# Improved Tools for Large-Scale Hypothesis Testing

by

**Zihao Zheng**

A dissertation submitted in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
(Statistics)
at the
University of Wisconsin-Madison
2022

Date of Final Oral Exam: May/10/2022

The dissertation is approved by the following members of the Final Committee:

Michael A Newton, Professor, Statistics, Biostatistics and Medical Informatics
Sunduz Keles, Professor, Statistics, Biostatistics and Medical Informatics
Karl Broman, Professor, Biostatistics and Medical Informatics
Chunming Zhang, Professor, Statistics
Miriam A Shelef, Associate Professor, Medicine

# Improved Tools for Large-Scale Hypothesis Testing

Zihao Zheng

## Abstract

Large-scale hypothesis testing, as one of the key statistical tools, has been widely studied and applied to high throughput bioinformatics experiments, such as high density peptide array studies and brain image data sets. The high dimensionality and small sample size of many experiments challenge conventional statistical approaches, including those aiming to control the false discovery rate (FDR). Motivated by this, in this dissertation, I develop several improved statistical and computational tools for large-scale hypothesis testing. The first method, `MixTwice`, advances an empirical-Bayesian tool that computes local false discovery rate statistics when provided with data on estimated effects and estimated standard errors. I also extend this method from two group comparison problems to multiple group comparison settings and develop a generalized method called `MixTwice-ANOVA`. The second method `GraphicalT` calculates local FDRs semiparametrically using available graph-associated information.

The first method, called `MixTwice`, introduces an empirical-Bayes approach that involves the estimation of two mixing distributions, one on underlying effects and one on underlying variance parameters. Provided with the estimated effect sizes and estimated errors, `MixTwice` estimates the mixing distribution and calculates the local false discovery rates via nonparametric MLE and constrained optimization with unimodal shape constraint of the effect distribution. Numerical experiments show that `MixTwice` can accurately estimate generative parameters and have good testing operating characteristics. Applied on a high density peptide array, it powerfully identifies non-null peptides to recover meaningful peptide markers when the underlying signal is weak, and has strong reproducibility properties when the underlying signal is strong.

The second contribution of this dissertation generalizes `MixTwice` from scenarios comparing two conditions to scenarios comparing multiple groups. Similar to `MixTwice`, `MixTwice-ANOVA` takes numerator and denominator statistics of F test to estimate two underlying mixing distributions. Compared with other large-scale testing tools for one-way ANOVA settings, `MixTwice-ANOVA` has better power prop-

erties and FDR control through numerical experiments. Applied to the peptide array study comparing multiple Sjogren-disease (SjD) populations, the proposed approach discovers meaningful epitope structure and novel scientific findings on Sjogren disease.

Numerical experiments support evaluation among testing tools. Besides the methodology contribution of `MixTwice` in large-scale testing, I also discuss generalized evaluation and computational aspects. For the former part, I propose an evaluation metric, in additional to FDR control, power, etc., called reproducibility, to provide a practical guide for different testing tools. For the latter part, I borrow the idea from pool adjacent violator algorithm (PAVA) and advance a computational algorithm called `EM-PAVA` to solve nonparametric MLE with isotonic partial order constraint. This algorithm is discussed through theoretical guarantees and computational performances.

The last contribution of this dissertation deals with large-scale testing problems with graph-associated data. Different from many studies that incorporate the graph-associated information through detailed modeling specifications, `GraphicalT` provides a semiparametric way to calculate the local false discovery rates using available auxiliary data graph. The method shows good performance in synthetic examples and in a brain-imaging problem from the study of Alzheimer's disease.

To mom and dad

# Acknowledgements

This work would not be possible without support from my supervisor, collaborators, colleagues, family and friends. I would like to say thank you to all of them for their generous help during my Ph.D. study.

Firstly, I would like to say thank you to my supervisor, Professor Michael A. Newton. Professor Newton is a fantastic statistician with many remarkable contributions in a variety of domains, including statistical theory, methodology, computation and application. He introduced me to the fantastic area of large-scale hypothesis testing and guided me to struggle with complicated data problems. I have been learning a lot during every single meeting with him, about statistics, research, writing, career, and life.

Second, I would like to thank Professor Miriam A. Shelef, my research assistant advisor. Professor Shelef is a rheumatologist and a faculty member of the School of Medicine and Public Health, with many terrific contributions in immunology diseases. She introduced me to the application research of high density peptide array, which motivated the majority of my statistical innovation during my Ph.D. study. Without her help, it would have been impossible for me to understand how statistical tools could be applied to real world challenges.

I would also like to thank my other final committee members: Professors Karl Broman, Chunming Zhang and Sunduz Keles. It is my great honor to have them on my thesis committee and I really appreciate their time in reading my dissertation. Their critical questions during the preliminary exam really guided the development of my dissertation.

I am also fortunate to work with faculty collaborators and other colleagues: Professor Irene Ong, Professor Sara McCoy, Dr.Aisha Mergaert, Dr.Xiuyu Ma, Tun Lee Ng, Peng Yu, Janna Bashar, Maya Amjadi, Maxwell Parker, Susan Glenn, Ilya Gurevic, Srishti Gupta, Ryan Adyniec. I appreciate their generous help during my research and life.

I would also like to express my appreciation to my fiancée Yutian Liu. It is my greatest fortune to meet you in high school, and thank you for your warm support in my life.

Finally, my greatest and deepest gratitude needs to be offered to my parents, Yongming Zheng and Wei Shi. Even though we are separated by oceans of almost 10,000 miles, your encouragement is always my biggest support.

# Contents

# List of Figures

# List of Tables

# Chapter 1

# Introduction

Large-scale hypothesis-testing tools deal with statistical inference problems when multiple hypotheses are simultaneously tested. These tools are needed and used in a wide variety of subject-matter domains. For example, in transcript analysis investigators might want to assess changes in gene expression between different cellular conditions [e.g., Van den Berge et al., 2017]. In brain imaging, investigators seek to understand changes in brain structure between different populations [e.g., Nichols and Hayasaka, 2003, Alberton et al., 2020]. Genetics researchers may be interested in differences in some phenotype between populations associated with some aspect of genetic structure [e.g., Dudoit et al., 2003]. Psychologists and empirical economists may be interested in understanding treatment effects of multiple arms or heterogeneity effects across different sub-populations [e.g., Shaffer, 1995, Bajgrowicz and Scaillet, 2012]. It is common, regardless of the domain, for the problem to be an unordered list of separate, uni-dimensional statistical testing questions. While a useful approach to such problems is to apply classical statistical testing procedures separately to each dimension, contemporary statistical theory tells us that overall operating characteristics can be improved by calculating decision rules for each test which rely on information extracted from the entire collection [Efron, 2012]. Broadly speaking, my thesis aims to contribute further to the statistical methodology for large-scale hypothesis testing.

In large-scale testing, the discovery list is a collection of hypotheses that are rejected based on a certain decision rule; i.e., these are the units inferred to be most statistically interesting in the comparison under consideration, and they often will be the subject of extensive follow-up experimentation and analysis. The false discoveries are units that are on the discovery list but that really ought not have

been placed there; they are incorrectly rejected (i.e., the type-I errors), and further, perhaps indefinite, experimentation and sampling would be futile in establishing significant differences for such units. Null statistical fluctuations cause these units to be placed on the discovery list. An important issue in constructing the list of discoveries is to control in some way the rate of these type-I errors. Efforts to control the type-I error rate in large-scale testing, such as Bonferroni correction [Bonferroni, 1936], started from methods that control the family-wise error rate (FWER), which is the probability of at least one false discovery. Its simplicity makes it popular for controlling FWER. However, such corrections are considered to be highly conservative, and may easily result in empty discovery lists, especially in cases with extremely high dimensionality and small sample size. Furthermore, the investigator may be comfortable sifting through lists that have more than one false discovery, as long as the rate of such is low, so aiming to control the FWER is not well-justified in some settings.

The false discovery rate (FDR), which is the expected proportion of false discoveries, was introduced to address the statistical control of type-I errors in large-scale testing [Benjamini and Hochberg, 1995]. Methods aiming to control the FWER also control the FDR, but the latter may be controlled with a less stringent decision rule. Methods proposed to control the FDR, such as Benjamini and Hochberg procedure (BH, [Benjamini and Hochberg, 1995]) and Storey's `qvalue` method [Storey, 2002], have been shown to have greater power to detect true positives. Beyond the question of the number of discoveries, it is important to recognize that the statistical methods to process large-scale data will affect the rank ordering of any list of discoveries. Relevant to my research contribution is the question of precisely what data are used from each testing unit in order to construct the required decision rules. The use of different data summaries affects both the rank-ordering of units and the assessment of false-discovery rate.

The empirical-Bayes mixture model has been applied in a variety of settings for large-scale hypothesis testing by treating all units as coming from a common population. In the context of an estimated mixture model, a useful empirical-Bayesian inference statistic to make the discovery list is local false discovery rate, sometimes denoted *lfdr* [Efron et al., 2001]. The local false discovery rate is the posterior probability of the null hypothesis given the statistics computed *locally* on each testing unit, where the required distributional forms are estimated *globally* from the full set of units.

One of the important questions in large-scale hypothesis testing concerns what

statistics are computed on each unit and then imported into the empirical-Bayes calculation. Methods such as `qvalue` work from unit-specific p-values that have been computed from two-sample t tests [Storey, 2002]. This method has improved power properties compared to the Bonferroni correction and BH procedure, but it still extracts a high penalty for dimensionality in many examples [Zheng et al., 2021]. One reason is because `qvalue` method enters quite late in data analysis; methods that reduce data less prior to empirical-Bayes mixing may be expected to have superior operating characteristics. For example, Efron's local FDR procedure (`locFDR`, [Efron et al., 2001]) intervenes on test statistics (or z-scores) prior to the p-value calculation, in order to avoid the reduction of sign information going from test statistics to p-values. Further improvements intervene on components of the local test statistics. Adaptive Shrinkage (`ASH`, [Stephens, 2017]) is a recent innovation that applies empirical-Bayesian modeling on estimated effect sizes and on estimated standard errors. `ASH` involves a mixture distribution for latent effect and further restricts the mixing distribution to be unimodal. Specifically, it encodes a nonparametric shape constraint that models the units with larger effects to be less common than units with smaller effects. Refinements of `ASH` include a two-step formulation [Lu and Stephens, 2019, 2016], addressing technical limitations of the standard-error modeling.

`ASH` and two-step `ASH` provide an effective general framework for utilizing two-dimensional effects and standard errors for large-scale testing. The methods have reasonable operating characteristics in many cases [e.g., Korthauer et al., 2019], especially when the amount of information (e.g. sample size) per testing unit is quite high, and thereby the estimated standard errors are close to the underlying standard errors and the parametric observation components are well supported. However, the `ASH` methods have deficiencies when the standard errors are poorly estimated. Motivated by this, I discuss and evaluate a novel empirical-Bayes method, named `MixTwice`, that intervenes after effect estimates and standard errors are computed on each testing unit (Chapter 2). By contrast with `ASH`, the proposed `MixTwice` method relies on separate nonparametric mixtures for variance and effect parameters. Combined with existing tools for nonparametric maximum likelihood estimation and constrained optimization, `MixTwice` is shown to have good operating characteristics through a variety of simulation studies and two data examples on high density peptide array.

Two-group comparisons constitute the majority of work in large-scale hypothesis testing. However, many applications involve comparisons among multiple groups

[e.g., Yang et al., 2016, Mergaert et al., 2022]. Large-scale testing tools based on p-values can be directly applied to scenarios comparing multiple groups. However, there is little research on how to improve power in these cases by incorporating more data per unit in empirical Bayes calculations. Motivated by this problem, I consider in Chapter 3 a variant of `MixTwice`, called `MixTwice-ANOVA`, that advances an empirical-Bayes calculation using the numerator and denominator of the unit-specific F statistic as a way to provide additional information for each test. It estimates the mixing distribution for both latent effect and for the variance parameter nonparametrically, and it enforces a monotonic shape constraint on the distribution of effects in order to regularize the semiparametric inference. `MixTwice-ANOVA` is evaluated on a variety of simulation examples and applied in a peptide-array study with multiple Sjogren-disease (SjD) populations.

Numerical experiments have guided the development of large-scale testing tools. In a particularly thorough review, Korthauer et al. [2019] provided a comprehensive study of various operating characteristics: FDR control, power, applicability, consistency, and usability. This review used a battery of simulated data sets and benchmark data sets for its extensive numerical studies. Besides the methods mentioned above (BH, `qvalue`, `ASH`), Korthauer et al. [2019] also evaluated testing tools that use unit-specific covariate data as inputs. This includes FDR regression (FDRreg, [Scott et al., 2015]), conditional local FDR (LFDR, [Cai and Sun, 2009]), Boca and Leek's FDR regression (BL, [Boca and Leek, 2018]), independent hypothesis weighting (`ihw`, [Ignatiadis et al., 2016]), and adaptive p-value thresholding (AdaPT, [Lei and Fithian, 2018])). Appreciating the utility of Korthauer et al. [2019]'s work, I seek to expand the review in order to both include `MixTwice` and also to examine reproducibility properties of the various methods. I present an expanded evaluation in Chapter 4. Besides those conventional evaluation metrics, I introduce another metric called *reproducibility* that measures the similarity between discovery lists from the analysis of replicate data sets. I also include `MixTwice` and two-step `ASH` in order to expand the comparison.

The nonparametric MLE of the mixing distribution is central in many local FDR calculations. The original `MixTwice` method (Chapter 2) relies on a gradient-based constrained optimization method. This is effective but computationally limiting in some examples. As a recent innovation for `MixTwice`, I propose in Chapter 4 an alternative algorithm that combines the Expectation-Maximization algorithm (EM) and the pool adjacent violator algorithm (PAVA). `EM-PAVA` efficiently solves the constrained optimization problem for several large-scale testing methods. In

Chapter 4, I include some background theory on how and why this algorithm works for certain partial-ordering constraints.

As a final example of semiparametric empirical-Bayes hypothesis testing, I study in Chapter 5 a novel method for *graph-associated data*. The data structure and inference problem are similar to what I study in previous chapters, but the testing units in this case are now associated with nodes of a known, undirected graph. This data type arises in many application domains. Examples include peptide arrays and structural brain imaging where the graph records peptide similarity, in one case, and spatial proximity, in the other. The aim of my work is to show how graph information can be used to improve power; I investigate an operationally simple technique that is based upon repeated t-tests for a two-group comparison problem. The proposed `GraphicalT` method does not rely on any detailed modeling specifications, in contrast to other approaches that have tried to improve large-scale testing in this setting [e.g., Sun and Cai, 2009, Liu et al., 2012, Vo et al., 2021]. Preliminary numerical experiments show that `GraphicalT` can usefully improve power without inflating the false discovery rate in graph-associated data problems.

# Chapter 2

# MixTwice: large-scale hypothesis testing for peptide arrays by variance mixing

*The material in this chapter was reported previously in Zheng et al. [2021], and represents a collaborative project with Drs. Mergaert, Ong, Shelef, and Newton, which I led.*

## 2.1 Introduction

Peptide microarray technology is used in biology, medicine, and pharmacology to measure various forms of protein interaction. Like other microarrays, a peptide array contains a large number of very small probes arranged on a glass or plastic chip. Each probe occupies a spatial position on the array, and is comprised of many molecular copies of a short amino-acid sequence (a peptide) anchored to the surface, perhaps 12 to 16 amino acids in length, depending on the design. In antibody profiling experiments, the array is exposed to serum derived from a donor's blood sample; antibodies in the sample that recognize an anchored peptide epitope may bind to the probe. In order to measure these antibody/antigen binding events, a second, fluorescently tagged antibody is applied, which binds to exposed sites on the already-bound antibodies, providing quantitative readout at probes where there has been sufficient binding of serum antibody recognizing the peptide epitopes. High-density peptide microarrays have emerged as a powerful technology in immunoproteomics, as they enable simultaneous antibody-binding measurements against millions of pep-

tide epitopes. Such arrays have guided the discovery of markers for viral, bacterial, and parasitic infections [Mishra et al., 2018, Tokarz et al., 2020, Bailey et al., 2020] and have illuminated the serological response to cancer [Yan et al., 2019] and cancer immunotherapy [Hoefges et al., 2020]. The photolithographic design allows for custom arrays, which have benefited studies of autoimmunity, for example, where various forms of post-translational modification (e.g., citrullination) create targets for autoantibodies [Bailey et al., 2017, Zheng et al., 2020].

The high dimensionality and small sample size of many peptide-array experiments challenge conventional statistical approaches. Zheng et al. [2020], for example, reported a custom peptide-array having 172,828 distinct features and array data from 60 human subjects across several disease subsets. This dimensionality is relatively high compared to gene-expression studies, but quite low compared to other peptide-array studies; arrays that probe the entire human proteome carry over 6 million peptide features, for example. Methods for large-scale hypothesis testing respond to these challenges, often aiming to control the false discovery rate (FDR) [e.g., Efron, 2012]. FDR-controlling procedures are more forgiving than techniques that control the probability of any type I errors (e.g., Bonferroni correction), but they still extract a high penalty for dimensionality in the peptide-array regime involving $10^5$-$10^6$ features. When additional data are available it may be possible to further limit penalties associated with large-scale testing.

Continuing with Zheng et al. [2020], the authors sought to identify peptides for which antibody binding levels differ between control subjects and rheumatoid arthritis (RA) patients expressing a specific disease marker combination (CCP+ and RF-). Sera from twelve subjects in each group were applied to their custom-built array. After pre-processing, a univariate statistic (t-statistic) measured statistical changes at each peptide. Peptides with the most extreme statistics (and smallest p-values) would be set aside for further validation. In the CCP+RF- RA example, no peptides had a FDR-adjusted p-value less than 10% by either the Benjamini-Hochberg (BH) method [Benjamini and Hochberg, 1995] or the more sensitive $q-$value method [Storey et al., 2003], although the latter method estimated that 21% percent of the peptides in fact have differential binding between the two groups.

Improving power while maintaining robustness and reproducibility is a theme of contemporary large-scale inference that we explore in the peptide-array setting. The BH and $q-$value procedures yield no discoveries in the CCP+RF- RA example at one conventional FDR level. If this is due to low statistical power, it may not be surprising since these procedures enter quite late in data analysis, after all p-values have

been computed. Procedures that intervene earlier have access to more information, and thereby may have better overall operating characteristics. Efron's local FDR approach, `locFDR`, intervenes on test statistics just prior to p-value computation and has improved power properties in some settings [Efron et al., 2001]. Independent filtering combines a selection statistic, such as marginal sample variance, and then applies an FDR-controlling procedure to the selected peptides [Bourgon et al., 2010]. Neither `locFDR` nor independent filtering at 50% yielded any results in the CCP+RF- RA example, as it happens. We have the same null finding by independent hypothesis weighting (IHW), which generalizes independent filtering in not requiring a specific selection rate [Ignatiadis et al., 2016].

