and of $\gamma^{(1)}, \gamma^{(2)}$ in the Ising mixture model with two components under Design C and Design D.

4.4 Real data applications

We examine the fit of the Ising model and of the Ising mixture model with two components for two eight-way binary contingency tables. The first example focuses on the Rochdale data – a dataset that has been analyzed numerous times in the existent literature. The pairwise interactions of the Rochdale data are considered to be well understood. The second dataset comes from a larger dataset, and has not been previously analyzed in this form. We chose it because of its much larger sample size leads to significantly larger counts in some of the cells compared to the largest counts in the Rochdale data. The presence of the larger counts will test the ability of the Ising mixture models to adequately capture the imbalance between the magnitude of the largest and the smallest counts in sparse contingency tables.

4.4.1 The Rochdale data

The Rochdale data (Whittaker, 1990) was collected to determine the relationships among factors affecting women's economic activity. It includes eight binary variables: 1) wife's economic activity (no, yes), 2) age of wife > 38 (no, yes), 3) husband's employment status (no, yes), 4) presence of children ≤ 4 years old (no, yes), 5) wife's education level, high-school+ (no, yes), 6) husband's education level, high-school+ (no, yes), 7) Asian origin (no, yes), and 8) presence of other working household members (no, yes). With a sample size of 665, the resulting 2^8 contingency table, shown in Table 10, is sparse, with 165 cells having 0 counts, 217 cells having small positive counts less than 3, and several cells with counts larger than 30 or even 50.

5	0	2	1	5	1	0	0	4	1	0	0	6	0	2	0
8	0	11	0	13	0	1	0	3	0	1	0	26	0	1	0
5	0	2	0	0	0	0	0	0	0	0	0	0	0	1	0
4	0	8	2	6	0	1	0	1	0	1	0	0	0	1	0
17	10	1	1	16	7	0	0	0	2	0	0	10	6	0	0
1	0	2	0	0	0	0	0	1	0	0	0	0	0	0	0
4	7	3	1	1	1	2	0	1	0	0	0	1	0	0	0
0	0	3	0	0	0	0	0	0	0	0	0	0	0	0	0
18	3	2	0	23	4	0	0	22	2	0	0	57	3	0	0
5	1	0	0	11	0	1	0	11	0	0	0	29	2	1	1
3	0	0	0	4	0	0	0	1	0	0	0	0	0	0	0
1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0
41	25	0	1	37	26	0	0	15	10	0	0	43	22	0	0
0	0	0	0	2	0	0	0	0	0	0	0	3	0	0	0
2	4	0	0	2	1	0	0	0	1	0	0	2	1	0	0
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0

Table 10: Rochdale data from Whittaker (1990). The cells counts appear row by row in lexicographical order with the levels of variable 8 varying fastest and the levels of variable 1 varying slowest.

The estimated posterior means of the association indicators γ are presented in Table 12 based on an Ising model with spike-and-slab prior with $\sigma_0 = 0.1$, $\sigma_1 = 1$ and $\beta = 0.5$. In what follows we consider an

interaction effect between variables v' and v to be significant if $E(\gamma_{v'v} \mid \boldsymbol{n}) > 0.5$. In Figure 10 we compare the set of non-zero interaction effects we identified with those of Whittaker (1990). We find that all the 14 significant pairwise associations found by Whittaker (1990) are also found by our Ising model. However, the Ising model determined two additional interactions: one interaction between variables 5 (wife's education level) and 7 (Asian origin) with a posterior mean of the corresponding association indicator of 0.86, and the interaction between variables 2 (age of wife > 38) and 7 (Asian origin) with a posterior mean of 0.65. The posterior means of these two extra interactions are much smaller then the posterior means of the 14 interactions that were also determined by Whittaker (1990). These two extra associations seem to be reasonable, and can be attributed to the wave of Asian immigration, particularly of Asian women, in the last century (Kim, 1977; Piper and Roces, 2004).

We also applied our Bayesian Ising mixture model with two components to the Rochdale data – see Table 12 and Figure 11. We fitted the mixture model with invariant main effects across components (Assumption 4.1). The number of components in the normal mixture sampling distribution is $J = 5$. The estimated posterior mean of the weight of the first component is $E(w^{(1)} \mid \boldsymbol{n}) = 0.14$. The 16 significant associations found by the Bayesian Ising model are also idenfied in both components of the Ising mixture model. However, each of the two components involve additional significant associations. We note that the estimated posterior means of all the association indicators are 1. Goodness-of-fit tests for the maximum likelihood estimators show that both the Ising model (4.1) and the Ising mixture model (4.2) fit the data well with p-values of 0.42 and 1, respectively. The likelihood ratio test shows that the two-component Ising mixture model fits the data significantly better than the Ising model with a p-value < 0.001.

The left panel of Table 11 shows the cells containing the 10 largest observed counts in the Rochdale data together with their expected cell counts in the Ising model and the Ising mixture model. We see that both models are able to capture the largest counts reasonably well.

4.4.2 The NLTCS data

We analyze a dataset extracted from the National Long Term Care Survey (NLTCS) created by the Center of Demographic Studies at Duke University (Manton et al., 1993). It includes eight binary variables that measure functional disability in daily living activities: 1) eating, 2) getting around inside, 3) dressing, 4) cooking, 5) grocery shopping, 6) getting about outside, 7) traveling, and 8) managing money. Each measure classifies study participants as healthy or disabled. The data comprise observations of elderly individuals aged 65 and above, pooled across four survey waves from 1982, 1984, 1989, and 1994. With a sample size of 21574, the resulting 2^8 contingency table, Table 13 is sparse, with 17.1% cells having 0 counts, 46.9%

Cell	Observed	Ising models	Ising mixtures	Cell	Observed	Ising models	Ising mixtures
10001100	57	56.78	58.63	00000000	4419	4181.60	4320.54
11001100	43	44.61	42.89	00010000	2063	2087.60	2134.16
11000000	41	36.40	36.99	00110000	1189	1324.38	1175.31
11000100	37	38.81	37.65	11111111	1056	1035.05	1055.71
10011100	29	33.29	29.02	00111111	764	702.14	752.47
00011100	26	20.37	21.84	00110100	667	607.96	658.85
11000101	26	23.69	23.33	00110101	654	702.29	657.78
11000001	25	28.13	29.30	00110001	601	571.91	597.43
10000100	23	22.70	23.25	00111101	549	577.10	561.45
11001101	22	22.85	21.43	00010100	529	565.21	518.92

Table 11: Expected cell counts for the top 10 largest counts cells for the Rochdale data (left panel) and the NLTCS data (right panel). Cells are identified through their sequence of level indicators with 0 as no and 1 as yes.

cells having small counts no larger than 5, and largest 1.6% of the cell counts accounting for 40.5% of the observations.

We employ our Bayesian framework to fit an Ising model and an Ising mixture model with two components. The prior specification, assumptions related to the invariance of main effects and the number of components in the mixture sampling distributions were the same as the ones used for the Rochdale data. The pairwise associations that have an estimated posterior mean of their indicators above 0.5 are shown as graphs in the right panel of Figure 12. There are 21 significant interaction effects identified in the Ising model. The forward stepwise function from the R package gRim (Højsgaard et al., 2012) identifies 20 of these 21 pairwise interactions – see the left panel of Figure 12. The additional interaction identified by our Bayesian framework involves variables 1 and 4, and has the smallest estimated posterior mean of 0.85 among the 21 associations. In the Ising mixture model, there are 17 significant interaction effects in the first component and 20 significant interaction effects in the second component – see Figure 13. The estimated posterior mean of the weight of the first component is 0.4. The patterns of significant interaction effects in both components of the Ising mixture model are sparser than the pattern inferred in the Ising model. One example of a key difference between the inferred association patterns relates to variables 1 and 4. The estimated posterior mean of the association indicator is 0.85 in the Ising model, while it is less than 0.5 in the first component of the Ising mixture model and it is equal with 1 in the second component. As such, this pairwise association is a combination of a weaker effect in one component and a very strong effect in the second component. Similar patterns with varying strength between components involve variables 1 and 3 and variables 4 and 5.

Goodness-of-fit tests for the maximum likelihood estimators show that the Ising model (4.1) does not

Model	γ_{12}	γ_{13}	γ_{14}	γ_{15}	γ_{16}	γ_{17}	γ_{18}	γ_{23}	γ_{24}	γ_{25}	γ_{26}	γ_{27}	γ_{28}	γ_{34}
Ising model	.23	**1.0**	**1.0**	**.96**	.22	**1.0**	.21	.29	**1.0**	**1.0**	.18	**.65**	**1.0**	.25
Ising mixture, Component 1	**1.0**	**1.0**	**1.0**	**1.0**	**1.0**	**1.0**	**1.0**	**1.0**	**1.0**	**1.0**	**1.0**	**1.0**	**1.0**	**1.0**
Ising mixture, Component 2	.11	**1.0**	**1.0**	**1.0**	**.92**	**1.0**	**1.0**	**1.0**	**1.0**	**1.0**	.29	**1.0**	**1.0**	**1.0**

Model	γ_{35}	γ_{36}	γ_{37}	γ_{38}	γ_{45}	γ_{46}	γ_{47}	γ_{48}	γ_{56}	γ_{57}	γ_{58}	γ_{67}	γ_{68}	γ_{78}
Ising model	**1.0**	**.95**	**.98**	.28	.30	.46	**.99**	**.99**	**1.0**	**.86**	.37	**1.0**	.44	.37
Ising mixture, Component 1	**1.0**	**1.0**	**1.0**	**1.0**	**1.0**	**1.0**	**1.0**	**1.0**	**1.0**	**1.0**	**1.0**	**1.0**	**1.0**	**1.0**
Ising mixture, Component 2	**1.0**	**1.0**	**1.0**	**1.0**	.24	**1.0**	.34	**1.0**	**1.0**	.29	**1.0**	**1.0**	**1.0**	.39

Table 12: Estimated posterior means of the association indicators in the Ising model and the Ising mixture model with two components for the Rochdale data.

fit the data well (p-value < 0.00001), while the Ising mixture model (4.2) with two components fits the data well (p-value 0.21). The right panel of Table 11 shows the cells containing the 10 largest observed counts observed in the NLTCS data together with their expected cell counts in the Ising model and the Ising mixture model. We see that the Ising mixture model seems to capture the size of the largest cell counts more faithfully than the Ising model.

4.5 Identifiability of Ising mixture models

In this chapter we focus on exploring the identifiability of Ising mixture models from a theoretical perspective. The non-identifiability of the probability mass function given the association indicators, i.e., $\pi(\boldsymbol{n} \mid \boldsymbol{\Gamma})$, arises from the non-identifiability of the probability mass function of Ising mixture models, i.e., $\pi(\boldsymbol{n} \mid \boldsymbol{\Theta}, \boldsymbol{w})$. We start with a thorough review of closely related existing results and methods in the literature, and explain why their application to Ising mixture models is challenging. Then we propose some specific sufficient conditions and necessary conditions for the identifiability of Ising mixture models. We also discuss several examples that illustrate specific cases of key interest. Proofs of all the theoretical results are given in the Appendix.

Manole and Khalili (2021, Corollary 1) provide a sufficient condition for the identifiability of finite mixtures of multinomial distributions. Ising mixture models assume that each cell count follows a multinomial distribution determined by cell probabilities, rather than a mixture of multinomial distributions – the situation studied in (Manole and Khalili, 2021). Thus their results are not directly applicable in our setting. Other related results in literature are based on the conditional independence assumption, such as Allman et al.