Adaptive Shrinkage (ASH) is a recent innovation for large-scale testing that intervenes after each peptide yields both an estimated effect and an estimated standard error [Stephens, 2017]. There are several variations of its empirical Bayesian formulation; when using the $t-$distribution sampling-model version of ASH (say ASH-t), we discover 76 peptides to have differential antibody binding in the CCP+RF- RA comparison, also at 10% FDR control. This may reflect increased power, and is consistent with numerical studies showing increased power of ASH in many settings. A recent report from Professor Stephens's group points out a technical limitation of ASH-t that could cause FDR inflation. It proposes a two-step ASH procedure that pre-processes the standard error estimates and then follows with the ASH-t procedure on modified input [Lu and Stephens, 2019]. It happens that we discover 12 peptides with differential binding affinity by two-step ASH at 10% FDR. The different behavior of FDR-controlling procedures in the CCP+RF- RA example exposes ongoing practical challenges that are also revealed in comprehensive numerical studies [Korthauer et al., 2019].

Data analysts face many issues as they filter high-dimensional measurements into short lists for experimental follow-up. In studying this problem, we propose and evaluate a flexible empirical Bayesian mixture method that, like ASH, intervenes after effect estimates and standard errors are computed on each testing unit. The proposed `MixTwice` procedure involves shape-constrained mixture distribution for latent effects and also a separate nonparametric mixture for variance parameters (Section 2). We leverage existing tools for constrained optimization in order to estimate the underlying mixing distributions, and we present a variety of comparative numerical experiments on the operating characteristics of `MixTwice`. The CCP+RF- RA peptide-array example happens to yield 44 peptides having significant differential antibody binding at 10% FDR. A closer look at the identified peptides reveals

binding patterns consistent with other biological information about RA, and thus provides a measure of confidence that these discoveries are not artifacts. In a second RA example where differential signals are stronger, `MixTwice` shows a higher level of reproducibility than other approaches when presented with two independent data sets on the same populations.

## 2.2 Mixture model

We index peptides by $i = 1, 2, \cdots, m$ and suppose that the two-group peptide-array data have been obtained and pre-processed in order to yield two summary statistics per peptide: $(x_i, s_i)$. The first component, $x_i$, is an estimated effect. It measures the difference between the two groups, such as a difference in sample means of log-transformed data, and is viewed a statistical estimate of an underlying effect, say $\theta_i$. In this view, $x_i$ is a random variable having some sampling distribution, which we take to be Gaussian centered at $\theta_i$; this is warranted noting the behavior of suitably-transformed fluorescence measurements coupled with central-limit effects for modest to large sample sizes. The second component, $s_i$, is an estimated standard error. In the Gaussian sampling model, $\mathbb{E}(x_i) = \theta_i$ and $\text{var}(x_i) = \sigma_i^2$, and $s_i^2$ is a sample-based estimate of the variance $\sigma_i^2$. We seek inference about the value of $\theta_i$ using local data $(x_i, s_i)$ as well as data $\{(x_{i'}, s_{i'})\}$ from all peptides, which informs the distribution of effect and variance parameters across the array.

Our formulation is common in large-scale inference, and we could infer $\theta_i$ values in a number of ways. For example, we could produce a peptide-specific p-value from the test statistic $t_i = x_i/s_i$ against the null hypothesis $H_{0,i} : \theta_i = 0$. We might refer $t_i$ to a Student-t distribution, obtain a two-sided p-value, and then process the p-values through the Benjamini-Hochberg (BH) or $q-$value methods to adjust for multiplicity [Benjamini and Hochberg, 1995, Storey et al., 2003]. Alternatively, we might use the collection $\{t_i\}$ and model their fluctuations as a discrete mixture of null and non-null cases, as in the `locFDR` procedure [Efron et al., 2001, Strimmer, 2008]. Both `locFDR` and $q-$value methods are based upon discrete mixtures; interestingly the reduction of $t_i$'s to two-sided p-values entails a loss of sign information that is enough to reduce statistical power in some settings. A more ambitious approach goes beyond null/non-null mixing to allow a full probability distribution of effects $\theta_i$ in order to account for fluctuations across all the peptides. Adaptive shrinkage (ASH) is appealing because it acquires robustness through a nonparametric treatment of this distribution, say $g(\theta)$, while using reasonable shape constraints to regularize the

estimation [Stephens, 2017]. Power advantages of ASH over other methods stem in part from its use of more data per peptide.

In the context of an estimated mixture model there are two useful empirical-Bayesian inference statistics. The first is local false discovery rate (lfdr), $l_i = \mathbb{P}(\theta_i = 0|x_i, s_i^2)$. The term *local false discovery rate* was coined by Professor Efron, and the statistic may be computed in various settings beyond the specific mixture deployed in Efron et al. [2001]. The list $\mathcal{L}$ of statistically significant peptides will be $\mathcal{L} = \{i : l_i \leq c\}$ for some threshold $c$. Notably, small $l_i$ warrants peptide $i$ to be placed in $\mathcal{L}$; but the value $l_i$ is also the probability (conditional on data) that such placement is erroneous [Newton et al., 2006]. Given the data, the expected rate of false discoveries in $\mathcal{L}$ is dominated by $c$. The local false sign rate (lfsr) is analogous to lfdr, but it avoids relying on effects being precisely zero; when the estimated effect is positive for example, the lfsr is $\mathbb{P}(\theta_i \leq 0|x_i, s_i^2)$. Lists controlling lfsr may be constructed in the same way as $\mathcal{L}$, and may be slightly smaller for the same value of $c$. (In the CCP+RF- RA example in Section 1, ASH lfsr and lfdr lists are the same at the 10% level.)

With modest sample sizes, differences between estimated standard errors $\{s_i\}$ and actual standard errors $\{\sigma_i\}$ can affect the performance of existing tools for lfdr and lfds. To better account for these differences we propose an additional mixture layer involving a sampling model $p(s_i^2|\sigma_i^2)$, which we derive from normal-theory considerations, and a flexible nonparametric mixing distribution $h(\sigma^2)$. For both nonparametric components – $g$ on effects $\theta_i$ and $h$ on squared standard errors $\sigma_i^2$ – we use finite grids and treat each distribution as a vector of probabilities. We estimate $g$ and $h$ by maximum likelihood, respecting unimodal shape constraints for $g$ (as in ASH), but otherwise allowing any distributional forms.

Suppose that effects take values in a finite, regular grid $\{a_{-K}, a_{-K+1}, \cdots, a_0, a_1, \cdots, a_K\}$ where $a_0$ is the presumed mode, taken to be $a_0 = 0$ in typical applications in which we aim to retain the null hypothesis of no group difference. We use $K = 15$ in numerical work reported here. Unimodality of the mixing distribution $g = (g_k)$ is expressed as a set of ordering constraints: $g_k \geq g_{k+1}$ for $k = 0, 1, \cdots, K$ and $g_k \leq g_{k+1}$ for $k = -K, -K+1, \cdots, -1$. We also set a second regular grid $\{0 < b_1, b_2, \cdots, b_L\}$ for squared standard errors, and impose no constraints on the mixing distribution $h = (h_l)$ aside from the basic nonparametric essentials: $h_l \geq 0$ and $\sum_l h_l = 1$.

The contribution to the likelihood objective from peptide $i$ is $p(x_i, s_i^2|g, h)$:

$$\begin{aligned} &= \sum_k \sum_l \mathbb{P}(\theta_i = a_k)\, \mathbb{P}(\sigma_i^2 = b_l)\, p(x_i, s_i^2|\theta_i = a_k, \sigma_i^2 = b_l) \\ &= \sum_k \sum_l g_k h_l\, p(x_i|\theta_i = a_k, \sigma_i^2 = b_l)\, p(s_i^2|\sigma_i^2 = b_l) \\ &= \sum_k \sum_l g_k h_l \frac{1}{\sqrt{b_l}} \phi\left(\frac{x_i - a_k}{\sqrt{b_l}}\right) \frac{\nu}{b_l} \chi_{2,\nu}\left(\frac{\nu s_i^2}{b_l}\right) \end{aligned} \tag{2.1}$$

where $\phi$ is the standard normal probability density, $\chi_{2,\nu}$ is the density of a chi-square random variable on $\nu$ degrees of freedom. Under a normal data model, $\nu$ is determined by design (e.g. total samples minus two in the traditional two-sample comparison). The chi-square model is accurate asymptotically for a wide range of non-normal sampling distributions, however the degrees of freedom needs estimation in these cases [O'Neill, 2014].

To estimate the mixing distributions $h$ and $g$ we use the log-likelihood objective function, with terms as in (2.1). In `MixTwice`, we solve the constrained optimization:

$$\begin{aligned} \min_{g,h} -l(g,h) &= -\sum_{i=1}^{m} \log p(x_i, s_i^2|g, h) \\ \text{Subject to:} \quad & g_k, h_l \geq 0 \quad \forall k, l \\ & \sum_k g_k = \sum_l h_l = 1 \\ & g_k \leq g_{k+1}, \quad k \in \{-K, -K+1, ..., -1\} \\ & g_k \geq g_{k+1}, \quad k \in \{0, 1, ..., K\} \end{aligned} \tag{2.2}$$

The gradient and Hessian of $l(g, h)$ are readily available, and so (2.2) may be solved efficiently using augmented Lagrangian for constrained optimization, using the BFGS algorithm for inner loop optimization, which is implemented in the R package alabama [Varadhan, 2015]. We extract lfdr and lfsr statistics from the peptide-specific posterior distributions at the optimized vectors $\hat{g}, \hat{h}$: $\mathbb{P}(\theta_i = a_k|x_i, s_i^2)$

$$\begin{aligned} &= \sum_l \mathbb{P}(\theta_i = a_k, \sigma_i^2 = b_l|x_i, s_i^2) \\ &\propto \hat{g}_k \sum_l \hat{h}_l \frac{1}{\sqrt{b_l}} \phi\left(\frac{x_i - a_k}{\sqrt{b_l}}\right) \frac{\nu}{b_l} \chi_{2,\nu}\left(\frac{\nu s_i^2}{b_l}\right). \end{aligned} \tag{2.3}$$

Proportionality is resolved by summation over the grid $k$, and we get:

$$\begin{aligned} \text{lfdr}_i &= \mathbb{P}(\theta_i = a_0 | x_i, s_i^2), \\ \text{lfsr}_i &= \min\left\{ \sum_{k \leq 0} \mathbb{P}(\theta_i = a_k | x_i, s_i^2), \sum_{k \geq 0} \mathbb{P}(\theta_i = a_k | x_i, s_i^2) \right\}. \end{aligned}$$

It may be helpful to recognize that by contrast to (2.3), ASH-normal would entail

$$\mathbb{P}(\theta_i = a_k | x_i, s_i^2) \propto \hat{g}_k \frac{1}{s_i} \phi\left( \frac{x_i - a_k}{s_i} \right), \tag{2.4}$$

and ASH-$t$ would replace the normal density $\phi$ in (2.4) with a Student $t$ density; in both cases the ASH-estimated mixing density $\hat{g}$ would come not from (2.2) but from an objective in which mixing over variances is not explicitly accommodated. The initial implementation of `MixTwice` invokes unimodality shape constraint, but not symmetry, and, for computational convenience, allows that a random subset of the testing units is used in the optimization. We investigate this approximation in Supplementary Material.

## 2.3 Simulation Study

We are interested in the performance of `MixTwice` in scenarios reflecting what might be expected to occur in practice and have performed numerical experiments involving different generative distributions of both effects ($g$) and variances ($h$). Noting the special role of the null value, $\theta = 0$, our experiments involve mixtures $g(\theta) = \pi_0 \delta_0 + (1 - \pi_0) g_{\text{alt}}(\theta)$, where $\pi_0 = \mathbb{P}(\theta_i = 0)$ and $g_{\text{alt}}$ provides various ways to distribute mass away from zero. Following Stephens [2017] and Lu and Stephens [2019], we entertain different general shapes, including so-called *big-variance*, *bi-modal*, *flattop*, *normal*, and *spiky*. `MixTwice` accounts for explicit differences between sample and underlying standard errors, and mixes nonparametrically over these underlying standard errors. Our numerical experiments consider the simplest case in which the data generating $h$ is a point mass, a case involving a finite mixture of two values, and also a continuous case of inverse-Gamma-distributed parameters. Patterns in the error of estimation and the hypothesis testing error rates are very comparable across different choices of $h$, and so for simplicity here we report only experiments when this true $h$ is a point mass distribution. Figures 1 and 2 summarize, respectively, properties of estimation accuracy and testing error rates. Experiments are based on Gaussian samples with

unit variance, $m = 1000$ peptides, and various sample size settings for the two-group comparison.

If a method tends to overestimate $\pi_0$, then power may be reduced; in case of underestimation the FDR may be inflated. Figure 1, Panel B, focuses on the estimation of this marginal null frequency for one choice of sample size, namely $n = 10$ observations per group. In each setting of $g$ (column), 500 data sets are generated, each drawn after its own $\pi_0$ value was uniformly drawn in $[0.5, 1]$. All methods respond appropriately to changes in $\pi_0$, though they exhibit different biases; `MixTwice` tracks the identity line (no bias) case closely in all scenarios except the challenging *spiky* case of $g_{\text{alt}}$. By contrast `locFDR` is conservatively biased, tending to over-estimate $\pi_0$ in most cases. Our experiments include an oracle case, namely ASH−normal, which takes the underlying standard errors as known. This numerical control helps us gauge the magnitude of statistical errors induced by estimation error of the variance profile.

Figure 2.1, Panel C, amplifies one case from the second row, when $\pi_0 = 0.9$, and shows how estimation error drops as the sample size per peptide grows. Most methods display a level of convergence in this setting, with `MixTwice` performing relatively well especially at low sample sizes. Going beyond the estimation of $\pi_0$, we compared methods by their 1-Wassertstein error in estimating the entire mixture distribution $g$; `MixTwice` showed relatively small error in this setting also (data not shown). `MixTwice` shares with other nonparametric mixture methods the identifiability problem that only an upper bound on $\pi_0$ may be reliably estimated from limited data [Efron et al., 2001, Stephens, 2017]. This may be appreciated by considering a single unit, $i$, on which the estimated effect $\hat{\theta}_i$ is a normal deviation from $\theta_i$, say with known variance $\sigma^2 = 1$, and ignoring the second level of mixing. If $g_{\text{alt}}$ concentrates enough mass near $\theta = 0$, then the null predictive density $\phi(x)$, of $\hat{\theta}_i$, may be partially absorbed by the alternative predictive density: i.e., there may be a $c > 0$ such that for all $x$, $c\phi(x) \leq \int \phi(x - \theta) g_{\text{alt}}(\theta)\, d\theta$, in which case an amount $c(1 - \pi_0)$ of putatively alternative mass could be swapped into the null component without changing the marginal predictive density. Sampling scenarios that allow for decreasing standard errors for at least a fraction of the units resolve this methodological issue. We can show, for symmetric $g_{\text{alt}}$ for example, that the gap $c$ vanishes to zero as the standard error $\sigma$ similarly converges (see Supplementary Material). This is consistent with numerical behavior of `MixTwice` in large samples (Fig. 1C), and is also consistent with work on mixture identification as information per unit increases [Ritchie et al., 2020, Aragam et al., 2020].

Figure 2.2 confirms that most methods are controlling FDR as advertised. The empirical false discovery rate is plotted against the controlled rate; the latter is the nominal target FDR value where we threshold the lfdr's; the former is what is evident from knowing the simulation states (in other terminology, it is the average, over simulated data sets, of the false discovery proportion). Colored lines are used to distinguish different levels of $\pi_0$, when the signal is dense (with a lower null proportion $\pi_0$) or when the signal is sparse (with a higher null proportion $\pi_0$). Recall we simulated independent data sets each governed by a randomly chosen $\pi_0$ from $[0.5, 1]$. In order to visualize the results, we stratified data sets into four groups and averaged internally: $0.5 \leq \pi_0 \leq 0.625$, $0.625 \leq \pi_0 \leq 0.75$, $0.75 \leq \pi_0 \leq 0.875$, $0.875 \leq \pi_0 \leq 1$. The FDR inflation by ASH-t at high $\pi_0$ is evident in this simulation.

## 2.4 Empirical studies

### 2.4.1 Antibodies in rheumatoid arthritis

Rheumatoid arthritis (RA) is a chronic autoimmune disease characterized by inflammation and pain, primarily in the joints. RA patients produce autoantibodies against many different "self" proteins. Most famously, they generate antibodies against proteins in which arginine amino acids have been post-translationally modified to citrullines [Schellekens et al., 1998] as well as antibodies that bind to antibodies, called rheumatoid factor (RF) [Waaler, 1940]. Both autoantibody types appear to be pathogenic [Sokolove et al., 2014] and both are used diagnostically [Aletaha et al., 2010], the former detected by the anti-cyclic citrullinated peptide (CCP) test. Most RA patients make both autoantibody types (CCP+RF+ RA), but some have only one type like in CCP+RF- RA. Little is known about why CCP+RF+ versus CCP+RF- RA develops. However, a better understanding of the autoantibody repertoires in each RA subset could provide insights, a task for which peptide arrays are perfect.

The custom high-density peptide array reported in Zheng et al. [2020] probed 172,828 distinct 12 amino acid length peptides derived from 122 human proteins suspected to be involved in RA, including peptides in which all arginines were replaced by citrullines. We reconsider here two distinct comparisons from that study, namely the comparison between CCP+RF- RA patients and controls, and a second comparison between CCP+RF+ RA patients and controls, in which differential signals are much stronger. Both comparisons have 12 subjects in each group. To assess

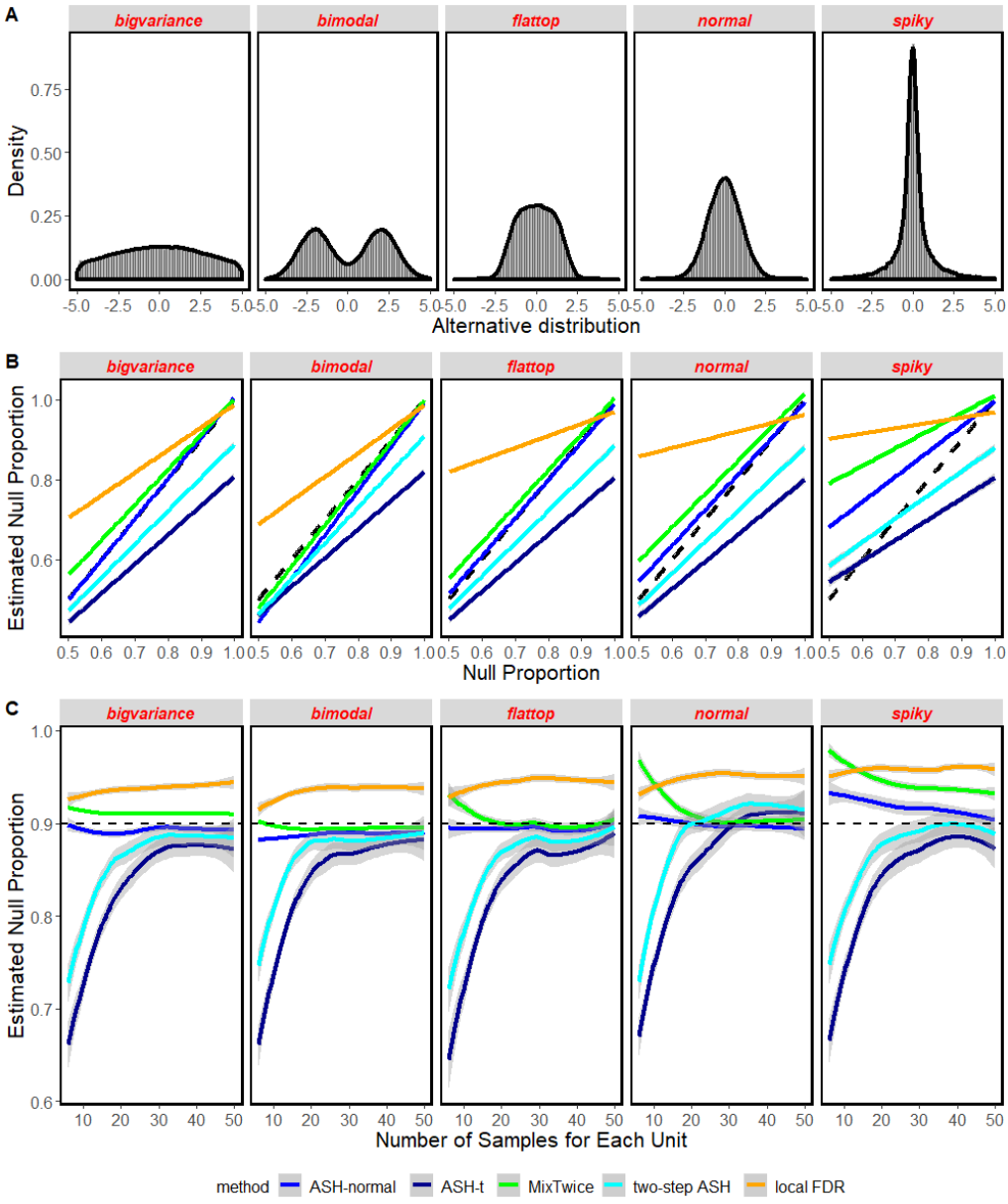


Figure 2.1: **Errors in Estimation of $\pi_0$.** Panel A shows distributions used for $g_{\text{alt}}(\theta)$. Panel B shows the estimation of null proportion $\pi_0$ in case of equal samples in each group of 10. Methods are distinguished by color, where we report average parameter estimates from 500 simulated data sets.The identity line (dashed) indicates no bias. ASH-normal is an oracle case in which $\sigma_i^2 = 1$ is provided to the algorithm. Panel C shows error estimation as the number of observations grows, in $\pi_0 = 0.9$.

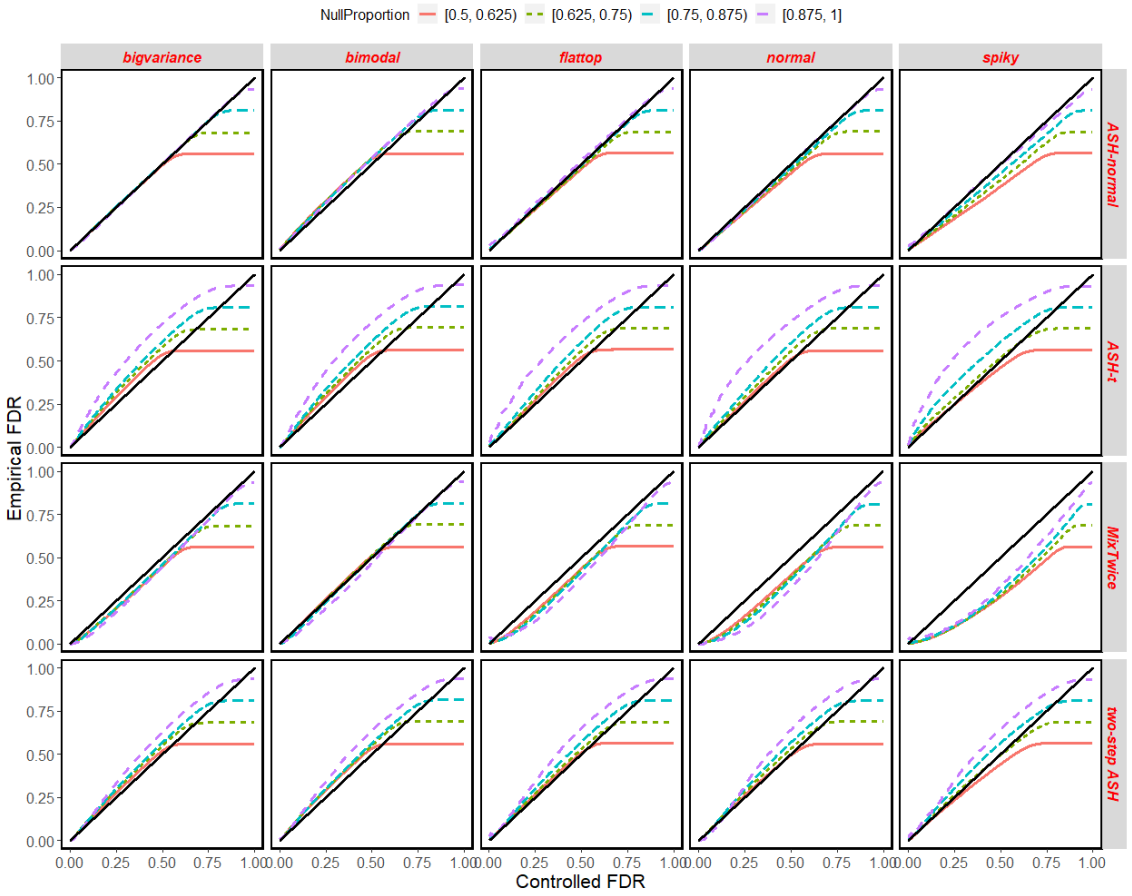


Figure 2.2: **Synthetic data and FDR control:** False discovery rates are shown by different methods (rows) under different alternative distributions $g_{\text{alt}}(\theta)$ (columns). Empirical FDR (vertical) is the achieved error rate in the simulation; controlled FDR (horizontal) is the rate targeted by the methodology. Results with different $\pi_0$ are coded using different colors. A method tends to inflate FDR over the target level if its curve is greater than the identity line; it is conservative when its curve is dominated by the identity line.

reproducibility, we take advantage of a second peptide array data set derived from an independent set of 8 controls and 8 CCP+RF+ RA patients.

### 2.4.2 CCP+RF- RA: weak signals

We applied `MixTwice` to fit the shape-constrained mixture model of Section 2. Fitted mixing distributions are visualized in Figure 2.3 and provide a measure of the magnitude of changes in mean antibody levels as well as the magnitude of sampling variation. For example, the effect-size distribution estimates no probability for effects larger than 0.037. Also, the median standard error is 0.10 (squared standard error 0.01), which is large compared to the probable effect sizes.

In Section 1 we presented summary counts of peptides identified at 10% FDR that exhibit differential binding between CCP+RF- RA patients and non-RA controls. `MixTwice`, ASH-t, and two-step ASH distinguish themselves in being the only methods among many standard large-scale tools to populate non-empty lists of discovered peptides at that FDR level. Recognizing that the magnitude of signal intensities on the peptide array is an important aspect of downstream analysis, Figure 2.4 shows a summary of the identified peptides by various methods. Notably, `MixTwice` and two-step ASH detect peptides in this case with higher average signal intensity than ASH-t; these may correspond to higher antibody abundance or affinity and potentially easier validation. ASH-t tends to select peptides with low standard errors, even when the estimated effects are very low.

Interestingly, the 44 peptides found by `MixTwice` have a strong pattern in their peptide sequences: all are citrulline ($B$)-containing peptides (which would be predicted for CCP+ RA patients) and contain citrulline next to glycine ($B$-$G$ or $G$-$B$), as shown in the motif in Figure 2.5. Binding of antigens in which citrulline is next to glycine is consistent with a growing body of literature on the reactivity of anti-citrullinated protein antibodies in RA [e.g., Burkhardt et al., 2002, Szarka et al., 2018, Steen et al., 2019, Zheng et al., 2020].

As a further negative control calculation, we applied `MixTwice` to each of 500 permuted data sets obtained by fixing the peptide data and randomly shuffling the 24 subject labels (12 control, 12 CCP+RF- RA). In 493 cases, the 10% FDR list is empty; 6 cases find a single peptide and one case finds 2 peptides at this threshold.

Among a number of large-scale testing methods applied to the CCP+RF- RA example, `MixTwice` identifies a comparatively large number of statistically significant peptides. By contrast to other methods, these peptides contain patterns in their

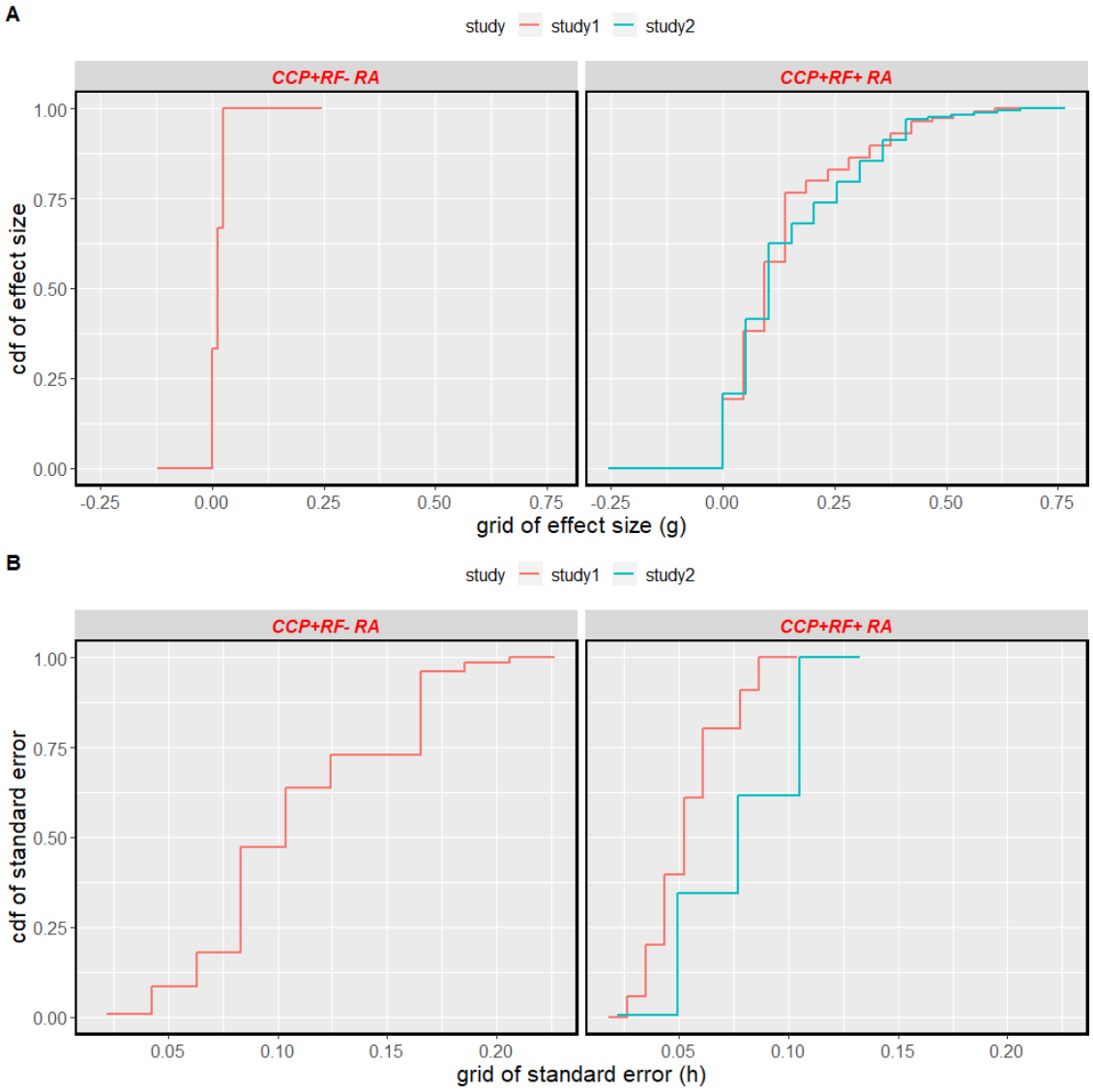


Figure 2.3: **Estimated mixing distributions:** For both effect distribution $g$ (Panel A) and squared-standard-error distribution $h$ (Panel B), shown are the maximum likelihood estimated mixing distributions as cumulative distribution functions (cdfs) in double natural log scale. The CCP+RF- RA example is shown on the left and the two CCP+RF+ RA examples are on the right.

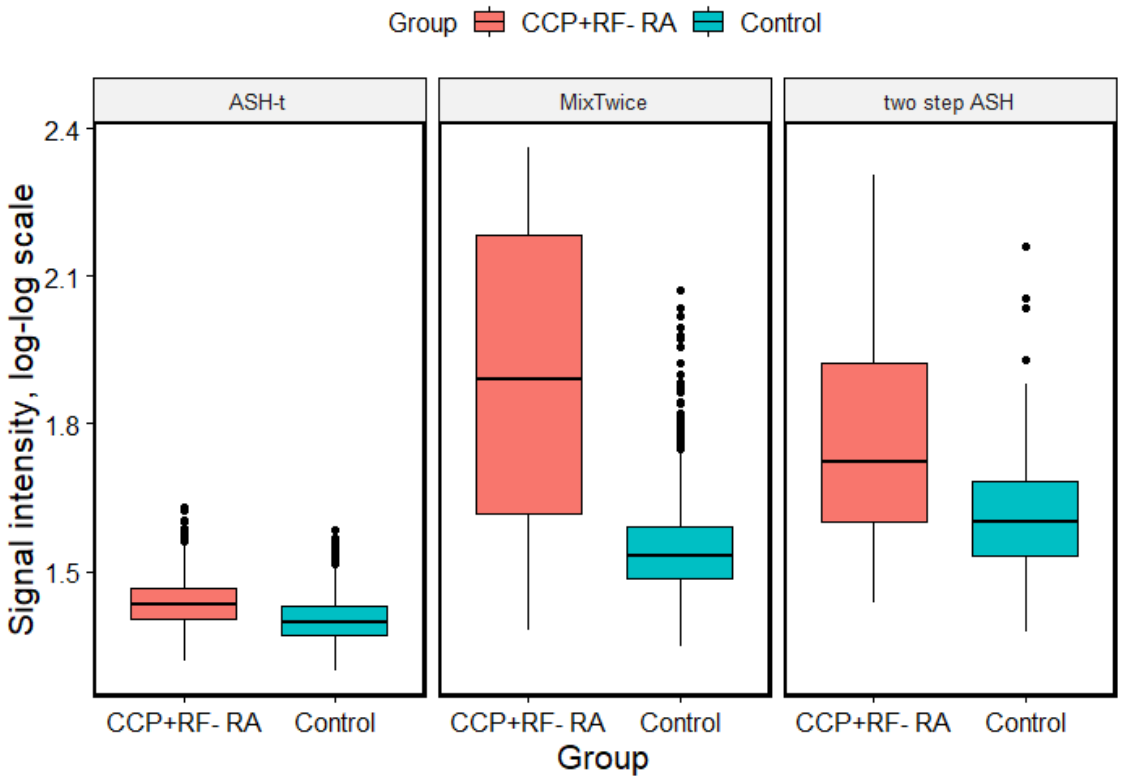


Figure 2.4: **Signal intensity of differentially abundant peptides:** Boxplots show averaged signal values on double natural log scale (both CCP+RF- RA and control subjects) for peptides found by ASH-t (76 peptides), MixTwice (44 peptides), and two-step ASH (11 peptides) all discovered at 10% FDR.

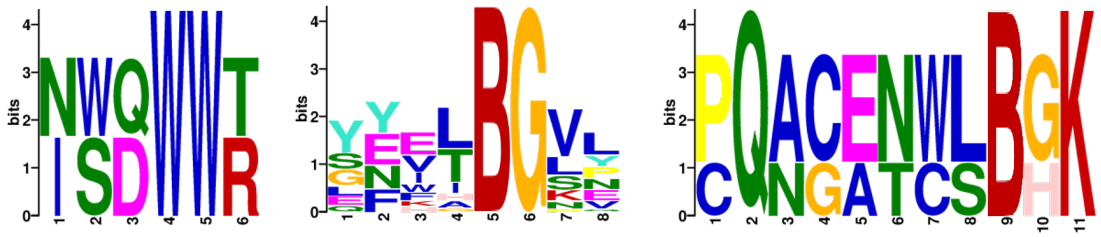


Figure 2.5: **Motif logo for significant peptides in CCP+RF- RA:** Consensus sequences were generated using online software MEME Suite [Bailey et al., 2009] and the significant peptides from the different methods: ASH-t (left), MixTwice (middle) and two-step ASH (right). Each position of the motif logo represents the empirical distribution of amino acids at that site, with size proportional to frequency. $B$ found in the middle and right panels is citrulline, a post-transitionally modified arginine. The overall height of each stack is an information measure (bits) related to the concentration of the empirical distribution on its support.

amino acid sequences consistent with emerging evidence on this disease, and they correspond to relatively high fluorescence intensity measurements. Together, these observations provide some assurance that the `MixTwice` findings are not artifacts.