4419	97	67	472	2063	55	335	44	313	18	33	76	1	5	2	6
119	115	1	16	0	4	1189	17	112	6	130	64	529	52	453	56
2	22	13	116	10	67	47	0	2	0	1	92	0	4	0	4
1	12	5	19	1	0	0	3	1	0	354	2	27	4	16	5
55	3	24	1	0	0	1	7	1	60	667	29	601	14	1	16
3	55	8	85	7	65	69	400	24	5	62	2	10	164	0	8
2	6	3	15	3	5	0	0	0	0	0	0	3	32	4	41
1	0	1	0	1	0	4	1	3	3	9	0	0	0	0	0
14	226	11	140	5	0	2	2	10	0	7	3	3	11	31	3
0	4	0	2	125	8	134	81	654	34	0	34	1	5	25	215
8	80	30	5	105	19	50	1	1	0	2	3	0	3	0	3
13	9	4	1	0	0	4	7	1	6	6	54	3	0	1	0
0	1	6	0	6	1	42	3	28	48	207	12	0	5	0	2
4	34	1	13	6	38	549	19	180	21	196	27	2	14	72	88
8	3	0	0	2	8	11	3	15	9	5	19	3	26	0	28
29	158	10	89	5	66	764	66	86	8	175	7	151	131	516	1056

Table 13: The NLTCS data. The cells counts appear row by row in lexicographical order with Variable 8 varying fastest and Variable 1 varying slowest.

(2009) and Xu (2017). In their work, the conditional independence assumption allows for the transfer of the identifiability question to an equivalent one with fewer variables and more levels with the help of the row-wise tensor product. In the case of three categorical variables, the conditional independence assumption further enables the identifiability question to be transformed into an equivalent one by considering the rank of matrices using the triple product. However, these methods are not directly applicable for Ising mixture models since the joint probability conditional on a mixture component cannot be written as a product of marginal probabilities.

4.5.1 Examples and main results related to identifiability

Definition 4.1. *An Ising mixture model parameterized by weights $\boldsymbol{w}$ as well as main and interaction terms $\boldsymbol{\Theta}$ is identifiable if and only if different $\boldsymbol{w}$, $\boldsymbol{\Theta}$ imply different cell probabilities $\boldsymbol{p}_{\text{mix}}$.*

Definition 4.2 (Definition 3 in Rothenberg (1971))**.** *A Ising mixture model is locally identifiable at a parameter point $\underline{\boldsymbol{w}}, \underline{\boldsymbol{\Theta}}$ if and only if there exists an open neighborhood of $\underline{\boldsymbol{w}}, \underline{\boldsymbol{\Theta}}$ containing no other parameter $\boldsymbol{w}, \boldsymbol{\Theta}$ implying the same cell probabilities $\boldsymbol{p}_{\text{mix}}$ as $\underline{\boldsymbol{w}}, \underline{\boldsymbol{\Theta}}$.*

In the sequel the identifiability is studied based on the following assumption.

Assumption 4.1. *The main effects vectors $(\theta_v^{(k)} : v \in [d])^T$ are identical for all $k \in [K]$.*

This assumption arises from the similarity among subpopulations. Although it is assumed that each cell probability is a mixture of different components, we don't want to assume that these components are entirely different from each other. By assuming identical main effects across all components, we can account

Model	γ_{12}	γ_{13}	γ_{14}	γ_{15}	γ_{16}	γ_{17}	γ_{18}	γ_{23}	γ_{24}	γ_{25}	γ_{26}	γ_{27}	γ_{28}	γ_{34}
Ising model	**1.0**	**.90**	**.85**	.32	.12	**1.0**	.11	**1.0**	**1.0**	.29	**.98**	**1.0**	.28	**1.0**
Ising mixture, Component 1	**1.0**	**1.0**	.16	**.70**	.35	.27	.46	**1.0**	**.97**	.33	**.72**	.28	**1.0**	**1.0**
Ising mixture, Component 2	**1.0**	.16	**1.0**	.18	.38	**1.0**	.22	**1.0**	**1.0**	.16	1.0	**1.0**	.21	**1.0**

Model	γ_{35}	γ_{36}	γ_{37}	γ_{38}	γ_{45}	γ_{46}	γ_{47}	γ_{48}	γ_{56}	γ_{57}	γ_{58}	γ_{67}	γ_{68}	γ_{78}
Ising model	.46	**1.0**	0.1	**1.0**	**.93**	**1.0**	**1.0**	**1.0**	**1.0**	**1.0**	**1.0**	**1.0**	**1.0**	**1.0**
Ising mixture, Component 1	**1.0**	.26	.18	**1.0**	**1.0**	**1.0**	**1.0**	**1.0**	.27	**1.0**	**1.0**	**1.0**	**1.0**	.16
Ising mixture, Component 2	.11	**1.0**	**.99**	.15	.30	**1.0**	**.99**	**1.0**	**1.0**	**1.0**	**1.0**	**1.0**	**1.0**	**1.0**

Table 14: The posterior means of $\boldsymbol{\gamma}^{(1)}$ and $\boldsymbol{\gamma}^{(2)}$ inferred by two-component Ising mixture models for the NLTCS data. The number of component in the normal mixture sampling distribution is $J = 5$. The posterior mean of the weight of the first component, i.e. $E(w^{(1)} \mid \boldsymbol{n})$, is 0.4.

for the similarity among components and also allow for heterogeneity of interaction effects in different components. Lemma 4.1 shows that this assumption simplifies the study of sufficient conditions and necessary conditions, allowing us to bypass main effects and focus on the identifiability of interaction effects.

Lemma 4.1. *When studying the identifiability of Ising mixture models, Assumption 4.1 allows us to assume that all main effects are equal to* 0 *without affecting the overall generality of the analysis.*

In what follows the parameter vector $\boldsymbol{\theta}$ includes only interaction effects for each component, i.e., $\boldsymbol{\theta}^{(k)} = (\theta^{(k)}_{v'v} : v' < v)^T$ for each $k \in [K]$.

The second assumption arises when interaction effects in only one component are unknown.

Assumption 4.2. *All interaction effects in every component except the first component are known. In other words,* $\boldsymbol{\theta}^{(k)}$ *are fixed and known for all* $k \geq 2$.

However, the following example shows that this assumption is not sufficient for the local identifiability of the Ising mixture model.

Example 4.2. Suppose $d = 2$ and $\theta^{(2)}_{12} = 0$. Then this mixture model is not locally identifiable for any $\theta^{(1)}_{12} \in \mathbb{R}$ and $w^{(1)} \in (0, 1)$.

Therefore, we need another assumption on the weights of components $\boldsymbol{w}$.

Assumption 4.3. *The weights of components* $w^{(k)} \in (0, 1), k \in [K]$ *are fixed and known.*

The following proposition states our first sufficient conditions for the local identifiability of Ising mixture models, which is particularly useful when some unknown subpopulation is mixed with other well-known populations.

Proposition 4.1. *Assumptions 4.1, 4.2 and 4.3 are sufficient for local identifiability of Ising mixture models.*

Next we connect Ising mixture models with mixtures of graphical structures. We represent an Ising model with parameters $\boldsymbol{\theta}$ by an undirected graph $G(\boldsymbol{\theta}) := G(V, \mathbb{E}(\boldsymbol{\theta}))$. In this representation, vertices $V = \{1, 2, \ldots, d\}$ are associated with each variable, and edges $\mathbb{E}(\boldsymbol{\theta}) := \{(v', v) : v' < v, \theta_{v'v} \neq 0\}$ are associated with each non-zero pairwise interaction. Each missing edge corresponds with a pairwise interaction effect that is zero. The edges in $\mathbb{E}(\boldsymbol{\theta})$ are called activation edges. We also define $\mathbb{V}(\boldsymbol{\theta})$ as the set of vertices with degree at least 1. The vertices in $\mathbb{V}(\boldsymbol{\theta})$ are called activation vertices or activation variables. We define the projection of the graph $G(\boldsymbol{\theta})$ onto its activation variables $\mathbb{V}(\boldsymbol{\theta})$ as the subgraph of $G(\boldsymbol{\theta})$ determined by $\mathbb{V}(\boldsymbol{\theta})$. This projection is denoted by $G(\boldsymbol{\theta} \mid \mathbb{V}(\boldsymbol{\theta}))$.

The graphical representation of Ising mixture models are constructed at the level of their mixture components. For component $k \in [K]$ with parameters $\boldsymbol{\theta}^{(k)}$, we construct its undirected graph $G(\boldsymbol{\theta}^{(k)})$. The set of activation variables and activation edges for component k are denoted by $\mathbb{V}(\boldsymbol{\theta}^{(k)})$ and $\mathbb{E}(\boldsymbol{\theta}^{(k)})$, respectively. To illustrate these definitions, consider the following example.

Example 4.3. Suppose $d = 4$, $K = 2$, $\theta^{(1)}_{v'v} = 0$ for all $(v', v) \neq (1, 2)$ and $\theta^{(2)}_{v'v} = 0$ for all $(v', v) \neq (3, 4)$. Then $G(\boldsymbol{\theta}^{(1)}) = (\{1, 2, 3, 4\}, \{(1, 2)\})$ and $G(\boldsymbol{\theta}^{(2)}) = (\{1, 2, 3, 4\}, \{(3, 4)\})$. The activation variables in component 1 are $\mathbb{V}(\boldsymbol{\theta}^{(1)}) = \{1, 2\}$. The activation variables in component 2 are $\mathbb{V}(\boldsymbol{\theta}^{(2)}) = \{3, 4\}$. The activation edges in component 1 are $\mathbb{E}(\boldsymbol{\theta}^{(1)}) = \{(1, 2)\}$. The activation edges in component 2 are $\mathbb{E}(\boldsymbol{\theta}^{(2)}) = \{(3, 4)\}$. $G(\boldsymbol{\theta}^{(1)} \mid \mathbb{V}(\boldsymbol{\theta}^{(1)})) = (\{1, 2\}, \{(1, 2)\})$ and $G(\boldsymbol{\theta}^{(2)} \mid \mathbb{V}(\boldsymbol{\theta}^{(2)})) = (\{3, 4\}, \{(3, 4)\})$. Please see Figure 14.

We now are prepared to formulate an essential assumption for identifiability based on graphical representations and activation variables.

Assumption 4.4. *The activation variables in different components of an Ising mixture model are mutually exclusive, i.e.,* $\bigcap_{k \in [K]} \mathbb{V}(\boldsymbol{\theta}^{(k)}) = \emptyset$.

The next result provides sufficient conditions for identifiability. This result is particularly useful when different components have different activation variables.

Proposition 4.2. *Assumptions 4.1, 4.3 and 4.4 are jointly sufficient for local identifiability of Ising mixture model.*

The following example illustrates that Assumption 4.4 alone does not guarantee local identifiability, and therefore Assumption 4.3 is required for the validity of Proposition 4.2.

Example 4.4. Suppose $K = 2$, $d = 4$. Only $\theta_{12}^{(1)}$, $\theta_{34}^{(2)}$ and $w^{(1)}$ are the nonzero unknown parameters. All other interaction effects are fixed at zero. The resulting mixture model is not locally identifiable for any $\theta_{12}^{(1)}, \theta_{34}^{(2)} \in \mathbb{R}$ and $w^{(1)} \in (0, 1)$.

Based on this example, it can be immediately inferred that any two-component Ising mixture model with at least one non-zero interaction effect in each component is not locally identifiable in general.

The following example shows that Assumption 4.3 is not sufficient enough for local identifiability.

Example 4.5. Let $d = 4$, $w^{(1)} = w^{(2)} = .5$ and $\theta_{13}^{(k)} = \theta_{14}^{(k)} = \theta_{23}^{(k)} = \theta_{24}^{(k)} = 0$ for $k = 1, 2$. The unknown parameters are $\theta_{12}^{(k)}, \theta_{34}^{(k)}$ for $k = 1, 2$ only. Then this mixture model is not locally identifiable by empirical verification.

4.6 Discussion

In this paper we developed finite mixtures of Ising within a Bayesian framework as an effective alternative to infer associations between binary variables. By combining Ising models with multivariate Bernoulli mixture models, our contribution addresses the current gap in the literature between loglinear models and various types of mixture models for categorical data. There are several key reasons why addressing this gap was a worthwhile effort. First, Ising mixture models not only effectively fit sparse data, but also offer interpretable results. If data are generated from an Ising mixture model with sparse interaction effects, using Ising models can result in denser and less confident interaction effects. Although Ising mixture models have more parameters than Ising models, they often infer fewer but more significant non-zero interaction effects. This feature of Ising mixture models was illustrated in the simulations experiments.

Second, Ising mixture models can be viewed as an extension of multivariate Bernoulli mixture models, breaking the conditional independence assumption by introducing interaction effects for each component. Inferring interaction effects from Ising mixture models can lead to the identification of mixtures of graphical loglinear models, providing insight into multivariate patterns of associations of subpopulations. As graphical models become increasingly popular in fields such as social networks, it is likely that Ising mixture models will also gain attention in these areas. Furthermore, the development of Ising mixture models can potentially contribute to the development of more general finite mixtures of graphical loglinear models.