### 2.4.3 CCP+RF+ RA: strong signals

One of the findings from Zheng et al. [2020] concerns the extensive antibody-profile differences between RA patients who are positive for both biomarkers (CCP+RF+) and control subjects. Statistically, it represents an interesting non-sparse, large-scale testing situation, and the immunological mechanisms driving this remain only partially understood. To check the reproducibility of peptide-array findings, a new experiment was performed using the same procedures and 172,828 peptide array to detect IgG binding as in Zheng et al. [2020], but with serum samples from 16 different subjects: 8 CCP+RF+ RA and 8 controls. CCP+RF+ RA and control subjects were similar in regards to age, sex, race, ethnicity, and overall health. Preprocessing followed the same protocol and provided a data set (*study 2*) for us to look at reproducibility of large-scale hypothesis testing methods.

Z-score histograms in Figure 2.6, Panel A, show that both studies reveal exten-

sive increased antibody binding in the CCP+RF+ RA group. The scatterplot in Panel B reveals concordance between the studies on this z-score metric. The color-coding highlights discovered peptides at the 0.1% FDR method by `MixTwice`, both uniquely in one study (green or yellow) and reproducibly in both studies (blue). Of course `MixTwice` uses more information than is in the z-score summary, but the scatterplot provides a convenient visualization. The lower panels in Figure 2.6 compare reproducibility statistics of different testing methods at various FDR thresholds. Denoting by $\mathcal{L}_j(\alpha)$ the list of significant peptides in study $j$ and FDR level $\alpha$, we have $|\mathcal{L}_1(\alpha) \cap \mathcal{L}_2(\alpha))|$ as the number of peptides identified in both studies (Panel D) and $\frac{|\mathcal{L}_1(\alpha) \cap \mathcal{L}_2(\alpha))|}{|\mathcal{L}_1(\alpha) \cup \mathcal{L}_2(\alpha))|}$ as the common fraction (Panel C). By connecting separate, independent studies of the same group difference, these statistics measure the reproducibility of various large-scale testing methods. `MixTwice` shows substantially better reproducibility than other testing methods, such as ASH-t, two-step ASH, and `locFDR` in this example.

## 2.5 Discussion

High-throughput biomedical experiments, such as those involving peptide arrays and immunological studies, continue to provide challenging problems for large-scale hypothesis testing. Readily applied techniques, such as $q-$value, `locFDR`, `IHW`, and ASH are often very effective at reporting lists of testing units (peptides) showing statistically significant effects at a targeted false discovery rate. In the case of high-density peptide arrays, we find several examples where these tools are deficient. One issue is the number of testing units, which is an order of magnitude larger than what is seen in transcript studies, for example. In the CCP+RF- RA comparison, most existing tools exhibit low power, which may stem in part from when they intervene in the data analysis. Methods that intervene earlier have access to more information and thereby may gain some advantage. The risk to intervening early is that more assumptions may be required to deliver relevant testing statistics (e.g., lfdr, lfsr). We rely on external validation, such as on sequence properties of the identified peptides, to assess practical utility. The CCP+RF+ RA example showcases a situation where power is high by all methods, and the differences boil down to how testing units are prioritized. The proposed `MixTwice` procedure shows impressive reproducibility in this case.

Structurally, `MixTwice` is similar to the ASH method for large-scale testing: it aims to estimate a mixing distribution of effects in an empirical Bayesian formu-

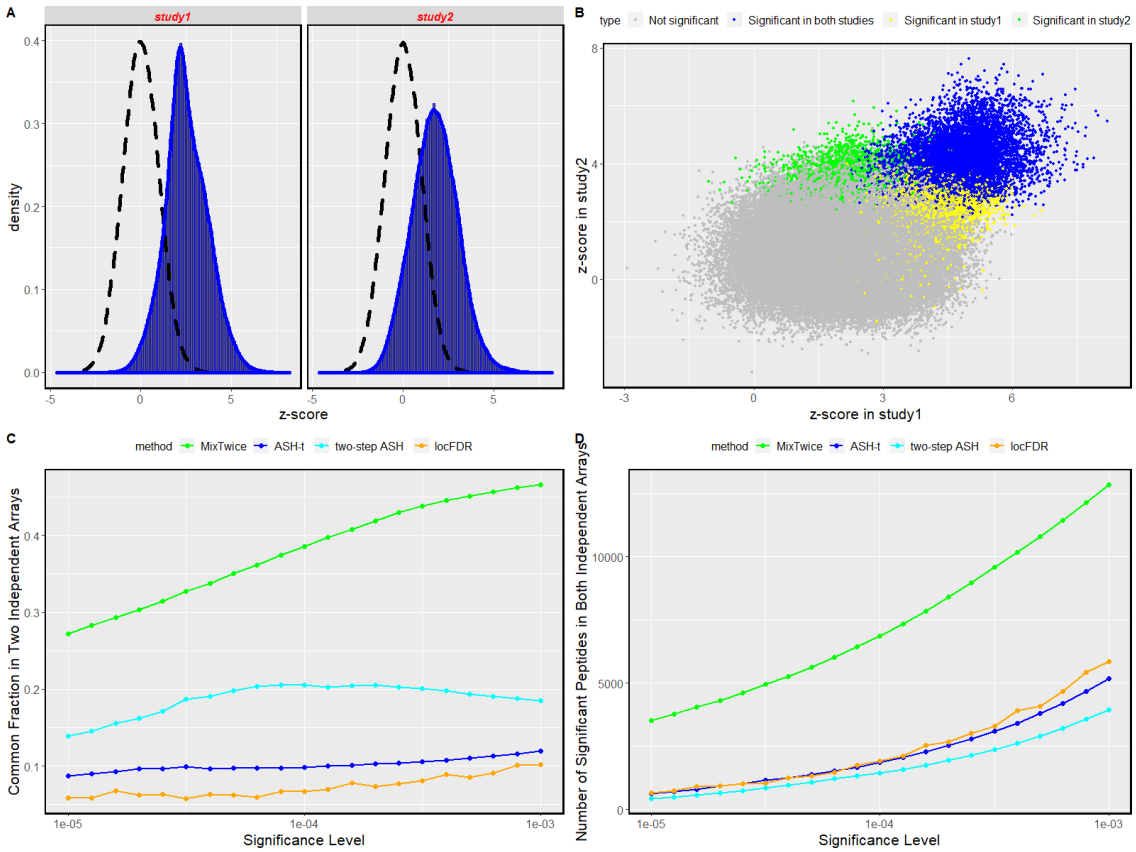


Figure 2.6: **Reproducibility comparison.** Panel A shows empirical z-score distributions for CCP+RF+ RA vs control at 172,828 peptides in two independent studies. The scatterplot in Panel B highlights peptides identified uniquely at 0.1% FDR by MixTwice in either study (yellow, green) and those reproducibly found in both studies (blue). Metrics in Panels C and D compare performance of MixTwice as a function of FDR threshold.

lation. It adopts ASH's nonparametric, shape-constrained model for effects, but deviates from that approach by incorporating a second mixing layer over underlying effect-variance parameters. A number of methodological issues deserve further study. For example, `MixTwice` treats the sampling model of squared standard errors as chi-square on a design-based degrees of freedom, which is rooted in a normal-data model. We expect that suitable transformation of the original data will make this treatment reasonable; for example, Zheng et al. [2020] proposed a double-log transform to stabilize variance. An interesting alternative is to use a bootstrap scheme to assess the sampling distributions directly, in order to thereby estimate the degrees of freedom that would be justified asymptotically for non-normal cases.

There are computational issues that warrant further investigation. The objective function (2.2) may not be convex in the pair of arguments $(g, h)$. Numerical experiments indicate good performance of the augmented Lagrangian optimization approach in a range of scenarios, though alternative approaches may have benefits. For example, the conditional optimizations of $g$ given $h$ or $h$ given $g$ are both convex, though attempts so far to leverage this have been less computationally efficient than the augmented Lagrangian method. Related to this are questions of grid sizes $K$ and $L$, which have to balance fidelity to the data and computational efficiency.

Though our presentation has focused on the classical two-group comparison problem, it should be evident that the core methodology is not restricted to this case. Estimated effects $x_i$, for example, could arise from a contrast of interest after adjusting for blocking variables or other covariates. These will be useful to consider as we expect them to emerge in experiments that further investigate mechanisms of immune-system disregulation.

Finally, we point out that other forms of information may be usefully integrated with the testing methodology. Peptides tile proteins, though we have treated them as anonymous testing units. More sophisticated peptide prioritization could leverage amino-acid structure, protein content, or other features of the immunological context.

# Chapter 3

# MixTwice-ANOVA: large-scale testing with multiple groups

## 3.1 Introduction

Two-group comparisons constitute the bulk of applied and methodological work in large-scale hypothesis testing. However, research interests might be generalized from scenarios comparing two groups to scenarios comparing multiple groups. For example, in antibody profiling research, investigators seek to understand epitope features across disease patients with multiple biomarkers [e.g. Zheng et al., 2020, Mergaert et al., 2022]. In the present chapter, I will introduce a new tool aiming to improve large-scale testing for multi-group comparisons.

One approach to improving the operating characteristics of large-scale testing methods is to increase the amount of summary data that is computed on each unit and fed into an empirical-Bayes mixture calculation. Storey's `q-value` [Storey, 2002], for example, works with unit-specific p-values; Efron's `locFDR` [Efron et al., 2001] works with unit-specific z-scores, and thereby retains sign-information that is lost to the p-value. Adaptive Shrinkage (`ASH`, [Stephens, 2017]) and `MixTwice` ([Zheng et al., 2021]) work with estimated effects and standard errors (roughly, the numerator and denominator of the z-score). In each case, the summary statistics per unit are acted upon by an empirical-Bayesian mixture computation, say to compute a local false discovery rate, a local false sign rate, or a tail-area q-value, from which a prioritized and error-rate controlled list of discoveries may be computed.

Those large-scale tools working with unit-specific p-values can be directly migrated from scenarios comparing two conditions to scenarios comparing multiple

conditions. However, those methods working with less reduced data (i.e., Efron's `locFDR`, `ASH`, `MixTwice`, etc.), even though are shown to have increased power properties under two-group comparison, have been less studied under the one-way ANOVA setting. Motivated by this, here we consider a variant of `MixTwice`, called `MixTwice-ANOVA`, that advances an empirical-Bayes calculation using the numerator and denominator of the unit-specific F statistic as a way to provide additional information for each test. The proposed procedure involves a nonparametric mixture distribution for latent effects with monotone shape constraint and also a separate nonparametric mixture for variance parameters. After a comprehensive simulation analysis for FDR control and power evaluation, it was applied on two peptide data examples: one compares three groups of Sjogren disease (SjD) patients where the signal is weak and the other compares rheumatoid arthritis (RA) patients with different biomarkers where the signal is strong.

## 3.2 Methodology

### 3.2.1 Data and sampling model

We consider the setting in which independent sampling units each provide high ($p$) dimensional observations, and we focus our notation on a single inference unit $i$ in $\{1, 2, \cdots, p\}$. We suppose that each sample is assigned to one of $m$ different populations, with say $n_j$ samples from population $j$. After any preprocessing, the data for testing unit $i$ are: $\{X_{i,j,k},\ j = 1, \cdots, m,\ k = 1, \cdots, n_j\}$. In total we have $n = \sum_{j=1}^{m} n_j$ samples. We are interested in hypothesis tests about expected values in the different populations. Denote the mean parameter within each group as $\mu_{i,j} = \mathbb{E}(X_{i,j,k})$, and let $\sigma_i^2 = \text{var}(X_{i,j,k})$ be the variance, assumed here to not depend on group $j$, but unknown and allowed to fluctuate among testing units. Of specific interest to us is testing the null hypothesis for unit $i$:

$$\text{H}_{0,i} : \mu_{i,1} = \mu_{i,2} = \cdots = \mu_{i,m}. \tag{3.1}$$

If preprocessing has stabilized variance, then an appropriate test of $\text{H}_{0,i}$ is the classical one-way ANOVA F test, with components as below, where SSB is the Sum of Squares in Between and SSE is the Sum of Squares in Error, and averages are indicated as in standard ANOVA notation:

$$F_i = \frac{\text{SSB}_i/(m-1)}{\text{SSE}_i/(n-m)},\ \text{SSB}_i = \sum_{j=1}^{m} n_j(\bar{X}_{i,j,\cdot} - \bar{X}_{i,\cdot,\cdot})^2,\ \text{SSE}_i = \sum_{j=1}^{m}\sum_{k=1}^{n_j}(X_{i,j,k} - \bar{X}_{i,j,\cdot})^2.$$

Recalling normal distribution theory [e.g., Casella and Berger, 2021, page 537], the sampling distribution of $F_i$ is available under both $\mathrm{H}_{0,i}$ and the alternative, and relies on independence and chi-square distributional forms of SSE and SSB, as well as on the effect-size parameter $\lambda_i = \frac{1}{n}\sum_{j=1}^{m} n_j(\mu_{i,j} - \bar{\mu}_i)^2$, where $\bar{\mu}_i = \frac{1}{n}\sum_{j=1}^{m} n_j\mu_{i,j}$:

$$\frac{\mathrm{SSB}_i}{\sigma_i^2}|\sigma_i^2, \lambda_i \sim \chi^2_{m-1,\frac{n\lambda_i}{\sigma_i^2}} \tag{3.2}$$

$$\frac{\mathrm{SSE}_i}{\sigma_i^2}|\sigma_i^2, \lambda_i \sim \chi^2_{n-m,0} \tag{3.3}$$

where $\chi^2_{k,\lambda}$ is a chi-squared distribution with degree of freedom $k$ and non-centrality parameter $\lambda$. These distributional forms are exact under normal theory for $\{X_{i,j,k}\}$, but they are known to be somewhat robust to non-normality, and they also are good approximations to the permutation distributions of the test statistics [Pearson, 1931].

### 3.2.2 Mixture model

From data, we first compute for each testing unit $i$ the pair of test statistics $(\mathrm{SSB}_i, \mathrm{SSE}_i)$. Beyond the chi-square sampling model, we treat the pair-specific parameters $(\lambda_i, \sigma_i^2)$ as themselves arising from a distribution. Specifically, we assume that parameters of interest $\lambda_i$ and the secondary parameters $\sigma_i^2$ fluctuate independently across the system according to unknown mixing distributions $g(\lambda)$ and $h(\sigma^2)$. We allow mixing mass on the point null $\lambda_i = 0$, and so $g$ is considered to put some probability mass there; computationally we treat $g$ as a finite mixing distribution supported on some grid of effect sizes.

Empirical-Bayesian inference relies on the local false discovery rate (lfdr):

$$l_i = \mathbb{P}(\lambda_i = 0|\mathrm{SSB}_i, \mathrm{SSE}_i) \tag{3.4}$$

$$\propto g(0)\int h(\sigma^2)\,\frac{1}{\sigma^2}\chi^2_{m-1,0}\left(\frac{\mathrm{SSB}_i}{\sigma^2}\right)\frac{1}{\sigma^2}\chi^2_{n-m,0}\left(\frac{\mathrm{SSE}_i}{\sigma^2}\right)d\sigma^2 \tag{3.5}$$

where proportionality is resolved by integrating (really summing) out over the mixing distribution $g$. The discovery list of significant units is $\mathcal{L}_c = \{i : l_i \leq c\}$ for some threshold $c$ and the controlled FDR is the arithmetic mean of $l_i$'s for $i \in \mathcal{L}_c$.

### 3.2.3 Estimation and Computation

To evaluate lfdr's requires us to estimate mixing distributions $g$ and $h$. As with `MixTwice`, we approximate these distributions with finite probability vectors each

on a grid of possible parameter values. Suppose that effects take values in a finite, regular grid $\{0 = a_0, a_1, \cdots, a_K\}$ where $a_0 = 0$ is for point mass indicating no group difference. We also set a second regular grid $\{0 < b_1, b_2, \ldots, b_L\}$ for $\sigma^2$. The mixing distribution $g$ and $h$ are then $(g_k)$ and $(h_l)$, to be estimated, with basic nonparametric essentials: $g_k \geq 0, h_l \geq 0, \sum_k g_k = \sum_l h_l = 1$.

We force the non-decreasing shape constraint on $g$ by $g_0 \geq \cdots \geq g_K$ as it is reasonable to expect that larger effect units would usually be less plausible, and so the distribution of effect will be non-decreasing. This shape constraint might be beneficial through a variety of perspectives. First, if the system really has such feature that the frequency of effects of a given size diminishes with size, then a statistical procedure that enforces this shape constraint is expected to have better statistical properties than one which does not (Marshall's lemma on Marshall [1970] and discussion on Groeneboom and Jongbloed [2018]). Second, the shape constraint provides a regularization on the nonparametric estimation without entailing any parametric assumptions to the model, and is understood to reduce estimation variability. Finally, there is one special motivation for the shape constraint in the FDR-controlling cases. Specifically, such non-decreasing constraint would shrink the estimates towards the null and therefore will not put an inference unit on the discovery list unless there is sufficiently strong evidence for that placement. Had we not enforced monotonicity, statistical fluctuations could more easily push uninteresting units on this list.

Mixing distribution $g = (g_k)$ and $h = (h_l)$ are estimated through non-parametric maximum likelihood. The contribution to the likelihood objective of unit $i$ is $p(\mathrm{SSB}_i, \mathrm{SSE}_i | g, h)$:

$$\begin{aligned}
&= \sum_k \sum_l \mathbb{P}(\lambda_i = a_k)\, \mathbb{P}(\sigma_i^2 = b_l)\, p(\mathrm{SSB}_i, \mathrm{SSE}_i | \lambda_i = a_k, \sigma_i^2 = b_l) \\
&= \sum_k \sum_l g_k h_l\, p(\mathrm{SSB}_i | \lambda_i = a_k, \sigma_i^2 = b_l)\, p(\mathrm{SSE}_i | \sigma_i^2 = b_l) \\
&= \sum_k \sum_l g_k h_l\, \frac{1}{b_l} \chi^2_{m-1, \frac{na_k}{b_l}} \left( \frac{\mathrm{SSB}_i}{b_l} \right) \frac{1}{b_l} \chi^2_{n-m,0} \left( \frac{\mathrm{SSE}_i}{b_l} \right). \qquad (3.6)
\end{aligned}$$

Nonparametric maximum likelihood provides an effective and computationally efficient strategy to estimate the underlying mixing distributions $g$ and $h$. In

`MixTwice-ANOVA`, I solve the constrained optimization:

$$\begin{aligned} \min_{g,h} -l(g,h) &= -\sum_{i=1}^{m} \log p(\text{SSB}_i, \text{SSE}_i | g, h) \\ \text{Subject to:} \quad & g_k, h_l \geq 0 \quad \forall k, l \\ & \sum_k g_k = \sum_l h_l = 1 \\ & g_0 \geq g_2 \geq \cdots \geq g_K. \end{aligned} \tag{3.7}$$

Similar as in `MixTwice`, this constrained optimization problem could be solved using Augmented Lagrangian algorithm. In the next chapter (Chapter 4), I will introduce another algorithm, combining Pool Adjacent Violator Algorithm (PAVA) with Expectation Maximization (EM) algorithm, that could solve this optimization problem more efficiently.