Ising mixture models are a powerful tool to handle multi-modal spike-and-slab posteriors in data. Research has shown that mixture models can be used to approximate these multi-modal posteriors in linear regressions (Ročková, 2018). Reporting a single model is a misleading reflection of overall model uncertainty. We studied Ising mixture models with spike-and-slab prior distributions to infer associations between

binary variables. We have shown that our framework is not only effective in fitting sparse contingency tables, but also leads to interpretable results. We established sufficient and necessary conditions for the identifiability of Ising mixture models without relying on the assumption of conditional independence. More work is certainly needed to propose general conditions for identifiability of Ising mixture models, that will lead to improving the interpretation of the inferred associations and reduce the risk of overfitting.

Our proposed framework can be extended in at least two key directions. It can be generalized to handle categorical random variables with more than two levels starting from the Potts model (Wu, 1982), although this should be done with care given the increase in the number of parameters. Furthermore, the inclusion of higher-order interaction terms can also be beneficial to allow the study of more complex interaction patterns of associations.

4.7 Appendix

Proof of Lemma 4.1. Suppose $\underline{\boldsymbol{w}}, \underline{\boldsymbol{\Theta}}, \underline{\boldsymbol{\Gamma}}$ are true values of parameters in the Ising mixture model. It follows from the likelihood equation of $\boldsymbol{X} = \boldsymbol{0}$ as well as $X_1 = 1, X_2 = \ldots = X_d = 0$ that

$$\sum_{k\in[K]} \frac{w_k}{Z_k} = \sum_{k\in[K]} \frac{\underline{w}_k}{\underline{Z}_k} \text{ as well as } \sum_{k\in[K]} \frac{w_k \exp(\theta_1)}{Z_k} = \sum_{k\in[K]} \frac{\underline{w}_k \exp(\underline{\theta}_1)}{\underline{Z}_k},$$

where $Z_k, \underline{Z}_k, k \in [K]$, are normalization constants. Dividing the second equation by the first equation, it follows that $\theta_1 = \underline{\theta}_1$. It then follows from analogous arguments that $\theta_k = \underline{\theta}_k$ for all $k \in [K]$. This lemma then follows immediately from the claim that all likelihood equations with the true value of main effects are equivalent with likelihood equations with main effects 0. □

Proof of Example 4.2. Let $\underline{\theta}_{12}^{(1)}$ and $\underline{w}^{(1)}$ be the true value of parameters. Then we only need to prove that the solutions to the following equations are not unique:

$$\frac{w^{(1)}}{3+\eta_{12}^{(1)}} + \frac{1-w^{(1)}}{4} = p_{00} = \frac{\underline{w}^{(1)}}{3+\underline{\eta}_{12}^{(1)}} + \frac{1-\underline{w}^{(1)}}{4} \text{ and } \frac{w^{(1)}\eta_{12}^{(1)}}{3+\eta_{12}^{(1)}} + \frac{1-w^{(1)}}{4} = p_{11} = \frac{\underline{w}^{(1)}\underline{\eta}_{12}^{(1)}}{3+\underline{\eta}_{12}^{(1)}} + \frac{1-\underline{w}^{(1)}}{4},$$

where $\eta_{12}^{(1)} = \exp(\theta_{12}^{(1)})$ and $\underline{\eta}_{12}^{(1)} = \exp(\underline{\theta}_{12}^{(1)})$. It follows from the second equation that

$$\eta_{12}^{(1)} = \underline{\eta}_{12}^{(1)} + \frac{(1-\underline{\eta}_{12}^{(1)})(w^{(1)}-\underline{w}^{(1)})}{\frac{\underline{w}^{(1)}}{3+\underline{\eta}_{12}^{(1)}} + \frac{w^{(1)}-\underline{w}^{(1)}}{4}}.$$

Replace $\eta_{12}^{(1)}$ with this identity in the first equation, it follows that

$$\frac{w^{(1)}}{3+\eta_{12}^{(1)}} + \frac{1-w^{(1)}}{4} = \frac{\underline{w}^{(1)}}{3+\underline{\eta}_{12}^{(1)}} + \frac{1-\underline{w}^{(1)}}{4}.$$

Therefore, this model is not locally identifiable. □

Proof of Proposition 4.1. Let $\underline{\boldsymbol{w}}$ denote the true value of the weights $\boldsymbol{w}$ and it is known by assumptions. Analogously, let $\underline{\boldsymbol{\theta}}^{(k)}$ denote the true value of interaction effects $\boldsymbol{\theta}^{(k)}$ in the k-th component. The interaction effect in the first component $\boldsymbol{\theta}^{(1)}$ is unknown and its true value is denoted by $\underline{\boldsymbol{\theta}}^{(1)}$.

It then follows from the identity of the cell probabilities, i.e.,

$$\boldsymbol{p}_{\text{mix}}(\boldsymbol{\theta}^{(1)}, \underline{\boldsymbol{w}}, \underline{\boldsymbol{\theta}}^{(2)}, \ldots, \underline{\boldsymbol{\theta}}^{(K)}) = \boldsymbol{p}_{\text{mix}}(\underline{\boldsymbol{\theta}}^{(1)}, \underline{\boldsymbol{w}}, \underline{\boldsymbol{\theta}}^{(2)}, \ldots, \underline{\boldsymbol{\theta}}^{(K)})$$

that

$$\underline{w}^{(1)}\boldsymbol{p}(\boldsymbol{\theta}^{(1)}) + \sum_{2\le k\le K} \underline{w}^{(k)}\boldsymbol{p}(\underline{\boldsymbol{\theta}}^{(k)}) = \underline{w}^{(1)}\boldsymbol{p}(\underline{\boldsymbol{\theta}}^{(1)}) + \sum_{2\le k\le K} \underline{w}^{(k)}\boldsymbol{p}(\underline{\boldsymbol{\theta}}^{(k)}),$$

and hence

$$\boldsymbol{p}(\boldsymbol{\theta}^{(1)}) = \boldsymbol{p}(\underline{\boldsymbol{\theta}}^{(1)})$$

given $\underline{w}^{(1)} > 0$. It then follows from the identifiability of Ising model that $\boldsymbol{\theta}^{(1)} = \underline{\boldsymbol{\theta}}^{(1)}$. □

Proof of Proposition 4.2. Without loss of generality we assume all main effects are zero by Lemma 4.1. Let $\underline{\boldsymbol{\theta}}^{(k)}$ denote the true value of $\boldsymbol{\theta}^{(k)}$ for $k \in [K]$. The key step is to show that

$$p_{\boldsymbol{i}}(\boldsymbol{\theta}^{(k)}) - p_{\boldsymbol{0}}(\boldsymbol{\theta}^{(k)}) = p_{\boldsymbol{i}}(\underline{\boldsymbol{\theta}}^{(k)}) - p_{\boldsymbol{0}}(\underline{\boldsymbol{\theta}}^{(k)}) \tag{4.6}$$

for any $\boldsymbol{i} \in I$ and $k \in [K]$. If these equations hold, we then sum them up over all $\boldsymbol{i} \in I$ and we immediately have $1 - 2^d p_{\boldsymbol{0}}(\boldsymbol{\theta}^{(k)}) = 1 - 2^d p_{\boldsymbol{0}}(\underline{\boldsymbol{\theta}}^{(k)})$ or $p_{\boldsymbol{0}}(\boldsymbol{\theta}^{(k)}) = p_{\boldsymbol{0}}(\underline{\boldsymbol{\theta}}^{(k)})$. As a result, we have $p_{\boldsymbol{i}}(\boldsymbol{\theta}^{(k)}) = p_{\boldsymbol{i}}(\underline{\boldsymbol{\theta}}^{(k)})$ and hence $\boldsymbol{\theta}^{(k)} = \underline{\boldsymbol{\theta}}^{(k)}$ by the identifiability of Ising models.

It suffices to prove (4.6) for $k = 1$. Arguments for remaining cases are analogous and hence omitted. For each $\boldsymbol{i} \in I$, define $\mathbb{V}(\boldsymbol{i}) := \{v : v \in [d], i_v = 1\}$ as activated variables for cell $\boldsymbol{i}$. Then $G(\boldsymbol{\theta} \mid \mathbb{V}(\boldsymbol{i}))$ is the projection of graph $G(\boldsymbol{\theta})$ on $\mathbb{V}(\boldsymbol{i})$ and $G(\boldsymbol{\theta} \mid \mathbb{V}(\boldsymbol{i}) \bigcap \mathbb{V}(\boldsymbol{\theta}))$ is the projection on activated variables $\mathbb{V}(\boldsymbol{i})$ and activation variables $\mathbb{V}(\boldsymbol{\theta})$. The probability of cell $\boldsymbol{i}$ can be written as

$$p_{\boldsymbol{i}}(\boldsymbol{\theta}) = \exp\Big(\sum_{(v',v)\in\mathcal{E}(G(\boldsymbol{\theta}|\mathbb{V}(\boldsymbol{i})))} \theta_{v'v}\Big)/Z = \exp\Big(\sum_{(v',v)\in\mathcal{E}(G(\boldsymbol{\theta}|\mathbb{V}(\boldsymbol{i})\bigcap\mathbb{V}(\boldsymbol{\theta})))} \theta_{v'v}\Big)/Z,$$

where $Z = Z(\boldsymbol{\theta})$ is a normalization constant depending on $\boldsymbol{\theta}$.

As a consequence, for any fixed $\boldsymbol{i} = (i_1, \ldots, i_d)^T$, we have

$$p_{\boldsymbol{i}}(\boldsymbol{\theta}^{(1)}) = \exp\Big(\sum_{(v',v)\in\mathcal{E}(G(\boldsymbol{\theta}^{(1)}|\mathbb{V}(\boldsymbol{i})\bigcap\mathbb{V}(\boldsymbol{\theta}^{(1)})))} \theta^{(1)}_{v'v}\Big)/Z^{(1)},$$

where $G(\boldsymbol{\theta}^{(1)})$ is the undirected graph associated with an Ising model with parameters $\boldsymbol{\theta}^{(1)}$, $\mathbb{V}(\boldsymbol{i})$ are activated variables for cell i, $\mathbb{V}(\boldsymbol{\theta}^{(1)})$ are activation variables for parameters $\boldsymbol{\theta}^{(1)}$, and $Z^{(1)} = Z(\boldsymbol{\theta}^{(1)})$ is the normalization constant.