The local false discovery rate statistic, $\text{lfdr}_i = \mathbb{P}(\lambda_i = 0 | \text{SSB}_i, \text{SSE}_i)$ could be extracted from the posterior distributions at the optimized vectors $\hat{g}, \hat{h}$:

$$\begin{aligned} \mathbb{P}(\lambda_i = a_k | \text{SSB}_i, \text{SSE}_i) &= \sum_l \mathbb{P}(\lambda_i = a_k, \sigma_i^2 = b_l | \text{SSB}_i, \text{SSE}_i) \\ &\propto \hat{g}_k \sum_l \hat{h}_l \frac{1}{b_l} \chi^2_{m-1, \frac{na_k}{b_l}} \left( \frac{\text{SSB}_i}{b_l} \right) \frac{1}{b_l} \chi^2_{n-m,0} \left( \frac{\text{SSE}_i}{b_l} \right) \end{aligned} \tag{3.8}$$

This formulation approximates the integral in Equation 3.5.

## 3.3 Simulation study

### 3.3.1 FDR control and power

In this section, I examine the FDR control and power of `MixTwice ANOVA` and other large-scale testing tools. Performance of `MixTwice-ANOVA` is compared with three approaches only requiring the p-values as statistic (Bonferroni [Bonferroni, 1936], BH [Benjamini and Hochberg, 1995] and `qvalue` [Storey, 2002]) and four other approaches that take p-values and covariates as input statistics: (LFDR [Cai and Sun, 2009], AdaPT [Lei and Fithian, 2018], BL [Boca and Leek, 2018], `ihw` [Ignatiadis et al., 2016]).

To understand the performance associated with levels of signal and noise, I vary the levels of effect size to either *large signal* (i.e., the expectation of non-null signal $\mathbb{E}(\lambda | \lambda > 0) = 4$) or *small signal* ($\mathbb{E}(\lambda | \lambda > 0) = 1$). Similarly, scenarios of *large*

*variance* ($\sigma = 3$) and *small variance* ($\sigma = 1$) are included in the experiments. Under each setting, simulation results are averaged over 30 independent trials with a randomly drawn $\pi_0 \sim U(0,1)$ on a data with $p = 3000$ number of testing units, $m = 5$ groups and $n_j = 4$ replicates within each group. A logistic function, $w(u) = \frac{1}{1+\exp(-10u+5)}, u \sim U(0,1)$ is used as probability weights to sample non-null units. For methods requiring covariates as input, I separate scenarios where a strong covariates ($w(u)$) are available or scenarios that only use non-informative covariates.

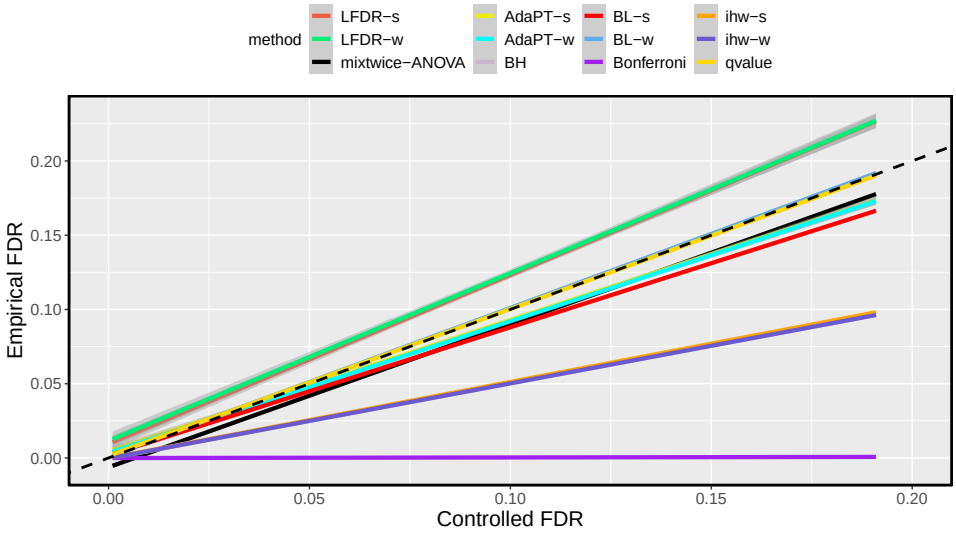


Figure 3.1: **Synthetic data and FDR control for variety of large-scale testing approaches:** Empirical FDRs (y-axes) versus controlled FDRs (x-axes) are shown for different approaches, including MixTwice-ANOVA (black), methods without covariates (Bonferroni, BH, q-value), methods with strong covariates (LFDR-s, AdaPT-s, BL-s and ihw-s) and methods with weak non-informative covariates (LFDR-w, AdaPT-w, BL-w and ihw-w). Reference line $y = x$ is in dotted.

The performance of FDR control is shown in Figure 3.1, which concludes that all methods, including `MixTwice-ANOVA` but excluding LFDR [Cai and Sun, 2009], control the false discovery rate. Some approaches, such as the Bonferroni correction, despite controlling the FDR, leave a larger *gap* between the nominated FDR and empirical FDR (lower than the identity line $y = x$). Considering that type one error and type two error are in conflict with each other, this *gap* might reduce the opportunity to achieve a satisfied power.

Though the majority of methods control the FDR no matter the level of signal size or variance, they perform differently for detecting true positives. The pro-

portion of true positives under a constant FDR level of 0.05 are plotted over the simulated null proportion $\pi_0$ under different levels of signal (row) and different levels of variance (column), as shown in panel A of Figure 3.2. Besides `MixTwice-ANOVA`, results of all other testing approaches are additionally averaged into three different groups: methods without covariates, methods with strong covariates and methods with weak/mis-specified covariates (panel B). Though strong covariates would certainly help in boosting power in many scenarios, `MixTwice-ANOVA` performs equally well compared with those with a strong covariates and even better under the *small signal, large variance* case.

## 3.4 High density peptide array data application

### 3.4.1 SSA+SjD vs SSA-SjD vs control: weak signal case

Sjogren's disease (SjD) is a systemic autoimmune disease with characteristic features of dry eye and dry mouth. In addition to characteristic dryness, SjD can affect many organ systems causing arthritis, interstitial lung disease, and increased risk of lymphoma, among other manifestations. Ultimately, SjD leads to a marked reduction in quality of life and healthcare costs more than twice those of healthy people [Callaghan et al., 2007, Lendrem et al., 2014]. Despite the impact of SjD, diagnosis is delayed more than two years [Huang et al., 2021].

B cells appear to play a major role in the pathogenesis of SjD based on the following features: i) germinal center-like reactions are present in SjD salivary glands that correlate with disease severity [He et al., 2017], ii) increased risk of B cell lymphoma [Masaki and Sugai, 2004], and iii) anti-SSA antibodies, which appear to be both diagnostic [Shiboski et al., 2017] and pathogenic given the presence of immune complexes of anti-SSA antibodies and necrotic cells that drive an interferon response[Båve et al., 2005]. Despite the importance of the anti-SSA antibody, up to 30% of SjD patients are anti-SSA antibody negative (SSA-) and require an invasive biopsy of the lip for diagnosis [Patel and Shahane, 2014]. SSA- SjD patients have a unique clinical phenotype including greater dryness, joint and nerve involvement [Brito Zerón et al., 2018, Park et al., 2019, Yazisiz et al., 2021, Relangi et al., 2021]. Novel biomarkers are a major unmet need in SjD and have the potential to provide novel pathogenic insights.

In order to identify novel SjD autoantibodies in SSA- and SSA+ SjD patients, my collaborators, Drs. Shelef and McCoy, designed a high density peptide array

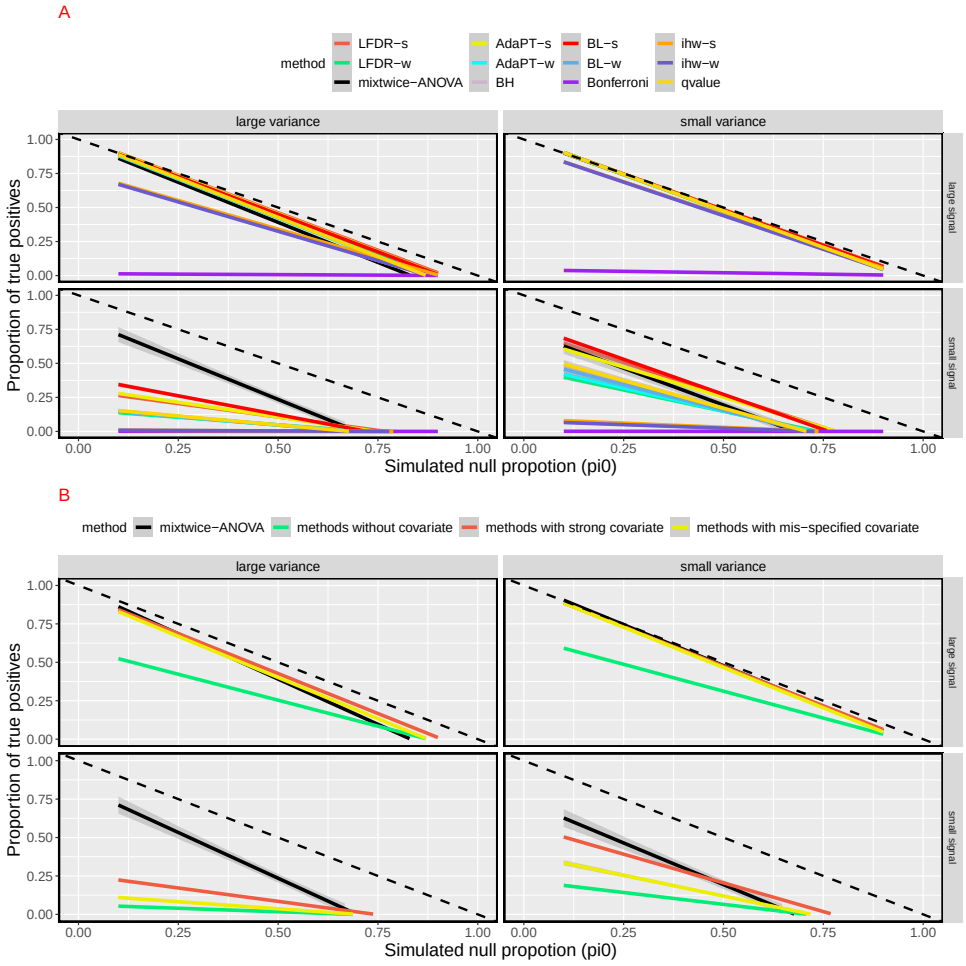


Figure 3.2: **Proportion of true positives versus simulated null proportions on synthetic data for a variety of large-scale testing approaches:** Panel A shows empirical proportion of true positives (y-axes) under a constant FDR control 0.05 vs simulated null proportions ($\pi_0$, x-axes) for different approaches, including MixTwice-ANOVA (black) under different levels of signal (row) and different levels of variance (column). Panel B specifically averages and plots the same metric for MixTwice-ANOVA (black), methods without covariates (green), methods with strong covariates (red) and methods with mis-specified covariates (yellow). Reference line $y = 1 - x$ is marked in dotted black.

experiment in which IgG binding to 172,828 peptides from 122 distinct human proteins that were selected based on possible importance for rheumatologic disease was quantified for 8 subjects with SSA+ SjD, 8 subjects with SSA- SjD and 8 control subjects whose clinical characteristics are described in Mergaert et al.

[2022]. I applied `MixTwice-ANOVA` and other large-scale testing tools (including BH, Storey's `qvalue`, `ihw`, LFDR, BL, AdaPT) on this data comparing these three groups. `MixTwice-ANOVA` is the only approach that could find significant peptides under the level of 0.05 over the comparison contrast even though `qvalue` reported the non-null proportion $\hat{\pi}_0$ is only 77 %. `MixTwice-ANOVA` identified 21 peptides bound belonging to 18 biologically relevant proteins, as reported in Table 3.1 with their local FDRs and basic statistics. Their F statistics were all ranked as 1% over peptides on the array (Figure 3.4). All the identified peptides followed a similar pattern; SSA- subjects had lower binding than controls and SSA+ subjects had higher binding than controls. We also find the low binding structure for SSA- subjects compared to control subjects in the peptide population; i.e., the median z-score comparing SSA- SjD and control for all peptides on the array is negative.

Table 3.1: Significant peptides by MixTwice-ANOVA comparing SSA+SjD vs SSA-SjD vs control subjects

| | Protein | Peptide sequence | Position | locFDR | Fstat | SSB | SSE |
|---|---|---|---|---|---|---|---|
| 1 | P11940 | PSQIAQLRPSPR | 425 | 0.010 | 12.60 | 0.04 | 0.04 |
| 2 | P08603 | SFTMIGHRSITC | 722 | 0.000 | 12.01 | 0.04 | 0.03 |
| 3 | P04114 | IKSPAFTDLHLR | 3978 | 0.016 | 11.72 | 0.04 | 0.04 |
| 4 | P02787 | SAHGFLKVPPRM | 317 | 0.003 | 11.41 | 0.02 | 0.02 |
| 5 | P36980 | RAMCQNGJLVYP | 254 | 0.021 | 11.26 | 0.04 | 0.04 |
| 6 | Q14739 | HKNTQEKFSLSQ | 146 | 0.001 | 10.54 | 0.03 | 0.03 |
| 7 | P00450 | EDRVKWYLFGMG | 278 | 0.000 | 9.76 | 0.03 | 0.03 |
| 8 | P10909 | BBPHFFFPKSBI | 214 | 0.009 | 9.64 | 0.02 | 0.02 |
| 9 | Q03591 | TAKQKLYLRTGE | 283 | 0.004 | 9.46 | 0.03 | 0.03 |
| 10 | P07305 | NADSQIKLSIKR | 63 | 0.006 | 9.14 | 0.03 | 0.03 |
| 11 | P01008 | NPMCIYRSPEKK | 50 | 0.005 | 9.05 | 0.02 | 0.02 |
| 12 | P35579 | LKERYYSGLIYT | 101 | 0.004 | 8.84 | 0.03 | 0.03 |
| 13 | P04114 | VSTAFVYTJNPN | 3704 | 0.013 | 8.49 | 0.03 | 0.03 |
| 14 | P00751 | NLFQVLPWLKEK | 744 | 0.022 | 8.46 | 0.02 | 0.02 |
| 15 | Q9BXR6 | FSCRKNLIRVGS | 173 | 0.010 | 7.83 | 0.02 | 0.03 |
| 16 | Q92496 | HGGLYYKSLRRL | 31 | 0.006 | 7.75 | 0.02 | 0.03 |
| 17 | P08603 | RCIRVJTCSJSS | 441 | 0.038 | 7.71 | 0.02 | 0.03 |
| 18 | Q16778 | RSRKESYSIYVY | 32 | 0.010 | 7.56 | 0.02 | 0.03 |
| 19 | P04114 | AYLMLMBSPSQA | 561 | 0.009 | 7.48 | 0.02 | 0.03 |
| 20 | P02751 | LTNFLVRYSPVK | 1295 | 0.010 | 7.26 | 0.02 | 0.03 |
| 21 | P01861 | FLYSRLTVDKSR | 285 | 0.011 | 7.24 | 0.02 | 0.03 |

We analyzed those selected peptides using a variety of techniques, including protein cluster analysis and peptide ELISA validation. From the protein cluster analysis, we identified complement regulation as a top theme. DAVID gene ontology (GO, [Huang et al., 2009a,b]) analysis yielded the protein cluster with the highest enrichment score is extracellular comprising complement. Of the identified

18 bound proteins, 6 (33%) were related to complement: P08603 (complement factor h), Q03591 (complement factor H-related 1), P36980 (complement factor H-related 2), Q92496 (complement factor H-related 4), P00751 (complement factor B), and P10909 (clusterin). Low complement is a feature of SjD associated with increased risk of mortality [Singh et al., 2016]. Complement factor B is upregulated in conjunctival cells of SjD patients and negatively correlates with tear breakup time [de Paiva et al., 2021]. Furthermore, complement factor B in cerebrospinal fluid discriminates between SjD subjects with and without fatigue [Larssen et al., 2019]. Complement factor H is pivotal to modulating complement activation and has been implicated in the pathogenesis of SjD. Complement factor H is reduced in SjD mouse model saliva [Li et al., 2021]. Furthermore, complement factor H along with clusterin were lower in SjD patients with neuromyelitis optica spectrum disease (NMOSD) than without NMOSD[Qiao et al., 2019]. Clusterin, controlling terminal complement-related damage, is also upregulated in SjD saliva compared to control[Sembler-Møller et al., 2020]. As a further negative control calculation, we performed the same protein cluster analysis using all proteins on the array (122 distinct human proteins) and GO does not yield the complement cluster as the top theme. Other interesting proteins that demonstrated reduced binding among SSA- SjD subjects and increased binding among SSA+ SjD subjects are included in Table 3.1.

Moreover, 2 peptides with the highest F statistic (PSQIAQLRPSPR and SFTMIGHRSITC, top 2 on Table 3.1) were validated by enzyme-linked immunosorbent assay (ELISA), a commonly used validation tool for high density peptide array. The former is part of P11940 (poly A binding protein cytoplasmic 1), an RNA-binding protein not yet described in SjD specifically but is implicated by its RNA binding as a general SjD pathogenic process. The latter is P08603, complement factor H, a component of the alternative complement cascade system. P-values for those two peptides comparing three groups on the ELISA signal intensities are all smaller than 0.1. Boxplots for those two peptides on array signal intensity and ELISA intensity are summarized in Figure 3.3.