Let $\boldsymbol{i}^{(1)}$ be the cell such that $\mathbb{V}(\boldsymbol{i}^{(1)}) = \mathbb{V}(\boldsymbol{i}) \bigcap \mathbb{V}(\boldsymbol{\theta}^{(1)})$ and we immediately have

$$G(\boldsymbol{\theta}^{(1)} \mid \mathbb{V}(\boldsymbol{i}^{(1)}) \bigcap \mathbb{V}(\boldsymbol{\theta}^{(1)})) = G(\boldsymbol{\theta}^{(1)} \mid \mathbb{V}(\boldsymbol{i}) \bigcap \mathbb{V}(\boldsymbol{\theta}^{(1)}) \bigcap \mathbb{V}(\boldsymbol{\theta}^{(1)})) = G(\boldsymbol{\theta}^{(1)} \mid \mathbb{V}(\boldsymbol{i}) \bigcap \mathbb{V}(\boldsymbol{\theta}^{(1)})).$$

Therefore, we have

$$\begin{aligned} p_{\boldsymbol{i}^{(1)}}(\boldsymbol{\theta}^{(1)}) &= \exp\Big(\sum_{(v',v) \in \mathcal{E}(G(\boldsymbol{\theta}^{(1)} | \mathbb{V}(\boldsymbol{i}^{(1)}) \bigcap \mathbb{V}(\boldsymbol{\theta}^{(1)})))} \theta^{(1)}_{v'v} \Big) / Z^{(1)} \\ &= \exp(\sum_{(v',v) \in \mathcal{E}(G(\boldsymbol{\theta}^{(1)} | \mathbb{V}(\boldsymbol{i}) \bigcap \mathbb{V}(\boldsymbol{\theta}^{(1)})))} \theta^{(1)}_{v'v}) / Z^{(1)} \\ &= p_{\boldsymbol{i}}(\boldsymbol{\theta}^{(1)}). \end{aligned}$$

Now let's consider $p_{\boldsymbol{i}^{(1)}}(\boldsymbol{\theta}^{(k)})$ for any $k \neq 1$. Note that $\mathbb{V}(\boldsymbol{i}^{(1)}) = \mathbb{V}(\boldsymbol{i}) \bigcap \mathbb{V}(\boldsymbol{\theta}^{(1)}) \subset \mathbb{V}(\boldsymbol{\theta}^{(1)})$ and $\mathbb{V}(\boldsymbol{\theta}^{(1)}) \bigcap \mathbb{V}(\boldsymbol{\theta}^{(k)}) = \emptyset$ by Assumption 4.4. We then have $\mathbb{V}(\boldsymbol{i}^{(1)}) \bigcap \mathbb{V}(\boldsymbol{\theta}^{(k)}) = \emptyset$. Therefore, for $k \neq 1$ we have

$$p_{\boldsymbol{i}^{(1)}}(\boldsymbol{\theta}^{(k)}) = \exp\Big(\sum_{(v',v) \in \mathcal{E}(G(\boldsymbol{\theta}^{(k)} | \mathbb{V}(\boldsymbol{i}^{(1)}) \bigcap \mathbb{V}(\boldsymbol{\theta}^{(k)})))} \theta^{(k)}_{v'v} \Big) / Z^{(k)} = \exp(0) / Z^{(k)} = 1 / Z^{(k)}.$$

Then the probability of $\boldsymbol{X} = \boldsymbol{i}^{(1)}$ is

$$\underline{w}^{(1)} p_{\boldsymbol{i}^{(1)}}(\boldsymbol{\theta}^{(1)}) + \sum_{k \geq 2} \underline{w}^{(k)} p_{\boldsymbol{i}^{(1)}}(\boldsymbol{\theta}^{(k)}) = \underline{w}^{(1)} p_{\boldsymbol{i}}(\boldsymbol{\theta}^{(1)}) + \sum_{k \geq 2} \underline{w}^{(k)} / Z^{(k)}.$$

This is also true with replacing $\boldsymbol{\theta}^{(k)}$ by its true value $\underline{\boldsymbol{\theta}}^{(k)}$. Therefore, the cell probability equation in the case of $\boldsymbol{X} = \boldsymbol{i}^{(1)}$, i.e.,

$$\underline{w}^{(1)} p_{\boldsymbol{i}^{(1)}}(\boldsymbol{\theta}^{(1)}) + \sum_{k \geq 2} \underline{w}^{(k)} p_{\boldsymbol{i}^{(1)}}(\boldsymbol{\theta}^{(k)}) = \underline{w}^{(1)} p_{\boldsymbol{i}^{(1)}}(\underline{\boldsymbol{\theta}}^{(1)}) + \sum_{k \geq 2} \underline{w}^{(k)} p_{\boldsymbol{i}^{(1)}}(\underline{\boldsymbol{\theta}}^{(k)})$$

can be written as

$$\underline{w}^{(1)} p_{\boldsymbol{i}}(\boldsymbol{\theta}^{(1)}) + \sum_{k \geq 2} \underline{w}^{(k)} / Z^{(k)} = \underline{w}^{(1)} p_{\boldsymbol{i}}(\underline{\boldsymbol{\theta}}^{(1)}) + \sum_{k \geq 2} \underline{w}^{(k)} / \underline{Z}^{(k)}, \tag{4.7}$$

where $\underline{Z}^{(k)}$ is the normalization constant corresponding with $\underline{\boldsymbol{\theta}}^{(k)}$. Considering the cell probability equation in the case of $\boldsymbol{X} = \mathbf{0} := (0, \cdots, 0)$, it follows from $p_{\mathbf{0}}(\boldsymbol{\theta}^{(k)}) = 1/Z^{(k)}$ that

$$\underline{w}^{(1)} p_{\mathbf{0}}(\boldsymbol{\theta}^{(1)}) + \sum_{k \geq 2} \underline{w}^{(k)} / Z^{(k)} = \underline{w}^{(1)} p_{\mathbf{0}}(\underline{\boldsymbol{\theta}}^{(1)}) + \sum_{k \geq 2} \underline{w}^{(k)} / \underline{Z}^{(k)}.$$

Then the difference between (4.7) and the identity above implies

$$\underline{w}^{(1)} p_{\boldsymbol{i}}(\boldsymbol{\theta}^{(1)}) - \underline{w}^{(1)} p_{\mathbf{0}}(\boldsymbol{\theta}^{(1)}) = \underline{w}^{(1)} p_{\boldsymbol{i}}(\underline{\boldsymbol{\theta}}^{(1)}) - \underline{w}^{(1)} p_{\mathbf{0}}(\underline{\boldsymbol{\theta}}^{(1)}),$$

which gives (4.6). □

Proof of Example 4.4. To simplify notations, let $\boldsymbol{\eta}^{(k)} := \exp(\boldsymbol{\theta}^{(k)})$, $k = 1, 2$. Suppose $\underline{\eta}_{12}^{(1)}, \underline{\eta}_{34}^{(2)}$ and $\underline{w}^{(1)}$ are true parameters. It then follows from the likelihood equations that

$$\begin{aligned}
\frac{w^{(1)}}{3+\eta_{12}^{(1)}} + \frac{1-w^{(1)}}{3+\eta_{34}^{(2)}} &= \frac{\underline{w}^{(1)}}{3+\underline{\eta}_{12}^{(1)}} + \frac{1-\underline{w}^{(1)}}{3+\underline{\eta}_{34}^{(2)}}, \\
\frac{w^{(1)}\eta_{12}^{(1)}}{3+\eta_{12}^{(1)}} + \frac{1-w^{(1)}}{3+\eta_{34}^{(2)}} &= \frac{\underline{w}^{(1)}\underline{\eta}_{12}^{(1)}}{3+\underline{\eta}_{12}^{(1)}} + \frac{1-\underline{w}^{(1)}}{3+\underline{\eta}_{34}^{(2)}}, \\
\frac{w^{(1)}}{3+\eta_{12}^{(1)}} + \frac{(1-w^{(1)})\eta_{34}^{(2)}}{3+\eta_{34}^{(2)}} &= \frac{\underline{w}^{(1)}}{3+\underline{\eta}_{12}^{(1)}} + \frac{(1-\underline{w}^{(1)})\underline{\eta}_{34}^{(2)}}{3+\underline{\eta}_{34}^{(2)}}.
\end{aligned}$$

From the first two equations it follows that

$$\eta_{12}^{(1)} = \underline{\eta}_{12}^{(1)} + \frac{(w^{(1)} - \underline{w}^{(1)})(1-\underline{\eta}_{12}^{(1)})(\underline{\eta}_{12}^{(1)}+3)}{w^{(1)}(\underline{\eta}_{12}^{(1)}+3) + \underline{w}^{(1)}(1-\underline{\eta}_{12}^{(1)})},$$

where $\eta_{12}^{(1)}$ is represented as a function of $w^{(1)}$. Analogously, it follows from the first and the third equations that

$$\eta_{34}^{(2)} = \underline{\eta}_{34}^{(2)} - \frac{(w^{(1)} - \underline{w}^{(1)})(1-\underline{\eta}_{34}^{(2)})(\underline{\eta}_{34}^{(2)}+3)}{(1-w^{(1)})(\underline{\eta}_{34}^{(2)}+3) + (1-\underline{w}^{(1)})(1-\underline{\eta}_{34}^{(2)})},$$

where $\eta_{34}^{(2)}$ is represented as a function of $w^{(1)}$. After some algebra we can see that $\eta_{12}^{(1)}$ and $\eta_{34}^{(2)}$ satisfy the first equation without other constraints. In other words, we can construct different values of parameters based on $w^{(1)} \in (0, 1)$ with the same cell probability. □

Proof of Example 4.5. In this example, the nonidentifiability is justified by the singularity of the information matrix based on the equivalence between the local identifiability and the rank of the information matrix. Some examples of Ising mixture models with singular information matrices are in Table 15. □

Index	$\theta_{12}^{(1)}$	$\theta_{34}^{(1)}$	$\theta_{12}^{(2)}$	$\theta_{34}^{(2)}$	eigenvalues of its Fisher information matrix
1	0.5	0.5	0.1	0.1	0.118389, 0.116171, 0.00220142, 3.04905e-17
2	0.5	0.5	0.1	1	0.117381, 0.102194, 0.00242613, -9.52222e-17
3	3	0.5	0.1	1	0.104261, 0.084175, 0.0228727, 1.59007e-16

Table 15: Examples of Ising mixture models with singular information matrices.

5 Contributions and discussions for future research

In this thesis, we introduce and study three models (a) Bi-s^*-concave models (b) nonparametric Poisson mixture models and (c) Ising mixture models. In this last chapter, we summarize the main contributions and discuss potential directions for further research for each of them.

5.1 Bi-s^*-concave models

We present a new class of distribution functions with shape constraints, the bi-s^*-concave classes, which extends the bi-log-concave class by including heavy-tailed distributions and the s-concave class by accommodating bimodal distributions. This class encompasses many commonly used parametric distributions, such as Student-t, F, Pareto, symmetrized Beta, and mixture of Gaussians shifted and shifted Student-t, among others. In this work, we provide a characterization of these new classes and connect them to the Csörgő - Révész constant. We also demonstrate how the bi-s^*-concave classes can be used to define refined confidence bands for distribution functions, taking advantage of the shape constraint to achieve more accurate and honest coverage. Our theoretical analysis shows that the refined confidence bands are Lipschitz continuous over the entire real line, with the lower band approaching zero polynomially fast as its input approaches negative infinity and the upper band approaching one polynomially fast as its input approaches positive infinity. Additionally, we establish the consistency and rate of convergence of the confidence bands for functionals with well-behaved integrands.

However, the current refined confidence bands are dependent on knowing the correct value of s^*, which is typically not available in practice. An interesting area for future research is the development of methods with theoretical guarantees for estimating s^*. Furthermore, while there are results on estimating log-concave and s-concave distributions, there is little work on estimating bi-s^*-concave distributions. Valuable questions to explore in this regard include: (a) does a maximum likelihood estimator exist, (b) if it exists, can it be characterized using finite parameters, and what is an efficient algorithm for computing this estimator, and (c) what are the asymptotic properties of this estimator, including its consistency, rates of convergence, and asymptotic distribution. Additionally, as both the log-concave and s-concave classes are naturally defined in multi-dimensional spaces, extending the bi-s^*-concave class to multiple dimensions is another important area for further research. One interesting question to investigate is whether the marginal or conditional distributions remain bi-s^*-concave if the joint distribution is bi-s^*-concave. Such properties play a critical role in the inference of the log-concave class.

5.2 Nonparametric Poisson mixture models

Our study focuses on nonparametric Poisson mixture models adapted for single-cell RNA sequencing data. In terms of theoretical contributions, we establish the existence and consistency of the adapted nonparametric maximum likelihood estimators. Additionally, we investigate and establish the minimax convergence rate of the maximum likelihood estimators of nonparametric Poisson mixture models. In terms of computation, we provide three algorithms for obtaining the nonparametric maximum likelihood estimators. We conduct time complexity analyses for each algorithm and perform simulations and real-data applications to demonstrate their effectiveness.

As a next step for further research, it would be interesting to extend the current model to multi-dimensional data. Possible areas of investigation include exploring potential estimators and studying their theoretical properties (such as existence, consistency, and rates of convergence). Additionally, we need to develop algorithms for computing these estimators and study their time complexity. Finally, it would be worthwhile to implement this approach on single-cell RNA sequencing data. Currently, we can only analyze data gene-by-gene, but it would be valuable if we could examine genes together and explore their associations. An additional area for exploration in this work is the rate of convergence for the nonparametric maximum likelihood estimator for a range of nonparametric power series mixture models, including nonparametric geometric mixture models and nonparametric negative binomial mixture models. Such an investigation would yield convergence rates for a nonparametric discrete family of distributions, rather than for a single specific distribution.