As noted above, the peptide array has 172,828 peptides from 122 distinct human proteins that were selected based on possible importance for rheumatologic disease. As a follow-up analysis, we utilized a larger array with the same subject samples but over 5.3 million peptides from the whole human proteome. `MixTwice-ANOVA` identified 881 peptides under FDR level 0.05 and the two peptides noted above (PSQIAQLRPSPR and SFTMIGHRSITC) are also on the discovery list.

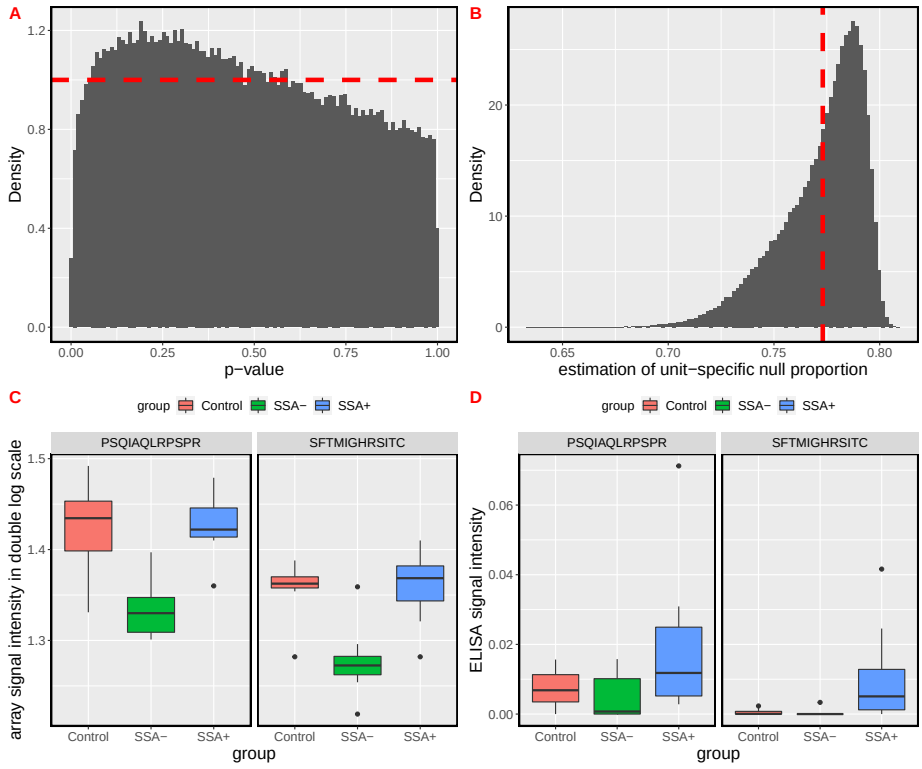


Figure 3.3: **One-way ANOVA analysis for Sjogren disease:** Panel A shows the histogram of p-values comparing SSA+ Sjogren (8) vs SSA- Sjogren (8) vs control (8). Panel B shows the histogram of estimated unit-specific null proportion ($\hat{\pi}_{0,i}$) estimated using Boca and Leek and the red dashed line highlighted the $\hat{\pi}_0$ using q-value. Boxplots of array signal intensity and ELISA intensity for two peptides selected by MixTwice-ANOVA are summarized in Panel C and Panel D.

### 3.4.2 CCP+RF+ RA vs CCP-RF- RA vs control: strong signal case

A reproducibility evaluation for `MixTwice` using a peptide array strong signal case, comparing signal intensity between RA patients who test positive for the anti-cyclic citrullinated peptide (CCP) antibody and rheumatoid factor (RF) biomarkers (CCP+RF+ RA) and control subjects in two independent studies was discussed in Zheng et al. [2021]. A similar reproducibility analysis could be performed for multiple-group comparison. In a first high density peptide array experiment[Zheng et al., 2020], signal intensity was measured for IgG binding to 172,828 peptides for 12 CCP+RF+ RA, 12 CCP-RF- RA, and 12 control subjects. A three-group comparison (CCP+RF+ RA vs CCP-RF- RA vs control) was also performed in

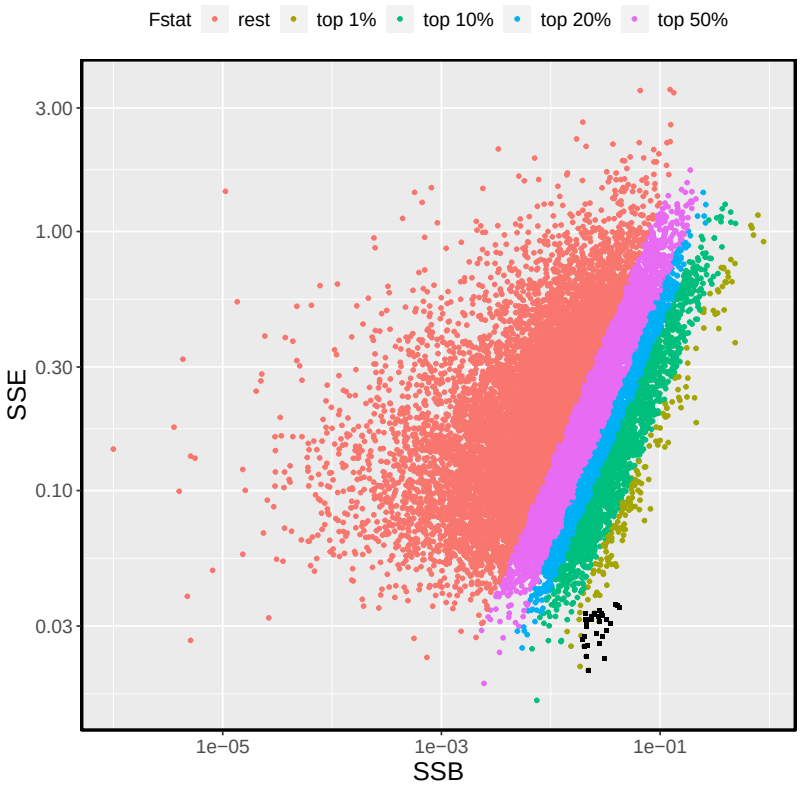


Figure 3.4: **Scatter plot of SSB vs SSE comparing three groups in SjD example:** Scatter plot shows Sum of Squares in Between (SSB, x-axes) and Sum of Squares in Error (SSE, y-axes) of all 172,828 peptides on the array, with those 21 peptides identified by MixTwice-ANOVA coded in black. Ranking of F statistics for peptides is coded using different colors.

a new experiment with the same peptide population but serum samples from different human subjects (8 in each group) described in detail in [Mergaert et al., 2022]. Figure 3.5 compares two reproducibility metrics between `MixTwice-ANOVA` and other tools. `MixTwice-ANOVA` shows better reproducibility compared to other testing methods in this example, which is consistent with the conclusion reported in Zheng et al. [2021] for a two-sample comparison. More data visualization and the results of a mixing distribution estimation are summarized in Supplementary material B.3. This reproducibility, as an additional metric to FDR control, power, applicability, etc., to compare large-scale testing tools [Korthauer et al., 2019], is discussed in the following chapter.

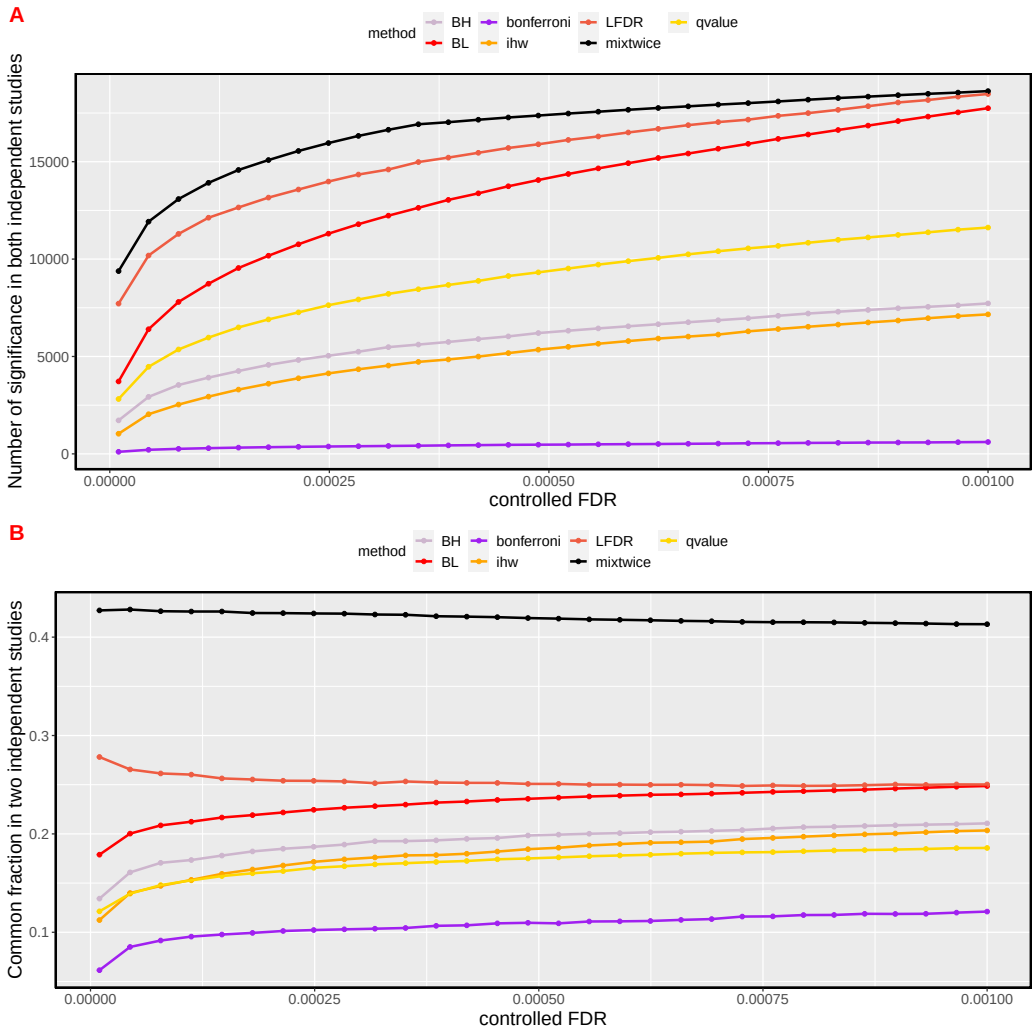


Figure 3.5: **Reproducibility comparison of CCP+RF+ RA vs CCP-RF-RA vs control:** two metrics compare MixTwice-ANOVA (black) and other testing methods as a function of FDR threshold.

# Chapter 4

# Numerical experiments and computational aspects of MixTwice

## 4.1 Introduction

Initial calculations reported in Zheng et al. [2021] and its follow-up generalization `MixTwice-ANOVA` show good performance in several case studies and in a range of simulations. For example, `MixTwice` estimates mixing distributions accurately and controls the FDR under a variety of alternative distribution settings. Also, it increases the power on the high density peptide array data where the signal is weak and increases the reproducibility where the signal is strong.

More extensive numerical experiments would help to establish the overall operating characteristics of `MixTWice` and to convey when and how the second mixing layer improves performance. I am guided by the recent review [Korthauer et al., 2019] in which a variety of large-scale testing tools was deployed on a large battery of numerical experiments, both using synthetic data and benchmark data sets. This review compared methods using different operating characteristics: FDR control, power, applicability, consistency and usability. In the first part of this chapter, I expand this review in order to include `MixTwice` and two-step `ASH` and also to examine reproducibility properties of the various methods. Besides those 5 metrics discussed in Korthauer et al. [2019], I introduce a reproducibility metric to measure the similarity between discovery lists from the analysis of replicate data sets.

Another innovation related to large-scale testing concerns computational algo-

rithms for constrained optimization. Nonparametric maximum likelihood estimation of the mixing distribution is essential in local false discovery rate calculations. The optimization problem in `MixTwice` and `MixTwice-ANOVA` can be solved through gradient-based optimization (i.e., Augmented Lagrangian method, `AugLag`). This is effective but sometimes computationally expensive, especially with large numbers of testing units. Motivated by this, in the second half of this chapter, I propose an alternative algorithm – `EM-PAVA`, – that combines the pool adjacent violator algorithm (PAVA) and the Expectation-Maximization (EM) algorithm in order to improve computational efficiency.

## 4.2 Generalized evaluation and reproducibility of large-scale testing tools

No single testing tool has proven to be uniformly superior to others on important operating metrics. Empirical-Bayes tools like `ASH` and `MixTwice` have improved power properties by using more unit-level data. However, they may be less applicable, since some analyses (such as single-cell RNA-seq data, ChIP-seq data, etc.) may not provide both effect sizes and standard errors. Korthauer et al. [2019] provides a practical guide to popular testing tools. The practical guide includes large-scale testing tools without the input of informative covariates such as BH [Benjamini and Hochberg, 1995], `qvalue` [Storey, 2002], `ASH` [Stephens, 2017] and methods with independent and informative covariates such as `IHW` [Ignatiadis et al., 2016], Boca and Leek method [Boca and Leek, 2018], Adaptive p-value thresholding method AdaPT [Lei and Fithian, 2018], Cai's conditional FDR method LFDR [Cai and Sun, 2009] and FDR regression [Scott et al., 2015]. Those methods assume basically the independent and informative covariates are correlated with the probability of testing units being null. For example, the Boca and Leek method uses the covariate to predict the unit-specific null proportion $\pi_{0,i}$ rather than the global null proportion $\pi_0$. Using a battery of synthetic and benchmark data sets, that guide compares methods from various angles. Besides the most popular evaluation metric FDR control and power, it also evaluates applicability, consistency and usability which evaluates *Can this method be applicable to most of the data sets?*, *Will the method provide similar result in different types of bioinformatics data sets?* and *Is there a convenient way to implement this method, such as R package?*, respectively.

Figure 4.1 summarizes an extension of the Korthauer study. Utilizing the same

collection of synthetic and benchmark data sets, I include `MixTwice` and two-step `ASH` to the comparison list. I also evaluate all methods in terms of reproducibility.

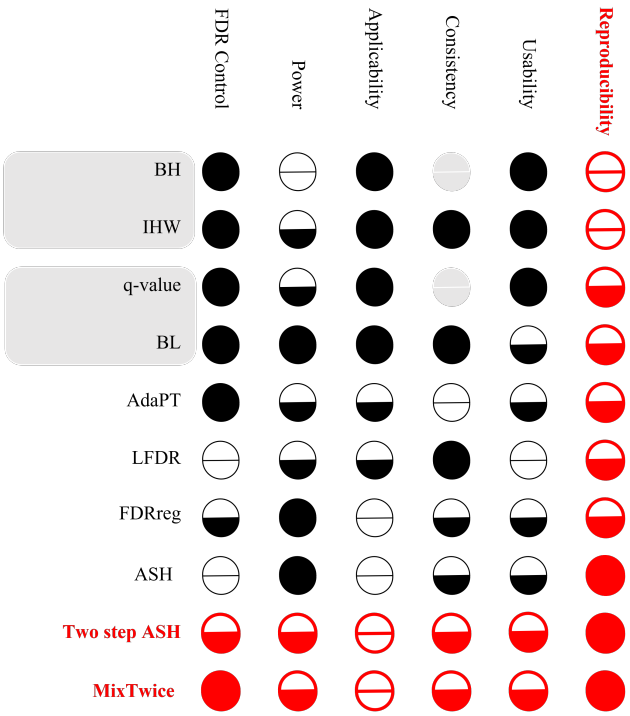


Figure 4.1: **Summary of recommendations comparing large-scale hypothesis testing tools:** A generalization figure of figure 6 from Korthauer et al. [2019] with (highlight in red) additional column evaluating the reproducibility of methods and two additional rows summarizing the performance of MixTwice [Zheng et al., 2021] and two-step ASH [Lu and Stephens, 2019]. Evaluation symbols are consistent with the definition in Korthauer et al. [2019].

I propose a reproducibility metric to measure the similarity between discovery lists from the analysis of replicate data sets. The evaluation of reproducibility, demonstrated in Figure 4.1, is based on the ranking of all those methods (top 30 % for a full score and top 80 % for a partial credit) in a yeast *in silico* experiment [Gierliński et al., 2015], as shown in Figure 4.2.

In the yeast *in silico* experiment [Gierliński et al., 2015], all samples that passed quality control are included. All genes with a mean count of at least 1 across all samples are included, for a total of 6553 genes. The full data compares genes on two conditions (WT vs Snf2-knockoff) with 48 replicates on each condition. This is a case where the signal is strong; under the 0.1 FDR, BH finds approximately 65% positives.

Though there are not two independent studies, as reported in Zheng et al. [2021], I evaluate the reproducibility on yeast *in silico* experiment by constructing pairs of *pseudo-independent* studies where each study of the pair compares randomly-selected 15 vs 15 samples. This procedure is repeated 50 times, and reproducibility evaluation metrics are calculated and then averaged using each pair of pseudo-independent studies.