5.3 Ising mixture models

Our proposed Ising mixture models, featuring spike-and-slab prior distributions, offer a means of inferring associations between binary variables. In our study, we explored the sufficient and necessary conditions for the identifiability of these models. Additionally, we investigated the meaning of the posterior mean of the indicator associations, even in cases where the model is non-identifiable. By combining the Ising model with multivariate Bernoulli mixture models, our proposed model not only effectively fits sparse contingency tables, but also yields interpretable results.

As a next step in our research, it would be valuable to propose and investigate conditions for the identifiability of the Ising mixture model. We could also extend our current proposal to multivariate categorical data with more than two levels, which would increase the number of parameters and present new challenges for identifying the model and developing efficient fitting algorithms. Additionally, we could incorporate higher order interaction effects beyond two-way interactions, thus further increasing the complexity of the model

and the number of parameters, which would again pose significant challenges for identifiability and fitting algorithms.

Another further research direction is to relax the constraint of fixing the number of components relax by assuming a prior distribution. The following notations are the same as in Chapter 4. For example, Hatjispyros et al. (2023) propose expressing decreasing weights $w^{(k)}k \geq 1$ as $w^{(k)} = P(Y \geq k)/E(Y)$, where Y is an auxiliary random variable that takes values in $\mathbb{N}^+$ with a probability mass function $pY(y)$. Together with another auxiliary random variable $\mathcal{K} \in \mathbb{N}^+$, which represents the component from which the binary random vector $\boldsymbol{X}$ is drawn, the probability mass function can be expressed as $P(\boldsymbol{X} = \boldsymbol{x}) = \sum_{y\geq 1} p(\boldsymbol{x}, y) = \sum_{y\geq 1}\sum_{k\geq 1} p(\boldsymbol{x}, y, k)$, where y and k are dummy variables for possible values of Y and $\mathcal{K}$, respectively. Here $p(\boldsymbol{x}, y, k) = p_Y(y)\mu^{-1}I(y \geq k)p_{\boldsymbol{x}}(\boldsymbol{\gamma}^{(k)})$ and $p(\boldsymbol{x}, y) = \sum_{k\geq 1} p(\boldsymbol{x}, y, k)$. Building upon the algorithm proposed by Hatjispyros et al. (2023), an adaptive Gibbs sampling approach can be derived, as outlined below. The assessment of the algorithm's performance is deferred to future work.

1. Initialize the latent variables $Y_l l \in [N]$ and $\mathcal{K} l \in [N]$ for observations $\boldsymbol{x}_l l \in [N]$.
2. Construct the decreasing weights via $w^{(k)} = \mathbb{P}(Y \geq k)/\mu$.
3. Update the locations of the mixture via $\pi(\boldsymbol{\gamma}^{(k)} \mid \cdots) \propto \pi(\boldsymbol{\gamma}^{(k)}) \prod_{l:\mathcal{K}_l=k} p_{\boldsymbol{x}_l}(\boldsymbol{\gamma}^{(k)})$. If there is no $\mathcal{K}_l = k$, then a sample for $\boldsymbol{\gamma}^{(k)}$ is taken from the prior $\pi(\boldsymbol{\gamma})$.
4. Sample the auxiliary variables from $\pi(Y_l = y_l \mid \mathcal{K}_l = k, \cdots) \propto p_Y(y_l)I(k \leq y_l)$ given the clustering variables.
5. Sample the discrete distribution for the clustering variables as $\pi(\mathcal{K}_l = k \mid Y_l = y_l, \cdots) \propto I(k \leq y_l)p_{\boldsymbol{x}_l}(\boldsymbol{\gamma}^{(k)})$.
6. Assuming $Y \sim \text{Poisson}(\lambda)$ with a prior $\lambda \sim \text{Gamma}(a, b)$, it follows that $\lambda \mid Y_l = y_l, \cdots \sim \text{Gamma}\left(\phi \mid a + \sum_{l=1}^{N} y_l - N, b + N\right)$.

References

Agresti, A. (2002). *Categorical Data Analysis, 2nd ed.* John Wiley & Sons, Hoboken.

Aickin, M. (1979). Existence of MLEs for discrete linear exponential models. *Annals of the Institute of Statistical Mathematics*, 31(1):103–113.

Allman, E. S., Matias, C., and Rhodes, J. A. (2009). Identifiability of parameters in latent structure models with many observed variables. *The Annals of Statistics*, 37(6A):3099–3132.

Anderson, M. J. (2001). A new method for non-parametric multivariate analysis of variance. *Austral Ecology*, 26(1):32–46.

Atienza, N., Garcia-Heras, J., Munoz-Pichardo, J., and Villa, R. (2007). On the consistency of MLE in finite mixture models of exponential families. *Journal of Statistical Planning and Inference*, 137(2):496–505.

Baisch, B., Cai, S., Li, Z., and Pinheiro, V. (2017). Reaction time of children with and without autistic spectrum disorders. *Open Journal of Medical Psychology*, 6:166–178.

Barndorff-Nielsen, O. E. (1978). *Information and Exponential Families in Statistical Theory*. Wiley.

Barrio, E. D., Giné, E., and Utzet, F. (2005). Asymptotics for L_2 functionals of the empirical quantile process, with applications to tests of fit based on weighted wasserstein distances. *Bernoulli*, 11(1):131–189.

Bartlett, M. S. (1935). Contingency table interactions. *Supplement to the Journal of the Royal Statistical Society*, 2(2):248–252.

Behboodian, J. (1970). On a mixture of normal distributions. *Biometrika*, 57(1):215–217.

Benachenhou, S., Etcheverry, A., Galarneau, L., Dubé, J., and Çaku, A. (2019). Implication of hypocholesterolemia in autism spectrum disorder and its associated comorbidities: A retrospective case–control study. *Autism Research*, 12(12):1860–1869.

Berry, K. J. and Mielke, P. W. (1983). Moment approximations as an alternative to the F test in analysis of variance. *British Journal of Mathematical and Statistical Psychology*, 36(2):202–206.

Bhattacharya, A. and Dunson, D. B. (2012). Simplex factor models for multivariate unordered categorical data. *Journal of the American Statistical Association*, 107(497):362–377.

Bi, Y. and Davuluri, R. V. (2013). NPEBseq: Nonparametric empirical bayesian-based procedure for differential expression analysis of RNA-seq data. *BMC Bioinformatics*, 14:262.

Billingsley, P. (1999). *Convergence of Probability Measures*. Wiley, 2nd edition.

Birch, M. W. (1963). Maximum likelihood in three-way contingency tables. *Journal of the Royal Statistical Society. Series B (Methodological)*, 25(1):220–233.

Bishop, Y. M. M., Fienberg, S. E., and Holland, P. W. (1975). *Discrete Multivariate Analysis Theory and Practice*. Springer.

Bobkov, S. and Ledoux, M. (2019). *One-dimensional empirical measures, order statistics, and Kantorovich transport distances*, volume 261. Mem. Amer. Math. Soc.

Böhning, D. (1982). Convergence of Simar's algorithm for finding the maximum likelihood estimate of a compound Poisson process. *Annals of Statistics*, 10(3):1006–1008.

Böhning, D. (1985). Numerical estimation of a probability measure. *Journal of Statistical Planning and Inference*, 11(1):57–69.

Böhning, D. (1986). A vertex-exchange-method in D-optimal design theory. *Metrika*, 33:337–347.

Boik, R. J. (1987). The Fisher-Pitman permutation test: A non-robust alternative to the normal theory F test when variances are heterogeneous. *British Journal of Mathematical and Statistical Psychology*, 40(1):26–42.

Borell, C. (1975). Convex set functions in d-space. *Period Math Hung*, 6:111–136.

Brascamp, H. J. and Lieb, E. H. (1976). On extensions of the Brunn-Minkowski and Prékopa-Leindler theorems, including inequalities for log concave functions, and with an application to the diffusion equation. *Journal of Functional Analysis*, 22(4):366 – 389.

Bricker, J., Miao, Z., Mull, K., Santiago-Torres, M., and Vock, D. M. (2023). Can a single variable predict early dropout from digital health interventions? comparison of predictive models from two large randomized trials. *J Med Internet Res*, 25:e43629.

Bricker, J. B., Mull, K. E., Santiago-Torres, M., Miao, Z., Perski, O., and Di, C. (2022). Smoking cessation smartphone app use over time: Predicting 12-month cessation outcomes in a 2-arm randomized trial. *Journal of Medical Internet Research*, 24(8).

Bro, R. (1997). PARAFAC. Tutorial and applications. *Chemometrics and intelligent laboratory systems*, 38(2):149–171.

Brooks, S. and King, R. (2001). Prior induction in log-linear models for general contingency table analysis. *The Annals of Statistics*, 29(3):715–747.

Brunel, V.-E. (2019). Learning rates for Gaussian mixtures under group action. In *Conference on Learning Theory*, pages 471–491. PMLR.

Calarge, C. A. and Schlechte, J. A. (2017). Bone mass in boys with autism spectrum disorder. *Journal of Autism and Developmental Disorders*, 47(6):1749–1755.

Cameron, J. M., Levandovskiy, V., Roberts, W., Anagnostou, E., Scherer, S., Loh, A., and Schulze, A. (2017). Variability of creatine metabolism genes in children with autism spectrum disorder. *International Journal of Molecular Sciences*, 18(8):1665.

Carreira-Perpinán, M. A. and Renals, S. (2000). Practical identifiability of finite mixtures of multivariate Bernoulli distributions. *Neural Computation*, 12(1):141–152.

Catchpole, E. A. and Morgan, B. J. T. (1997). Detecting parameter redundancy. *Biometrika*, 84(1):187–196.

Chandra, S. (1977). On the mixtures of probability distributions. *Scandinavian Journal of Statistics*, 4(3):105–112.

Chapuy, G. (2007). Random permutations and their discrepancy process. *2007 Conference on Analysis of Algorithms, AofA 07*, pages 457–470.

Chawarska, K., Campbell, D., Chen, L., Shic, F., Klin, A., and Chang, J. (2011). Early generalized overgrowth in boys with autism. *Archives of General Psychiatry*, 68(10):1021–1031.

Chen, G., Ning, B., and Shi, T. (2019). Single-cell RNA-seq technologies and related computational data analysis. *Frontiers in Genetics*, 10:317.

Chen, J. (1995). Optimal rate of convergence for finite mixture models. *Annals of Statistics*, 23(1):221–233.

Chen, J. (2017). Consistency of the mle under mixture models. *Statist. Sci.*, 32(1):47–63.

Cheng, R. and Liu, W. (2001). The consistency of estimators in finite mixture models. *Scandinavian Journal of Statistics*, 28(4):603–616.

Christensen, R. (1997). *Log-Linear Models and Logistic Regression*. Springer.

Chung, E. and Romano, J. P. (2013). Exact and asymptotically robust permutation tests. *Annals of Statistics*, 41(2):484–507.

Cochran, W. G. (1952). The χ^2 Test of Goodness of Fit. *The Annals of Mathematical Statistics*, 23(3):315–345.

Cover, T. M. and Thomas, J. A. (2006). *Elements of Information Theory*. Wiley-Interscience, 2nd edition.

Crawford, S. L. (1994). An application of the Laplace method to finite mixture distributions. *Journal of the American Statistical Association*, 89(425):259–267.

Crovella, M. E., Taqqu, M. S., and Bestavros, A. (1998). *Heavy-Tailed Probability Distributions in the World Wide Web*, page 3–25. Birkhauser Boston Inc., USA.

Csörgő, M. and Révész, P. (1978). Strong approximations of the quantile process. *Ann. Statist.*, 6(4):882–894.

Dadaneh, S. Z., Qian, X., and Zhou, M. (2018). BNP-seq: Bayesian nonparametric differential expression analysis of sequencing count data. *Journal of the American Statistical Association*, 113(521):81–94.

Darroch, J. N. (1962). Interactions in multi-factor contingency tables. *Journal of the Royal Statistical Society. Series B (Methodological)*, 24(1):251–263.

Dawid, A. P. and Lauritzen, S. L. (1993). Hyper-Markov laws in the statistical analysis of decomposable graphical models. *The Annals of Statistics*, 21(3):1272–1317.