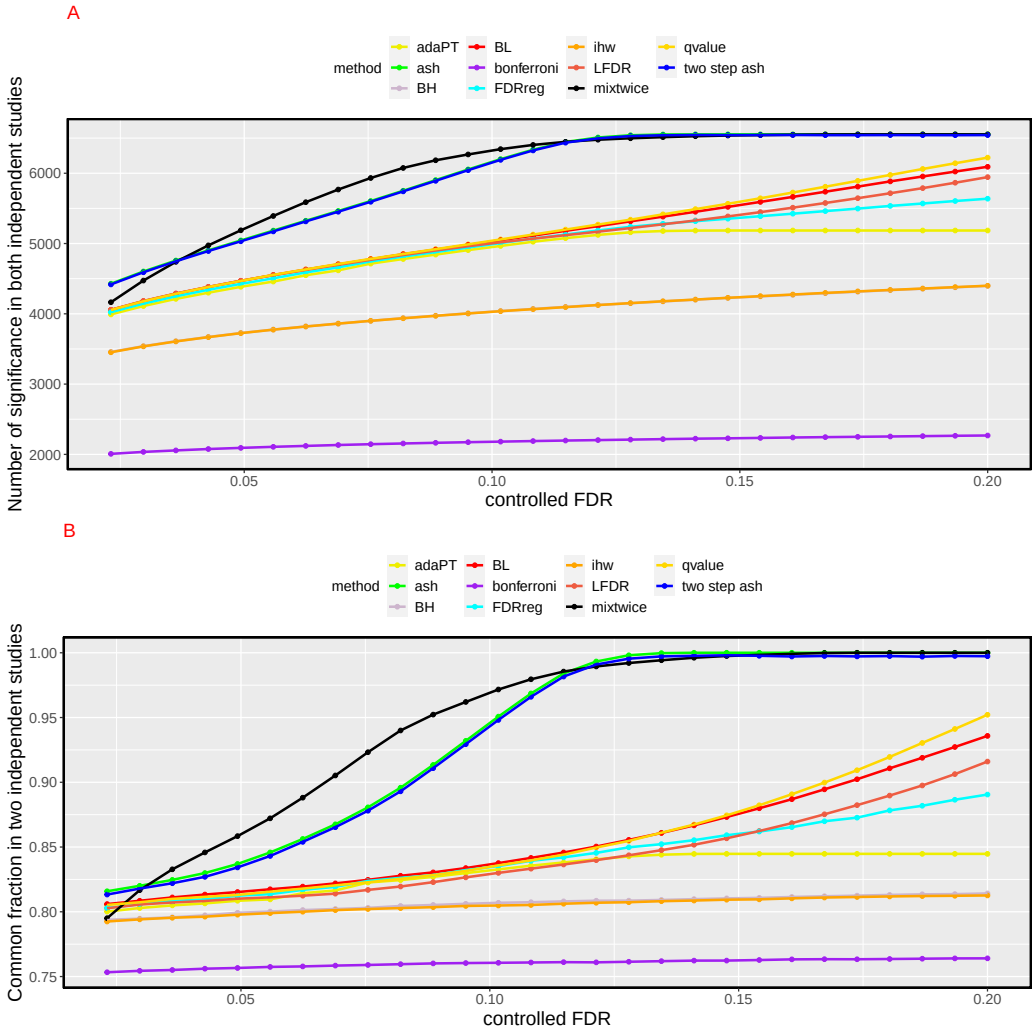


Figure 4.2: **Two reproducibility metrics among all testing approaches on 50 times of paired pseudo-independent studies of yeast in silico experiment:** Panel A shows the number of discoveries in both independent studies and panel B shows the common fraction in both independent studies. Metrics are evaluated under defined level of FDR control from 0 to 0.2 and averaged among the 50 times of simulations.

Something interesting to note is the ranking of reproducibility is almost consistent with the ranking of power, which is reversed from the ranking of applicability. On the one hand, methods with the most data reduction are those that only require the p-values (besides covariates) to make testing multiplicity adjustment. Those methods, like `BH` and `qvalue`, are applicable to almost all data sets as long as the collections of p-values are available. On the other hand, the power as well as reproducibility evaluation prefers more information (i.e., less reduction) from the data. This is further discussed through a toy example in Supplementary material C.1.

As noted above, there is no uniform superior for any testing tools in all evaluation metrics thus the evaluation table Figure 4.1 might provide additional recommendations in choosing the most appropriate testing tools in different bioinformatics data analysis. For example, those methods with high applicability might be the only choices for data sets where the p-values are the only available statistics. When the independent and informative covariates are possible, methods incorporating them might have improved power properties. When the less-reduced data such as effect sizes and standard errors are available, methods such as two-step `ASH` and `MixTwice` would give higher power and better reproducibility.

## 4.3 Computation of MixTwice: nonparametric MLE with shape constraint

### 4.3.1 Optimization problem

Recall the optimization problem we need to solve in `MixTwice` and in `MixTwice-ANOVA`. In `MixTwice`, we solve the constrained optimization:

$$\begin{aligned} \min_{g,h} -l(g,h) &= -\sum_{i=1}^{m} \log p(x_i, s_i^2|g,h) \qquad (4.1) \\ \text{Subject to:} \quad & g_k, h_l \geq 0 \quad \forall k,l \\ & \sum_k g_k = \sum_l h_l = 1 \\ & g_k \leq g_{k+1}, \quad k \in \{-K, -K+1, ..., -1\} \\ & g_k \geq g_{k+1}, \quad k \in \{0, 1, ..., K\}. \end{aligned}$$

In `MixTwice-ANOVA`, we solve the constrained optimization:

$$\begin{aligned} \min_{g,h} -l(g,h) &= -\sum_{i=1}^{m} \log p(\mathrm{SSB}_i, \mathrm{SSE}_i|g,h) \qquad (4.2) \\ \text{Subject to:} \quad & g_k, h_l \geq 0 \quad \forall k,l \\ & \sum_k g_k = \sum_l h_l = 1 \\ & g_0 \leq g_1 \leq \cdots \leq g_K. \end{aligned}$$

This optimization problem could be solved directly using the Augmented Lagrangian approach (with option `method = "AugLag"` in the `MixTwice` package). However, when the number of testing units and the number of support points are large, Augmented Lagrangian approach could be less efficient, especially with numerous constraints for the unimodality or monotonicity restriction. The next section describes an alternative algorithm, called `EM-PAVA`, that could solve the optimization problem more efficiently.

### 4.3.2 EM-PAVA algorithm

Without the shape constraint, the problem could be solved efficiently using Expectation-Maximization (EM) approach. Within each iteration, the expectation step calculates the posterior probability at a given grid ($\tilde{g}_k = \mathbb{E}\left(\mathbb{P}(\theta = a_k|x, s^2)\right)$ in `MixTwice` or

similarly, $\tilde{g}_k = \mathbb{E}\left(\mathbb{P}(\lambda = a_k|\text{SSB}, \text{SSE})\right)$ in `MixTwice-ANOVA`) and the maximization step solves the optimization problem by minimizing the negative log loss:

$$\begin{aligned} \min_g -l(g) &= -\sum_k \tilde{g}_k \log g_k \\ \text{Subject to:} \quad & g_k \geq 0 \quad \forall k \\ & \sum_k g_k = 1. \end{aligned}$$

It is easier to verify with the given property that $\sum_k \tilde{g}_k = 1$, the optimizer of the problem just takes $g_k^* = \tilde{g}_k$ at each iteration step. The shape constraint does not change the E-step while it adds another constraint when we try to minimize the negative log loss in M-step. For example, in the `MixTwice-ANOVA` problem, the M-step is:

$$\begin{aligned} \min_g -l(g) &= -\sum_k \tilde{g}_k \log g_k \qquad (4.3) \\ \text{Subject to:} \quad & g_k \geq 0 \quad \forall k \\ & \sum_k g_k = 1 \\ & g_0 \leq g_1 \leq \cdots \leq g_K. \qquad (4.4) \end{aligned}$$

We refer the contribution of pool adjacent violator algorithm (PAVA) to solve this problem [Ayer et al., 1955, Robertson et al., 1988]. It is useful to review some important definitions, problems, algorithms and theorems related to the shape constraint, as present in the book Robertson et al. [1988] and Turner [2020b].

**Definition 1 (Simple order)** *A binary relation $\prec$ on $\mathcal{X}$ is a simple order on $\mathcal{X}$:*

1. *it is reflexive: $x \prec x$ for $x \in \mathcal{X}$;*
2. *it is transitive: $x, y, z \in \mathcal{X}$, $x \prec y$ and $y \prec z$ imply $x \prec z$;*
3. *it is anti-symmetric: $x, y \in \mathcal{X}$, $x \prec y$ and $y \prec x$ imply $x = y$;*
4. *every two elements of $\mathcal{X}$ are comparable: $x, y \in \mathcal{X}$ implies that either $x \prec y$ or $y \prec x$.*

*Furthermore, a binary relation $\prec$ on $\mathcal{X}$ is a partial order if it is reflexive, transitive and anti-symmetric, but there may be non-comparable elements. A quasi-order is reflexive and transitive. It need not to be anti-symmetric, and it may admit non-comparable elements. Every simple order is a partial order and every partial order is a quasi-order.*

**Definition 2 (Isotonic)** *A real valued function, $g$, on $\mathcal{X}$ is isotonic with respect to the quasi-ordering $\prec$ on $\mathcal{X}$ if $x, y \in \mathcal{X}$ and $x \prec y$ imply $g(x) \leq g(y)$*

In our context of solving optimization 4.3, we want to estimate the isotonic function $g$ where $\mathcal{X}$ is just the support $\mathcal{X} = \{1, 2, \cdots, K\}$ and our estimator is $g_k = g(k)$. The monotone non-decreasing constraint restricted a function with respect to a simple order however the unimodal constraint with a given mode is only a quasi-order.

**Definition 3 (Isotonic regression)** *Let $g$ be a given function on $\mathcal{X}$ and $w$ a given positive function on $\mathcal{X}$. An isotonic function $g^*$ on $\mathcal{X}$ is an isotonic regression of $g$ with weights $w$ if and only if*

$$\sum_{x \in \mathcal{X}} [g(x) - g^*(x)]^2 w(x) \leq \sum_{x \in \mathcal{X}} [g(x) - f(x)]^2 w(x)$$

*for all functions $f$ on $\mathcal{X}$ which are isotonic.*

This defines the problem of isotonic regression. We can also view this problem, closely related to our question, as solving the optimization problem to find $g_k^* = g^*(k)$ given input of $\tilde{g} = g(k)$ but on a squared error loss rather than a log loss. We will discuss in the next a few paragraphs about their connection.

**Theorem 1 (Optimality condition of constraint optimization)** *Suppose $\mathscr{C}$ is any convex set of functions on $\mathcal{X}$ and $g$ and $w$ are given functions on $\mathcal{X}$ with $w(x) > 0$ for all $x \in \mathcal{X}$. If $g^* \in \mathscr{C}$ and $g^*$ solves*

$$\min \sum_{x \in \mathcal{X}} [g(x) - g^*(x)]^2 w(x)$$

*subject to $f \in \mathscr{C}$, then for every $f \in \mathscr{C}$,*

$$\sum_{x \in \mathcal{X}} [g(x) - g^*(x)][g^*(x) - f(x)] w(x) \geq 0$$

*and*

$$\sum_{x \in \mathcal{X}} [g(x) - f(x)]^2 w(x) \geq \sum_{x \in \mathcal{X}} [g(x) - g^*(x)]^2 w(x) + \sum_{x \in \mathcal{X}} [f(x) - g^*(x)]^2 w(x)$$

*Conversely, if $u \in \mathscr{C}$ and*

$$\sum_{x\in\mathcal{X}}[g(x)-u(x)][u(x)-f(x)]w(x)\geq 0$$

*for all $f\in\mathscr{C}$ then $u$ solves the problem. There is at most one such function.*

With that in mind, we will then separately discuss computational algorithms and connections among:

1. Simply-ordered isotonic regression;
2. Quasi-ordered isotonic regression and isotonic regression with unimodal constraint with fixed mode;
3. Connections between isotonic regression and the optimization problem with negative log loss objective.

**Simply-ordered isotonic regression**

Pool adjacent violator algorithm (PAVA) solves the simply-ordered isotonic regression efficiently.

**Definition 4 (CSD and GCM)** *Plot the points $P_j=(W_j,G_j); j=0,1,\cdots,K$ with $W_j=\sum_i w(x_i)$ and $G_j=\sum_i g(x_i)w(x_i)$ and $P_0=(0,0)$. The plot of these points is called cumulative sum diagram (CSD) for the given function $g$ with weights $w$. Let $G^*$ be the greatest convex minorant (GCM) of the CSD on the interval $[0,W_K]$. The value, $G^*(t)$, is the supremum of the values, at $t$, for all convex functions which lie entirely below the CSD. Let $g^*(x_i)$ be the left derivative of $G^*$ at $W_i$ for $i=1,2,\cdots,K$.*

**Theorem 2 ($g^*$ solves the simply-ordered isotonic regression)** *If $\mathcal{X}$ is simply ordered, the left derivative or left-hand slope, $g^*$, of the GCM furnishes the isotonic regression of $g$.*

The proof of this theorem is by verifying

$$\sum_{x\in\mathcal{X}}[g(x)-g^*(x)]^2w(x)\geq\sum_{x\in\mathcal{X}}[g(x)-g^*(x)]^2w(x)+\sum_{x\in\mathcal{X}}[f(x)-g^*(x)]^2w(x)$$

for $f$ simply-ordered isotonic.

On the one hand, this theorem states the theoretical guarantees of solving the simply-ordered isotonic regression. On the other hand, the pool adjacent algorithm

(PAVA), firstly published by Ayer et al. [1955], provides an efficient way of finding the GCM. We can also prove that the solution of PAVA is optimal by verifying the KKT condition.

### Quasi-ordered isotonic regression and unimodality

The algorithm for the quasi-ordered isotonic regression is much complicated compared to algorithm solving simply-ordered isotonic regression. However, there are still some general results.

**Theorem 3** *If $\mathscr{H}$ is a convex cone of functions on $\mathcal{X}$ and $g$ and $w$ are given functions on $\mathcal{X}$ with $w > 0$ then a function $g^*$ on $\mathcal{X}$ solves the isotonic regression with respect any quasi-order if and only if $g^* \in \mathscr{H}$ and*

$$\sum_{x\in\mathcal{X}}[g(x) - g^*(x)]g^*(x)w(x) = 0$$

*and*

$$\sum_{x\in\mathcal{X}}[g(x) - g^*(x)]f(x)w(x) \leq 0, \forall f \in \mathscr{H}$$

The proof of this theorem is simply followed by Theorem 1. Followed by this algorithm, it is immediately to get if $\mathscr{H}$ is any convex cone of functions on $\mathcal{X}$ and if $\mathscr{H}$ contains all the constant functions on $\mathcal{X}$ and if $g^*$ solves the isotonic regression problem, then $\sum_{x\in\mathcal{X}} g(x)w(x) = \sum_{x\in\mathcal{X}} g^*(x)w(x)$.

**Definition 5 (Level set)** *Suppose $g$ and $w$ are functions defined on $\mathcal{X}$, set*

$$\mathrm{Av}(\mathcal{A}) := \frac{\sum_{x\in A} w(x)g(x)}{\sum_{x\in\mathcal{A}} w(x)}$$

*for those $\mathcal{A}$ which are nonempty subsets of $\mathcal{X}$ and let $[g^* = c]$ denote $\{x \in \mathcal{X} : g^*(x) = c\}$.*

**Theorem 4** *If $c$ is any real number and if the set $[g^* = c]$ is nonempty then $c = \mathrm{Av}([g^* = c])$.*

This theorem can be proven by contradiction and by observing

$$\sum_{x\in\mathcal{X}}[g(x) - g^*(x)]^2w(x) = \sum_{[g^*\neq c]}[g(x) - g^*(x)]^2w(x) + \sum_{[g^*=c]}[g(x) - c]^2w(x)$$

Also, this is one of the most critical theorems as it reduces the problem of isotonic regression (solving for $g^*$) to the problem of finding the sets on which $g^*$ is constant (i.e., level sets). For the simple order the sides of GCM determine the level sets using PAVA. For the quasi-order, it might be a little more complicated. The efficient algorithm is not guaranteed to be known in general for any quasi-order isotonic regression, but it is well-developed for unimodal constraint with a fixed mode.

The following result is summarized and reproduced from Turner [2020b]. For $\mathcal{X} = \{-K, \cdots, 0, \cdots, K\}$, the unimodal constraint with a fixed mode at 0 is a quasi-order (even a partial order but not simple order) where we define the binary relation $x \prec y$ as $x \leq y \leq 0$ or $x \geq y \geq 0$. If $x \leq 0$ and $y \geq 0$ or vice versa then $x, y$ are not comparable hence the relation is not simply-ordered.

Define $\mathcal{X}_1$ and $\mathcal{X}_2$ to be $\mathcal{X}_1 := \{k \in \mathcal{X} | k \neq 0\}$ and $\mathcal{X}_2 := \{0\}$. Let $g_1$ be the restriction of $g$ to $\mathcal{X}_1$ and let $g_1^*$ be the isotonic regression of $g_1$. A direct but important corollary result of Theorem 4 is the following:

**Corollary 1** *The isotonic regression with fixed-mode unimodal isotonic regression $g^*$ takes the form:*

$$g^*(x) = c_i \quad \text{on} \quad \mathcal{L}_i, \quad i = 1, 2, \cdots, r$$

.

where the collection of $\mathcal{L}_i$ form a disjoint and exhaustive collection of subsets of $\mathcal{X}$, and $c_1 < c_2 \cdots < c_r$. Moreover, $c_i = \text{Av}(\mathcal{L}_i) := \frac{\sum_{x \in \mathcal{L}_i} w(x) g(x)}{\sum_{x \in \mathcal{L}_i} w(x)}$ is the level sets defined previously.

Let the level sets and level values for $g_1^*$ be $\mathcal{L}_i$ and $c_1 < c_2 < \cdots < c_r$ and let $\mathcal{L}_{r+1} = 0$ and $c_{r+1} = g(0)$ (in our problem, $\tilde{g}_0$). Define the function $f$ on $\{1, \cdots, r+1\}$ by $f(t) = c_t$ for $t = 1, 2, \cdots, r+1$ and a weight function $u(t) = \sum_{x \in \mathcal{L}_t} w(x)$. The computational algorithm and the main theorem is as follows.

**Theorem 5** *Let $f^*$ be the isotonic regression of $f$ with respect to the simple order $1, 2, \cdots, r+1$ and weight function $u$. Then the isotonic regression $g^*$ is:*

$$g^*(x) = f^*(t), \quad \text{for} \quad s \in \mathcal{L}_t$$

The proof of this theorem is heavily based on the result of Theorem 1.

**Isotonic regression with log loss objective**

PAVA solves the isotonic regression in the context of a simple order, which is related to what we need to solve in `MixTwice-ANOVA`. The algorithm in Turner [2020b] solves the isotonic regression with unimodal constraint for a known mode, one special case of quasi-order, which is related to what we need to solve in `MixTwice`. There is a final gap between the isotonic regression and our optimization problem: our optimization minimizes the negative log loss while the isotonic regression minimizes the weighted squared error loss. However, there is a direct mitigation proposed in Robertson et al. [1988] between the log loss and the weighted square error loss, using the following theorems. And this mitigation is valid for all quasi-orders.

Recall the problem is to maximize $\sum_{x\in\mathcal{X}} \tilde{g}(x) \log g(x)$ subject to $g(x)$ being quasi-order on $\mathcal{X}$ and $\sum_x \tilde{g}(x) = \sum_x g(x) = 1$. $\mathcal{X}$ is $1, 2, \cdots, K$ for the `MixTwice-ANOVA` case and $\{-K, \cdots, 0, \cdots, K\}$ for the `MixTwice` case. $\tilde{g}(x) = g_k$ on support $k$ is the observed expected probability after the E-step of each iteration and $g^*(x) = g_k^*$ is the optimizer after the M-step of each iteration.