De Angelis, R., Capocaccia, R., Hakulinen, T., Soderman, B., and Verdecchia, A. (1999). Mixture models for cancer survival analysis: application to population-based data with covariates. *Statistics in medicine*, 18(4):441–454.

De Lathauwer, L., De Moor, B., and Vandewalle, J. (2000). A multilinear singular value decomposition. *SIAM Journal on Matrix Analysis and Applications*, 21(4):1253–1278.

Deb, N. and Sen, B. (2021). Multivariate rank-based distribution-free nonparametric testing using measure transportation. *Journal of the American Statistical Association*, (in press).

Dellaportas, P. and Forster, J. J. (1999). Markov chain Monte Carlo model determination for hierarchical and graphical log-linear models. *Biometrika*, 86(3):615–633.

Dharmadhikari, S. and Joag-Dev, K. (1988). *Unimodality, convexity, and applications*. Academic Press.

Dobra, A. and Lenkoski, A. (2011). Copula Gaussian graphical models and their application to modeling functional disability data. *Annals of Applied Statistics*, 5:969–993.

Dobra, A. and Massam, H. (2010). The mode oriented stochastic search (MOSS) algorithm for log-linear models with conjugate priors. *Statistical Methodology*, 7(3):240–253. SPECIAL ISSUE ON STATISTICAL METHODS FOR THE SOCIAL SCIENCES Honoring the 10th Anniversary of the Center for Statistics and the Social Sciences at the University of Washington.

Dobra, A., Tebaldi, C., and West, M. (2006). Data augmentation in multi-way contingency tables with fixed marginal totals. *Journal of Statistical Planning and Inference*, 136:355–372.

Dümbgen, L., Kolesnyk, P., and Wilke, R. A. (2017). Bi-log-concave distribution functions. *Journal of Statistical Planning and Inference*, 184:1–17.

Dümbgen, L. and Wellner, J. A. (2014). Confidence bands for distribution functions: A new look at the law of the iterated logarithm. Technical report, Department of Statistics, University of Washington.

Durrett, R. (2019). *Probability—Theory and Examples*, volume 49 of *Cambridge Series in Statistical and Probabilistic Mathematics*. Cambridge University Press, Cambridge. Fifth edition of [MR1068527].

Erdős, P. and Stone, A. (1970). On the sum of two Borel sets. *Proceedings of the American Mathematical Society*, 25(2):304–306.

Erosheva, E. A., Fienberg, S. E., and Joutard, C. (2007). Describing disability through individual-level mixture models for multivariate binary data. *The Annals of Applied Statistics*, 1(2):346–384.

Ezegwui, I., Lawrence, L., Aghaji, A., Obiekwe, O., Okoye, O., Onwasigwe, E., and Ebigbo, P. (2014). Refractive errors in children with autism in a developing country. *Nigerian Journal of Clinical Practice*, 17:467–70.

Fedorov, V. (1972). *Theory of Optimal Experiments Designs*. Academic Press.

Feller, W. (1943). On a general class of “contagious” distributions. *The Annals of Mathematical Statistics*, 14(4):389 – 400.

Fienberg, S. E. (2000). Contingency tables and log-linear models: Basic results and new developments. *Journal of the American Statistical Association*, 95(450):643–647.

Fienberg, S. E. and Rinaldo, A. (2012). Maximum likelihood estimation in log-linear models. *The Annals of Statistics*, 40(2):996–1023.

Fisher, R. A. (1925). *Statistical Methods for Research Workers*. Oliver and Boyd.

Fisher, R. A. (1935). *Design of Experiments*. Oliver and Boyd.

Frühwirth-Schnatter, S. (2006). *Finite mixture and Markov switching models*. Springer, Berlin, 1st edition.

Fukumoto, A., Hashimoto, T., Mori, K., Tsuda, Y., Arisawa, K., and Kagami, S. (2011). Head circumference and body growth in autism spectrum disorders. *Brain and Development*, 33(7):569–75.

Gamerman, D. and Lopes, H. F. (2006). *Markov chain Monte Carlo: stochastic simulation for Bayesian inference*. CRC press.

Gardner, R. J. (2002). The Brunn-Minkowski inequality. *Bull. Amer. Math. Soc. (N.S.)*, 39(3):355–405.

Good, I. J. (1963). Maximum entropy for hypothesis formulation, especially for multidimensional contingency tables. *The Annals of Mathematical Statistics*, 34(3):911–934.

Goodman, L. A. (1963). On methods for comparing contingency tables. *Journal of the Royal Statistical Society. Series A (General)*, 126(1):94–108.

Goodman, L. A. (1964). Simple methods for analyzing three-factor interaction in contingency tables. *Journal of the American Statistical Association*, 59(306):319–352.

Grenander, U. (1956). On the theory of mortality measurement. II. *Skand. Aktuarietidskr.*, 39:125–153 (1957).

Gupta, S. and Chintagunta, P. K. (1994). On using demographic variables to determine segment membership in logit mixture models. *Journal of Marketing Research*, 31(1):128–136.

Haberman, S. (1974). *The Analysis of Frequency Data*. The University of Chicago Press.

Hall, P. (1984). On unimodality and rates of convergence for stable laws. *J. London Math. Soc. (2)*, 30(2):371–384.

Han, F., Miao, Z., and Shen, Y. (2021). Nonparametric mixture mles under gaussian-smoothed optimal transport distance. *arXiv preprint arXiv:2112.02421*.

Han, Q. and Wellner, J. A. (2016). Approximation and estimation of s-concave densities via Rényi divergences. *The Annals of Statistics*, 44(3):1332 – 1359.

Han, Y. and Shiragur, K. (2021). On the competitive analysis and high accuracy optimality of profile maximum likelihood. In *Proceedings of the 2021 ACM-SIAM Symposium on Discrete Algorithms (SODA)*, pages 1317–1336.

Haque, A., Engel, J., Teichmann, S. A., and Lönnberg, T. (2017). A practical guide to single-cell rna-sequencing for biomedical research and clinical applications. *Genome Medicine*, 9(1):75.

Hatjispyros, S. J., Merkatas, C., and Walker, S. G. (2023). Mixture models with decreasing weights. *Computational Statistics & Data Analysis*, 179:107651.

Hengartner, N. W. (1997). Adaptive demixing in poisson mixture models. *Annals of Statistics*, 25(3):917–928.

Hilbert, M. (2016). Big data for development: A review of promises and challenges. *Development Policy Review*, 34(1):135–174.

Hoeffding, W. (1952). The large-sample power of tests based on permutations of observations. *Annals of Mathematical Statistics*, pages 169–192.

Hohn, V. D., de Veld, D., Mataw, K., van Someren, E., and Begeer, S. (2019). Insomnia severity in adults with autism spectrum disorder is associated with sensory hyper-reactivity and social skill impairment. *Journal of Autism and Developmental Disorders*, 49(5):2146–2155.

Hoirisch-Clapauch, S. and Nardi, A. (2019). Autism spectrum disorders: Let's talk about glucose? *Translational Psychiatry*, 9(1):51.

Højsgaard, S., Edwards, D., and Lauritzen, S. (2012). *Graphical Models with R*. Springer, New York. ISBN 978-1-4614-2298-3.

Hoorfar, A. and Hassani, M. (2008). Inequalities on the lambert W function and hyperpower function. *Journal of Inequalities in Pure and Applied Mathematics*, 9(2).

Hosseinzadeh, D. and Krishnan, S. (2008). Gaussian mixture modeling of keystroke patterns for biometric applications. *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)*, 38(6):816–826.

Huang, M., Wang, J., Torre, E., Dueck, H., Shaffer, S., Bonasio, R., Murray, J., Raj, A., Li, M., and Zhang, N. R. (2018). SAVER: Gene expression recovery for UMI-based single cell RNA sequencing. *bioRxiv*.

Ishwaran, H. and Rao, J. S. (2005). Spike and slab variable selection: Frequentist and Bayesian strategies. *The Annals of Statistics*, 33(2):730 – 773.

Jewell, N. P. (1982). Mixtures of exponential distributions. *Annals of Statistics*, 10(2):479–484.

Jiang, W. and Zhang, C.-H. (2019). Rate of divergence of the nonparametric likelihood ratio test for gaussian mixtures. *Bernoulli*, 25(4B):3400–3420.

Jiao, J., Han, Y., and Weissman, T. (2018). Minimax estimation of the ℓ_1 distance. *IEEE Transactions on Information Theory*, 64(10):6672–6706.

Jiao, J., Venkat, K., Han, Y., and Weissman, T. (2015). Minimax estimation of functionals of discrete distributions. *IEEE Transactions on Information Theory*, 61(5):2835–2885.

Johndrow, J. E., Bhattacharya, A., and Dunson, D. B. (2017). Tensor decompositions and sparse log-linear models. *The Annals of Statistics*, 45(1):1–38.

Joshi, G., Biederman, J., Petty, C., Goldin, R. L., Furtak, S. L., and Wozniak, J. (2012). Examining the comorbidity of bipolar disorder and autism spectrum disorders: A large controlled analysis of phenotypic and familial correlates in a referred population of youth with bipolar I disorder with and without autism spectrum disorders. *Journal of Clinical Psychiatry*.

Juan, A. and Vidal, E. (2002). On the use of Bernoulli mixture models for text classification. *Pattern Recognition*, 35(12):2705–2710.

Juan, A. and Vidal, E. (2004). Bernoulli mixture models for binary images. In *Proceedings of the 17th International Conference on Pattern Recognition, 2004. ICPR 2004.*, volume 3, pages 367–370. IEEE.

Kiefer, J. and Wolfowitz, J. (1956). Consistency of the maximum likelihood estimator in the presence of infinitely many incidental parameters. *Ann. Math. Statist.*, 27(4):887–906.

Kim, A. K. (2014). Minimax bounds for estimation of normal mixtures. *Bernoulli*, 20(4):1802–1818.

Kim, B.-L. C. (1977). Asian wives of U.S. servicemen: Women in shadows. *Amerasia Journal*, 4(1):91–115.

Kindermann, R. and Snell, J. L. (1980). *Markov Random Fields and their Applications*, volume 1. American Mathematical Society Providence, RI.

Kleiber, C. and Kotz, S. (2003). *Statistical Size Distributions in Economics and Actuarial Sciences*. John Wiley & Sons.

Knuiman, M. W. and Speed, T. P. (1988). Incorporating prior information into the analysis of contingency tables. *Biometrics*, 44(4):1061–1071.

Kriston, L. (2013). Dealing with clinical heterogeneity in meta-analysis. Assumptions, methods, interpretation. *International journal of methods in psychiatric research*, 22(1):1–15.

Laha, N., Miao, Z., and Wellner, J. A. (2021). Bi-s^*-concave distributions. *Journal of Statistical Planning and Inference*, 215:127–157.

Laird, N. (1978). Nonparametric maximum likelihood estimation of a mixing distribution. *Journal of the American Statistical Association*, 73(364):805–811.

Lambert, D. and Tierney, L. (1984). Asymptotic properties of maximum likelihood estimates in the mixed Poisson model. *Annals of Statistics*, 12(4):1388–1399.

Lauritzen, S. (1996). *Graphical Models*. Oxford University Press, Oxford, UK.

Leisch, F. (2004). Exploring the structure of mixture model components. *Proceedings in Computational Statistics*.

Lesperance, M. L. and Kalbfleisch, J. D. (1992). An algorithm for computing the nonparametric mle of a mixing distribution. *Journal of the American Statistical Association*, 87(417):120–126.

Letac, G. and Massam, H. (2012). Bayes factors and the geometry of discrete hierarchical log-linear models. *The Annals of Statistics*, 40(2):861–890.

Li, H. and Reynolds, J. F. (1995). On definition and quantification of heterogeneity. *Oikos*, 73(2):280–284.

Lindsay, B. G. (1983a). The geometry of mixture likelihoods: A general theory. *Annals of Statistics*, 11(1):86–94.

Lindsay, B. G. (1983b). The geometry of mixture likelihoods, part ii: The exponential family. *Annals of Statistics*, 11(3):783–792.

Lindsay, B. G. (1995). Mixture models: Theory, geometry and applications. *NSF-CBMS Regional Conference Series in Probability and Statistics*, 5:I–163.

Lindsay, B. G. and Roeder, K. (1993). Uniqueness of estimation and identifiability in mixture models. *The Canadian Journal of Statistics / La Revue Canadienne de Statistique*, 21(2):139–147.

Liu, S., Jiang, Y., and Yu, T. (2019). Modelling RNA-Seq data with a zero-inflated mixture Poisson linear model. *Genetic Epidemiology*, 43(7):786–799.

Loh, W.-L. and Zhang, C.-H. (1996). Global properties of kernel estimators for mixing densities in discrete exponential family models. *Statistica Sinica*, 6(3):561–578.

Loh, W.-L. and Zhang, C.-H. (1997). Estimating mixing densities in exponential family models for discrete variables. *Scandinavian Journal of Statistics*, 24(1):15–32.

Love, M., Huber, W., and Anders, S. (2014). Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biology*, 15:550.

Lu, M. (2018). *Generalized Adaptive Shrinkage Methods and Applications in Genomics Studies*. University of Chicago.

Lubke, G. H. and Muthén, B. (2005). Investigating population heterogeneity with factor mixture models. *Psychological methods*, 10(1):21.

Mannering, F. L., Shankar, V., and Bhat, C. R. (2016). Unobserved heterogeneity and the statistical analysis of highway accident data. *Analytic Methods in Accident Research*, 11:1–16.

Manole, T. and Khalili, A. (2021). Estimating the number of components in finite mixture models via the Group-Sort-Fuse procedure. *The Annals of Statistics*, 49(6):3043 – 3069.

Manton, K. G., Corder, L., and Stallard, E. (1993). Estimates of change in chronic disability and institutional incidence and prevalence rate in the us elderly populations from 1982 to 1989. *Journal of Gerontology Social Sciences*, 48:S153–S166.

Marascuilo, L. A. and McSweeney, M. (1977). *Nonparametric and Distribution-Free Methods for the Social Sciences*. Brooks/Cole Publishing Company.

Mardia, J., Jiao, J., Tànczos, E., Nowak, R. D., and Weissman, T. (2019). Concentration inequalities for the empirical distribution of discrete distributions: beyond the method of types. *Information and Inference: A Journal of the IMA*.

Massam, H., Liu, J., and Dobra, A. (2009). A conjugate prior for discrete hierarchical log-linear models. *The Annals of Statistics*, 37(6A):343–3467.

Massart, P. (1990). The tight constant in the Dvoretzky-Kiefer-Wolfowitz inequality. *Ann. Probab.*, 18(3):1269–1283.

McLachlan, G. J. and Basford, K. E. (1988). *Mixture Models: Inference and Application to Clustering*.

Miao, Z., Chen, Y.-C., and Dobra, A. (2023). Bayesian finite mixtures of ising models.

Miao, Z., Kong, W., Vinayak, R. K., Sun, W., and Han, F. (2022). Fisher-pitman permutation tests based on nonparametric poisson mixtures with application to single cell genomics. *Journal of the American Statistical Association*, 0(0):1–13.

Mielke, P. W. and Berry, K. J. (2007). *Permutation Methods: A Distance Function Approach*. Springer.

Mielke Jr, P. (1984). 34 Meteorological applications of permutation techniques based on distance functions. In *Handbook of Statistics*, volume 4, pages 813–830. Elsevier.

Mielke Jr, P. W., Berry, K. J., and Johnson, E. S. (1976). Multi-response permutation procedures for a priori classifications. *Communications in Statistics-Theory and Methods*, 5(14):1409–1424.

Nesterov, Y. (2003). *Introductory lectures on convex optimization: A basic course*. Springer.

Nguyen, X. (2013). Convergence of latent mixing measures in finite and infinite mixture models. *Annals of Statistics*, 41(1):370–400.

Olkin, I. and Rubin, H. (1964). Multivariate beta distributions and independence properties of the Wishart distribution. *The Annals of Mathematical Statistics*, 35(1):261 – 269.

Owen, A. B. (1995). Nonparametric likelihood confidence bands for a distribution function. *Journal of the American Statistical Association*, 90(430):516–521.

Panaretos, V. M. and Zemel, Y. (2019). Statistical aspects of Wasserstein distances. *Annual Review of Statistics and Its Application*, 6(1):405–431.

Papathomas, M. and Richardson, S. (2016). Exploring dependence between categorical variables: Benefits and limitations of using variable selection within bayesian clustering in relation to log-linear modelling with interaction terms. *Journal of Statistical Planning and Inference*, 173:47–63.

Pearson, K. (1894). Contributions to the mathematical theory of evolution. *Philosophical Transactions of the Royal Society of London A*, 185:71–110.

Petersen, A. and Müller, H.-G. (2019). Fréchet regression for random objects with Euclidean predictors. *Annals of Statistics*, 47(2):691–719.

Pfanzagl, J. (1988). Consistency of maximum likelihood estimators for certain nonparametric families, in particular: Mixtures. *Journal of Statistical Planning and Inference*, 19(2):137–158.

Piper, N. and Roces, M. (2004). *Wife or worker?: Asian women and migration*. Rowman & Littlefield Publishers.

Pitman, E. J. G. (1938). Significance tests which may be applied to samples from any populations III. The analysis of variance test. *Biometrika*, 29(3/4):322–335.

Polyanskiy, Y. and Wu, Y. (2020). Self-regularizing property of nonparametric maximum likelihood estimator in mixture models. *arXiv preprint arXiv:2008.08244*.

Rau, A., Maugis-Rabusseau, C., Martin-Magniette, M.-L., and Celeux, G. (2015). Co-expression analysis of high-throughput transcriptome sequencing data with Poisson mixture models. *Bioinformatics*, 31(9):1420–1427.

Ravikumar, P., Wainwright, M. J., and Lafferty, J. D. (2010). High-dimensional Ising model selection using ℓ_1-regularized logistic regression. *The Annals of Statistics*, 38(3):1287–1319.

Ray, S. and Lindsay, B. G. (2005). The topography of multivariate normal mixtures. *The Annals of Statistics*, 33(5):2042 – 2065.

Rebafka, T. and Roueff, F. (2015). Nonparametric estimation of the mixing density using polynomials. *Math. Meth. Stat.*, 24:200–224.

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

Rinott, Y. (1976). On convexity of measures. *Ann. Probab.*, 4(6):1020–1026.

Robertson, C. A. and Fryer, J. G. (1969). Some descriptive properties of normal mixtures. *Scandinavian Actuarial Journal*, 1969(3-4):137–146.

Robertson, T., Wright, F. T., and Dykstra, R. L. (1988). *Order Restricted Statistical Inference*. Wiley & Sons.

Robinson, J. (1973). The large-sample power of permutation tests for randomization models. *Annals of Statistics*, 1(2):291–296.

Robinson, M. D., McCarthy, D. J., and Smyth, G. K. (2010). edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics*, 26(1):139–140.

Rockafellar, R. T. (1970). *Convex Analysis*. Princeton University Press.

Ročková, V. (2018). Particle EM for variable selection. *Journal of the American Statistical Association*, 113(524):1684–1697.

Ročková, V. and George, E. I. (2014). EMVS: The EM approach to Bayesian variable selection. *Journal of the American Statistical Association*, 109(506):828–846.

Ročková, V. and George, E. I. (2018). The Spike-and-Slab LASSO. *Journal of the American Statistical Association*, 113(521):431–444.

Rothenberg, T. J. (1971). Identification in parametric models. *Econometrica*, 39(3):577–591.

Roueff, F. and Rydén, T. (2005). Nonparametric estimation of mixing densities for discrete distributions. *Annals of Statistics*, 33(5):2066–2108.

Roy, S. N. and Kastenbaum, M. A. (1956). On the hypothesis of no "interaction" in a multi-way contingency table. *The Annals of Mathematical Statistics*, 27(3):749 – 757.

Samworth, R. J. (2018). Recent progress in log-concave density estimation. *Statist. Sci.*, 33(4):493–509.

Samworth, R. J. and Sen, B. (2018). Editorial: special issue on "Nonparametric inference under shape constraints". *Statist. Sci.*, 33(4):469–472.

Sarkar, A. and Stephens, M. (2021). Separating measurement and expression models clarifies confusion in single-cell rna sequencing analysis. *Nature Genetics*, 53(6):770–777.

Saumard, A. (2019). Bi-log-concavity: some properties and some remarks towards a multi-dimensional extension. *Electron. Commun. Probab.*, 24:Paper No. 61, 8.

Scheffé, H. (1959). *The Analysis of Variance*. Wiley.

Serfling, R. J. (1980). *Approximation Theorems of Mathematical Statistics*. Wiley.

Shaked, M. (1980). On mixtures from exponential families. *Journal of the Royal Statistical Society. Series B (Methodological)*, 42(2):192–198.

Shi, H., Drton, M., and Han, F. (2020). Distribution-free consistent independence tests via center-outward ranks and signs. *Journal of the American Statistical Association*, (in press).

Shorack, G. R. (2017). *Probability for Statisticians*. Springer.

Shorack, G. R. and Wellner, J. A. (2009). *Empirical Processes with Applications to Statistics*, volume 59 of *Classics in Applied Mathematics*. Society for Industrial and Applied Mathematics (SIAM), Philadelphia, PA. Reprint of the 1986 original [MR0838963].

Silva, A., Rothstein, S. J., McNicholas, P. D., and Subedi, S. (2019). A multivariate Poisson-log normal mixture model for clustering transcriptome sequencing data. *BMC Bioinformatics*, 20:394.

Simar, L. (1976). Maximum likelihood estimation of a compound poisson process. *Annals of Statistics*, 4(6):1200–1209.

Snipen, L., Almøy, T., and Ussery, D. W. (2009). Microbial comparative pan-genomics using binomial mixture models. *BMC genomics*, 10(1):1–8.

Still, A. and White, A. (1981). The approximate randomization test as an alternative to the F test in analysis of variance. *British Journal of Mathematical and Statistical Psychology*, 34(2):243–252.

Swami, A. (2000). Non-gaussian mixture models for detection and estimation in heavy-tailed noise. In *2000 IEEE International Conference on Acoustics, Speech, and Signal Processing. Proceedings (Cat. No. 00CH37100)*, volume 6, pages 3802–3805. IEEE.

Tan, X., Chen, J., and Zhang, R. (2007). Consistency of the constrained maximum likelihood estimator in finite normal mixture models. *Proceedings of the American Statistical Association, American Statistical Association, Alexandria, VA*, pages 2113–2119.

Teicher, H. (1960). On the mixture of distributions. *The Annals of Mathematical Statistics*, 31(1):55 – 73.

Teicher, H. (1961). Identifiability of mixtures. *Ann. Math. Statist.*, 32(1):244–248.

Teicher, H. (1963). Identifiability of finite mixtures. *The Annals of Mathematical Statistics*, 34(4):1265 – 1269.

Teicher, H. (1967). Identifiability of mixtures of product measures. *The Annals of Mathematical Statistics*, 38(4):1300–1302.

Tian, K., Kong, W., and Valiant, G. (2017). Learning populations of parameters. *arXiv preprint arXiv:1709.02707*.

Timan, A. F. (2014). *Theory of Approximation of Functions of a Real Variable*, volume 34. Elsevier.

Titterington, D. M., Smith, A. F. M., and Makov, U. E. (1985). *Statistical Analysis of Finite Mixture Distributions*. Wiley.

van Borkulo, C. D., Borsboom, D., Epskamp, S., Blanken, T. F., Boschloo, L., Schoevers, R. A., and Waldorp, L. J. (2014). A new method for constructing networks from binary data. *Scientific Reports*, 4(1):5918.

van de Geer, S. (1996). Rates of convergence for the maximum likelihood estimator in mixture models. *Journal of Nonparametric Statistics*, 6(4):293–310.

van de Geer, S. (2003). Asymptotic theory for maximum likelihood in nonparametric mixture models. *Computational Statistics and Data Analysis*, 41(3):453–464.

van Eeden, C. (1956). *Maximum likelihood estimation of ordered probabilities*. Statist. Afdeling S 188 (VP 5). Math. Centrum Amsterdam.

Velmeshev, D., Schirmer, L., Jung, D., Haeussler, M., Perez, Y., Mayer, S., Bhaduri, A., Goyal, N., Rowitch, D. H., and Kriegstein, A. R. (2019). Single-cell genomics identifies cell type–specific molecular changes in autism. *Science*, 364(6441):685–689.

Verbeek, A. (1992). The compactification of generalized linear models. *Statistica Neerlandica*, 46(2-3):107–142.

Vinayak, R. K., Kong, W., Valiant, G., and Kakade, S. M. (2019). Maximum likelihood estimation for learning populations of parameters. *arXiv preprint arXiv:1902.04553*.

Vu, T. N., Wills, Q. F., Kalari, K. R., Niu, N., Wang, L., Rantalainen, M., and Pawitan, Y. (2016). Beta-Poisson model for single-cell RNA-seq data analyses. *Bioinformatics*, 32(14):2128–2135.

Walther, G. (2009). Inference and modeling with log-concave distributions. *Statist. Sci.*, 24(3):319–327.

Wang, N., Rauh, J., and Massam, H. (2019). Approximating faces of marginal polytopes in discrete hierarchical models. *The Annals of Statistics*, 47(3):1203–1233.

Whittaker, J. (1990). *Graphical Models in Applied Multivariate Statistics*. Wiley.

Wooldridge, J. M. (2000). Ceosal2. Instructional Stata datasets for econometrics ceosal2, Boston College Department of Economics.

Wu, C.-F. (1978a). Some algorithmic aspects of the theory of optimal designs. *Annals of Statistics*, 6(6):1286–1301.

Wu, C.-F. (1978b). Some iterative procedures for generating nonsingular optimal designs. *Communications in Statistics-Theory and Methods*, 7(14):1399–1412.

Wu, F. Y. (1982). The Potts model. *Review of Modern Physics*, 54:235–268.

Wu, H., Qin, Z., and Zhu, Y. (2013). PM-Seq: Using finite Poisson mixture models for RNA-seq data analysis and transcript expression level quantification. *Statistics in Biosciences*, 5(1):71–87.

Wu, Y. and Yang, P. (2016). Minimax rates of entropy estimation on large alphabets via best polynomial approximation. *IEEE Transactions on Information Theory*, 62(6):3702–3720.

Wu, Y. and Yang, P. (2020a). Optimal estimation of gaussian mixtures via denoised method of moments. *Annals of Statistics*, 48(4):1981–2007.

Wu, Y. and Yang, P. (2020b). Polynomial methods in statistical inference: Theory and practice. *Foundations and Trends in Communications and Information Theory*, 17(4):402–586.

Xu, G. (2017). Identifiability of restricted latent class models with binary responses. *The Annals of Statistics*, 45(2):675 – 707.

Yakowitz, S. J. and Spragins, J. D. (1968). On the identifiability of finite mixtures. *The Annals of Mathematical Statistics*, 39(1):209–214.

Yu, G., Sapiro, G., and Mallat, S. (2011). Solving inverse problems with piecewise linear estimators: From gaussian mixture models to structured sparsity. *IEEE Transactions on Image Processing*, 21(5):2481–2499.

Zhang, C.-H. (1995). On estimating mixing densities in discrete exponential family models. *Annals of Statistics*, 23(3):929–945.

Zhang, M., Liu, S., Miao, Z., Han, F., Gottardo, R., and Sun, W. (2022). Ideas: individual level differential expression analysis for single-cell rna-seq data. *Genome Biology*, 23(33).

Zhang, M., Liu, S., Miao, Z., Han, F., Gottardo, R., and Sun, W. (2022). Individual level differential expression analysis for single cell RNA-seq data. *Genome Biology*, 23(33):1–11.

Zhang, M. J., Ntranos, V., and Tse, D. (2020). Determining sequencing depth in a single-cell RNA-seq experiment. *Nature Communications*, 11:774.

Zhou, J., Bhattacharya, A., Herring, A. H., and Dunson, D. B. (2015). Bayesian factorizations of big sparse tensors. *Journal of the American Statistical Association*, 110(512):1562–1576.

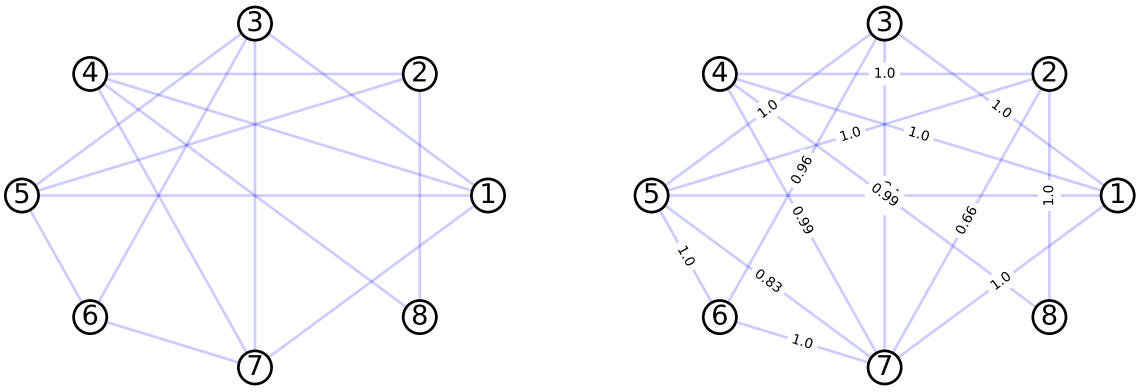

Figure 10: Significant pairwise associations determined by Whittaker (1990) (left panel) and the Bayesian Ising model we proposed (right panel). Each association is shown as an edge between vertices associated with variables it involves. The labels of the edges indicate the estimated posterior means of their indicators.

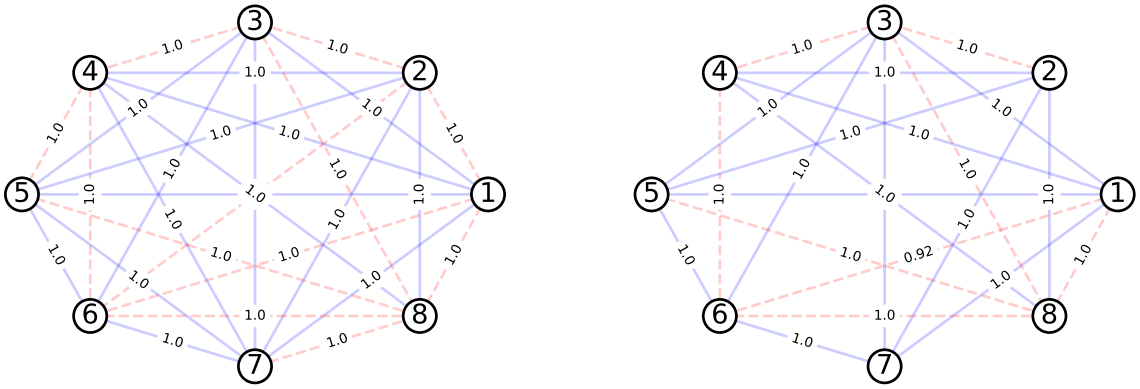

Figure 11: Significant pairwise associations determined by the Bayesian Ising mixture model. The 28 associations in the first component are presented in the left panel, while the 22 associations in the right component are shown in the right panel. Each association is shown as an edge between vertices associated with variables it involves. The labels of the edges indicate the estimated posterior means of their indicators.

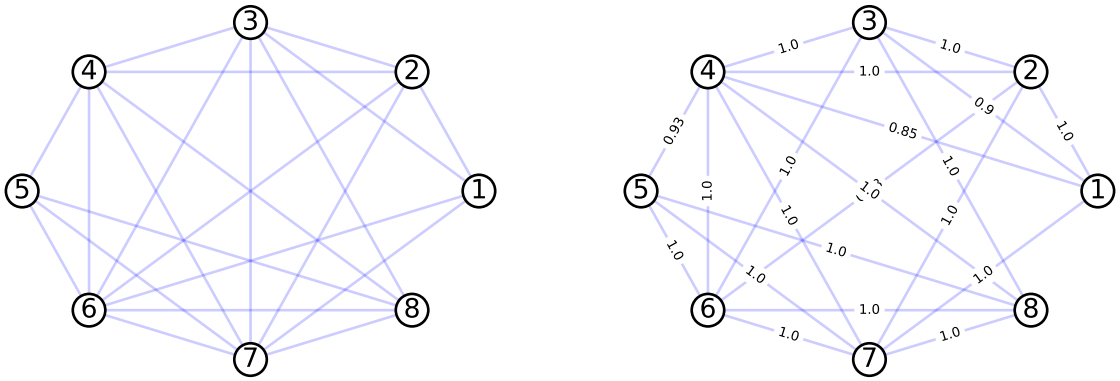

Figure 12: Significant pairwise associations determined by the forward stepwise function based on AIC in the R package gRim (Højsgaard et al., 2012) (left panel) and the Bayesian Ising model (right panel) in the NLTCS data. Pairwise associations are shown as edges between vertices associated with variables they involve. The labels of the edges indicate the estimated posterior means of their indicators.

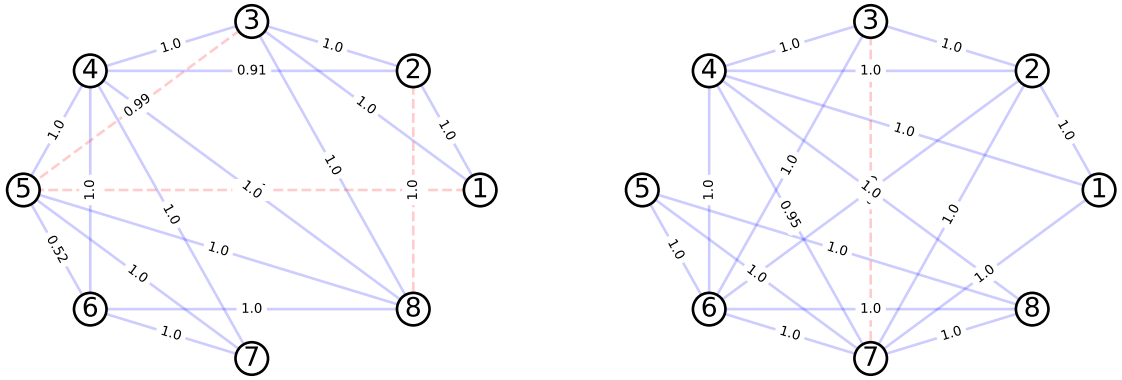

Figure 13: Significant pairwise associations determined by the Bayesian Ising mixture model with two components in the NLTCS data. Each association is shown as an edge between vertices associated with variables it involves. There are 18 associations in the first component (shown in the left panel), and 19 associations in the second component (shown in the right panel). The labels of the edges indicate the estimated posterior means of their indicators.

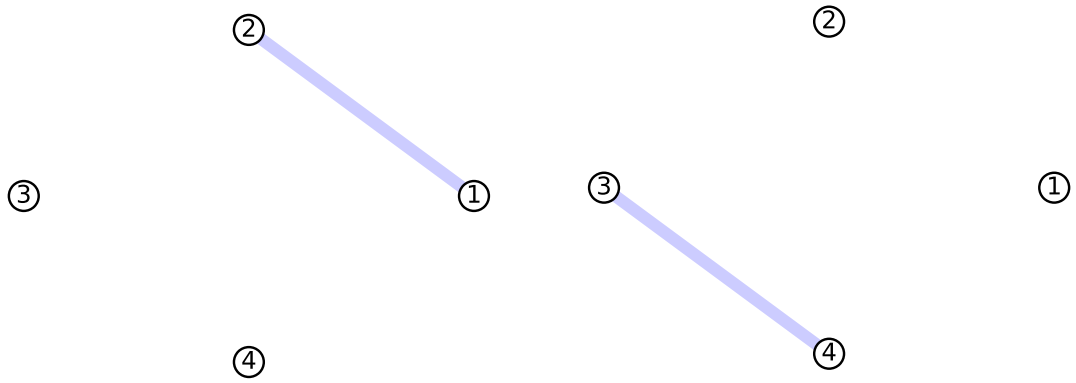

Figure 14: Graph representations for Example 4.3. Component 1 is shown on the left and component 2 is shown on the right.