**Theorem 6** *For an arbitrary real valued function, $\Psi$, defined on reals,*

$$\sum_{x\in\mathcal{X}} [g(x) - g^*(x)]\Psi[g^*(x)]w(x) = 0$$

.

This is a direct application of Theorem 4.

**Theorem 7** *Suppose $\Phi$ is a convex function which is finite on an interval $I$ containing the range of function $g$ on $\mathcal{X}$ and define $\Delta_\Phi(u, v) = \Phi(u) - \Phi(v) - (u - v)\phi(v)$ where $\phi$ is the left derivative of $\Phi$. If $f$ is isotonic on $\mathcal{X}$ then*

$$\sum_{x\in\mathcal{X}} \Delta_\Phi[g(x), f(x)]w(x) \geq \sum_{x\in\mathcal{X}} \Delta_\Phi[g(x), g^*(x)]w(x) + \sum_{x\in\mathcal{X}} \Delta_\Phi[g^*(x), f(x)]w(x)$$

*Consequently, $g^*$ minimizes*

$$\sum_{x\in\mathcal{X}} \Phi_\Delta[g(x), f(x)]w(x)$$

*in the class of all isotonic $f$.*

These two theorems, Theorem 6 and Theorem 7, together provide a direct conclusion mitigating the isotonic regression with our optimization problem. Specifically, take $\Phi(u) = u \log u$ in Theorem 7 and hence $\Delta_{\Phi(g,\tilde{g})} = \tilde{g} \log \tilde{g} - \tilde{g} \log g - (\tilde{g} - g)$. Therefore, the $g^*$, the isotonic regression of $g$, maximizes

$$\sum_{x \in \mathcal{X}} (\tilde{g}(x) \log g(x) + \tilde{g}(x) - g(x))$$

However, the latter part, $\sum_{x \in \mathcal{X}} (\tilde{g}(x) - g(x)) = 0$ by taking a naive $\Psi$ from Theorem 6. The desired result follows immediately.

### 4.3.3 Empirical performance of AugLag and EM-PAVA

The implementation of both options to solve the constrained optimization problem is available in package `MixTwice` [Zheng and Newton, 2022] where we refer package `alabama` [Varadhan, 2015] to solve the problem directly using `method = "AugLag"` and refer package `Iso` to solve the problem using `method = "EM-PAVA"` [Turner, 2020a]. The `Iso` package provides two helpful functions, `PAVA` and `ufit`, for isotonic regression with monotonic constraint or unimodal constraint, which would be used in `MixTwice-ANOVA` and `MixTwice`, respectively.

Here I illustrate one simulation example to compare those two computational options. From the following Figure 4.3, `EM-PAVA` approach would converge to the direct `AugLag` optimization approach, in both the pointmass estimator and the objective function quantity (negative log likelihood). For implementation, I define the error of convergence by the $L_2$ norm between iterations, $e_t := ||\boldsymbol{g}^t - \boldsymbol{g}^{t+1}||_2^2$ and stop the iteration when $e_t$ is smaller than $10^{-6}$. `EM-PAVA` is much faster (around 30 times) compared to `AugLag`, as reported in the figure. Furthermore, the number of iterations required to achieve convergence decreases with the number of testing units. Though the increasing number of units may introduce higher computational complexity [1] in other steps (e.g., matrix multiplication), the lower number of iterations to achieve convergence, as an additional bonus, makes it more applicable to be implemented for more testing units.

[1] time complexity is CPU time evaluated with Inter® Core™ i5-7400HQ CPU processor

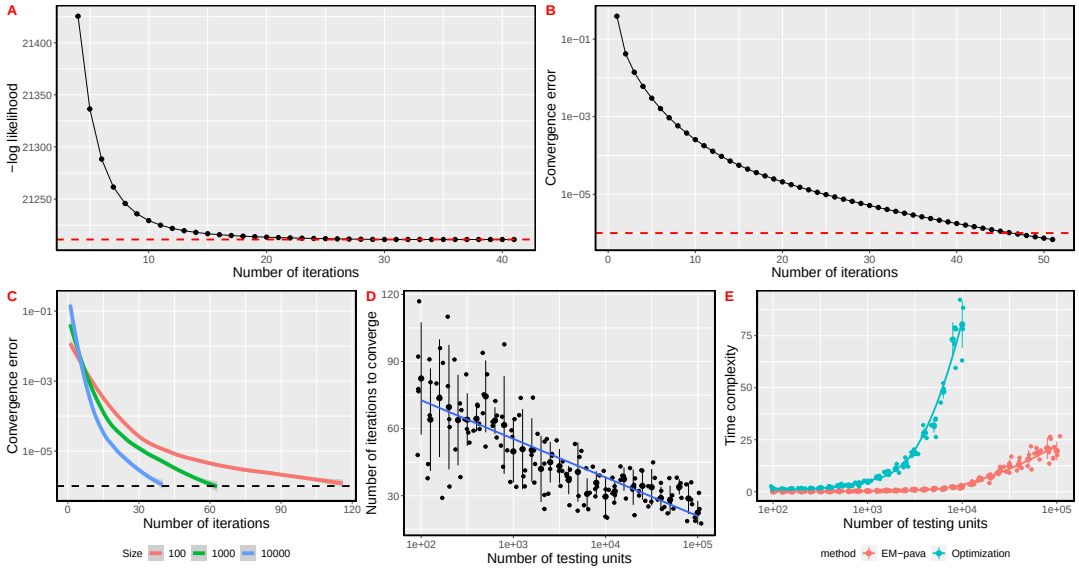


Figure 4.3: **EM-PAVA algorithm and its comparison with the direct optimization approach:** Upper panel shows a toy example with 3000 number of testing units where the convergence of objective function (with red dotted line as reference) and the convergence of error (with red dotted line as reference $10^{-6}$) are shown in panel A and panel B. Panel C and D show the number of iterations required for convergence with different numbers of testing units. The comparison of time complexity (CPU time in seconds evaluated with Inter® Core™ i5-7400HQ CPU processor) between EM-PAVA (red) and the direct optimization approach (Augmented Lagrangian, blue) is shown in panel E, with units from 100 to $10^5$ (Augmented Lagrangian approach is only evaluated up to $10^4$ due to extreme complexity).

# Chapter 5

# GraphicalT: latent clustering for local FDR computation

## 5.1 Introduction

The empirical-Bayes methods developed in earlier chapters treat all the inference units as being exchangeable; their latent parameter values are treated as draws from a distribution estimated from the entire collection. In many applied settings there is auxiliary information relating the units. I am particularly interested in examples where this auxiliary information takes the form of an undirected (and unweighted) graph. Nodes of the graph constitute the basic inference units that we aim to test for some condition effects. Edges convey some additional information that the nodes share. For example, for peptide-arrays the amino-acid sequence content of the peptides induces a graph with neighboring peptides having sufficiently similar sequence. For brain imaging data, nodes are voxels and edges convey spatial neighborhood information.

Some large-scale testing methods take advantage of such graph-associated auxiliary data to improve power. For example, Sun and Cai [2009] built a latent hidden Markov model (HMM) for the sequential dependence structure. The method was applied to an influenza-like illness surveillance study for detecting the timing of epidemic periods. Liu et al. [2012] generalized the hidden Markov model (HMM) into a Markov random field (MRF) which does not assume the specific underlying line graph structure. The proposed method was shown to have good operating characteristics through simulation studies and the application of Genome-wide Association Studies (GWAS). An empirical-Bayes method focused on local false discovery rates

and brain-image data was presented in [Vo et al., 2021]. That modeling approach used the idea that neighbors in the graph may share expected values, and thereby power can be improved by constraining the dimension of the unknown parameter space. There was empirical support for the model in a brain imaging study of Alzheimer's disease, though the parametric assumptions and the computational complexity may limit the utility of this approach.

In this chapter, I propose a flexible semiparametric method to utilize graph-associated information. It is compelling in part because it relies in a straightforward way on repeated calculation of t-statistics; it is also based on an important statistical fact that local false discovery rate, as an expected value, can be usefully represented as an average of conditional expected values, where we average over some relevant piece of missing data. More specifically, take the two-group comparison for concreteness, and at one unit $i$, let $\mathrm{H}_{0,i}$ be the null hypothesis that the difference in mean parameters between two conditions equals 0. Local FDR ($l_i$) computations considered so far evaluate $\mathbb{P}(\mathrm{H}_{0,i}|\mathbf{x}_i, \mathbf{y}_i)$ where $\mathbf{x}_i, \mathbf{y}_i$ records data on the two samples from unit $i$. The full data set ($\mathbf{x} = \{\mathbf{x}_i\}$, $\mathbf{y} = \{\mathbf{y}_i\}$) is used to estimate the involved distributions (e.g., $g$ and $h$ in `MixTwice`), but each inference statistic is computed from the local data $\mathbf{x}_i, \mathbf{y}_i$ alone using the globally-estimated distributions. A more ambitious calculation is $\mathbb{P}(\mathrm{H}_{0,i}|\mathbf{x}, \mathbf{y})$ where $\mathbf{x}, \mathbf{y}$ records all the data in the system. This would seem to require an explicit model specification relating all components, which might be very difficult to assess or validate. If the intrinsic dimension of the space of expected values is sufficiently low, a curious method presents itself. It is based on the idea of clustering of units according to equalities in their latent expected values. Theoretically, it is based on a simple fact about local false discovery rate:

$$l_i \;=\; \mathbb{P}(\mathrm{H}_{0,i}|\mathbf{x}, \mathbf{y}) = \mathbb{E}\left(\mathbb{P}(\mathrm{H}_{0,i}|\mathcal{G}, \mathbf{x}, \mathbf{y})|\mathbf{x}, \mathbf{y}\right). \tag{5.1}$$

Here $\mathcal{G}$ could represent any aspect of the modeled system, but a convenient one has $\mathcal{G}$ to be an undirected graph, with an edge between components $i$ and $j$ if both $\mu_i = \mu_j$ and $\nu_i = \nu_j$. Were the graph $\mathcal{G}$ known, then data from units sharing means in both conditions could be combined to assess changes between conditions.

In this chapter I present a technique to compute local FDRs semiparametrically using the available graph information. The results are limited, but I am able to demonstrate a computationally efficient method that has good operating characteristics in several simulation settings.

## 5.2 Statistical methodology

### 5.2.1 Inference problem, data, and input of the algorithm

Suppose for each testing unit $i = 1, 2, \cdots, m$, the null hypothesis compares the mean parameter between two groups $\mathrm{H}_{0,i} : \mu_i = \nu_i$ where $\mu_i$ is the mean parameter of samples $\mathbf{x}_i = \{x_{i,1}, \cdots, x_{i,n_1}\}$ and $\nu_i$ is the mean parameter of samples $\mathbf{y}_i = \{y_{i,1}, \cdots, y_{i,n_2}\}$. Sample sizes for the two groups are $n_1$, and $n_2$, respectively. The matrices $\mathbf{x}$ (of dimension $m$ by $n_1$) and $\mathbf{y}$ (of dimension $m$ by $n_2$) form the input data $\mathbf{D}$. Auxiliary information $G_{\text{aux}}$ is also available. The inference problem is to calculate for each unit $i$ the local false discovery rate $l_i = \mathbb{P}(\mathrm{H}_{0,i}|\mathbf{x}, \mathbf{y})$.

Let $\mathcal{G}$ denote the latent undirected graph connecting nodes with same expected values. Specifically, node $i$ and $j$ are connected in the graph $\mathcal{G}$ if their mean parameters are same, i.e., if $\mu_i = \mu_j$ and $\nu_i = \nu_j$. My model is that $\mathcal{G}$ is a subgraph of the auxiliary information graph $G_{\text{aux}}$, with the same nodes but potentially (and probably) fewer edges. This assumption is a way to encode the idea that dimension constraints may act locally relative to the auxiliary information. The following section discusses the algorithm of `GraphicalT` provided with the data matrix $\mathbf{D}$ and auxiliary graph $G_{\text{aux}}$.

### 5.2.2 Algorithm of GraphicalT

Based on the input $(\mathbf{D}, G_{\text{aux}})$, `GraphicalT` consists of three parts, sketched in a toy example in Figure 5.1:

1. node binding,
2. graph simulation and clique statistics,
3. locFDR averaging.

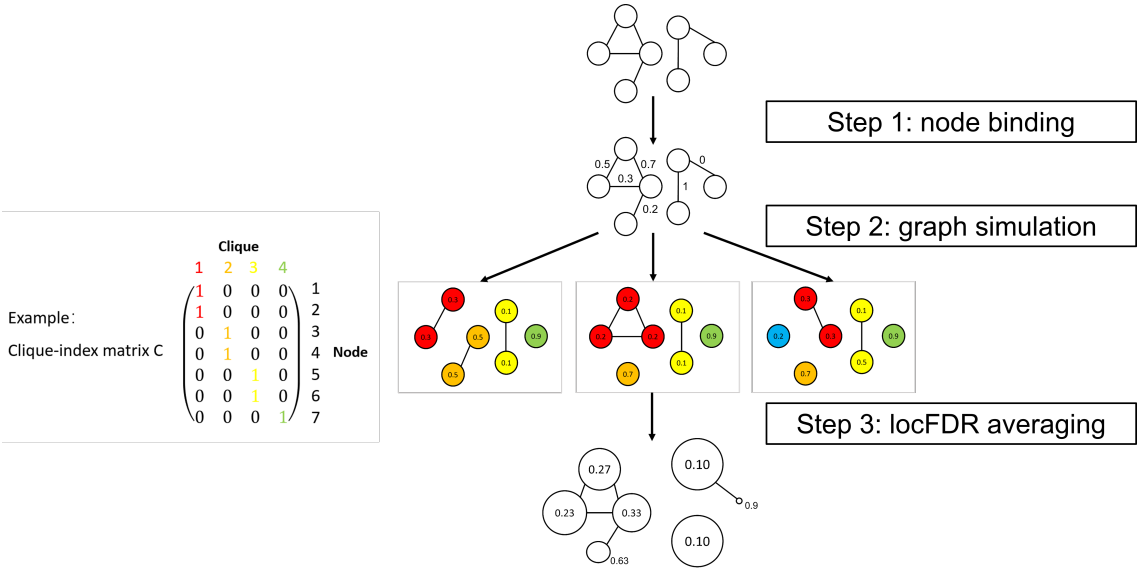


Figure 5.1: **Algorithm of GraphicalT through a toy example.** This example starts from the input of the algorithm, $\mathbf{D}$ and $G_{\text{aux}}$, a small graph with 7 nodes and 6 edges. The first step, node binding, calculates the pairwise locFDR's $\tilde{l}_e$ on each edge $e$ in $G_{\text{aux}}$. The second step, graph simulation, simulates three different realizations of $\mathcal{G}$, where each clique is coded using different color and second-stage locFDR's are calculated among cliques. The final locFDR's are calculated by average over different random graphs (node size is proportional to 1 minus the locFDR).

### Step 1: node binding

I assume that $\mathcal{G}$ is a subgraph of $G_{\text{aux}}$ with the same nodes but possibly fewer edges. Each potential edge is inferred by pairwise testing using data from the nodes incident to that edge. Specifically, suppose $G_{\text{aux}} = (V, E)$ where $V$ contains inference units that are connected in $E$. For any edge $e \in E$ where $e$ connects two nodes $j_1 \in V$ and $j_2 \in V$, the binding step examines the following null hypothesis:

$$\begin{aligned} \tilde{\mathrm{H}}^{\mathbf{x}}_{0,e} &: \mu_{j_1} = \mu_{j_2} \\ \tilde{\mathrm{H}}^{\mathbf{y}}_{0,e} &: \nu_{j_1} = \nu_{j_2}. \end{aligned}$$

Denote data vectors collected from edge $e$ connecting nodes $j_1, j_2$ be $\mathbf{x}_{j_1}, \mathbf{x}_{j_2}$ and $\mathbf{y}_{j_1}, \mathbf{y}_{j_2}$ and local false discovery rate for $e$ be $\tilde{l}_e$:

$$\begin{aligned} \tilde{l}_e &= \mathbb{P}(\tilde{\mathrm{H}}^{\mathbf{x}}_{0,e} \cap \tilde{\mathrm{H}}^{\mathbf{y}}_{0,e} | \mathbf{x}_{j_1}, \mathbf{x}_{j_2}, \mathbf{y}_{j_1}, \mathbf{y}_{j_2}) \\ &= \mathbb{P}(\tilde{\mathrm{H}}^{\mathbf{x}}_{0,e} | \mathbf{x}_{j_1}, \mathbf{x}_{j_2})\ \mathbb{P}(\tilde{\mathrm{H}}^{\mathbf{y}}_{0,e} | \mathbf{y}_{j_1}, \mathbf{y}_{j_2}) \\ &= \tilde{l}^{\mathbf{x}}_e\ \tilde{l}^{\mathbf{y}}_e. \end{aligned} \tag{5.2}$$

where the probability of intersection is simplified by the product of probabilities assuming independence between $\tilde{\mathrm{H}}^{\mathbf{x}}_{0,e}$ $\tilde{\mathrm{H}}^{\mathbf{y}}_{0,e}$. Conveniently, each test here may be performed via a classical paired t-test. We may use different options to calculate the local false discovery rates $\tilde{l}_e$ after the collection of statistics for the paired t-test (implemented as `pair.method`). Tools that are sensitive may be preferred in order not to overestimate the binding probabilities. The collection of $\{\tilde{l}_e\}$ returned by the node-binding step would be used in the next step to simulate random graphs.

**Step 2: graph simulation and clique statistics**

I use the posterior probabilities $\{\tilde{l}_e\}$ from the node-binding step to represent the averaging operation in (5.1). Simply, I treat uncertainty in $\mathcal{G}$ as equivalent to a random graph having an edge at $e$ from the $G_{\text{aux}}$ with probability $\tilde{l}_e$. A random graph $\mathcal{G}$ is simulated repeatedly where each edge $e$ is simulated through an independent Bernoulli trial with probability $\tilde{l}_e$. Then this random graph $\mathcal{G}$ guides a combination of inference units prior to comparison between groups.

On any realized $\mathcal{G}$, consider node $i$ and let $C(i)$ denote a maximal clique that contains $i$. That is, all nodes $j \in C(i)$ are connected in $\mathcal{G}$, and the set cannot be expanded while retaining complete connections. All nodes in $C(i)$ are deemed to have the same mean parameter values in both conditions, taken separately. In other words, $\mu_j$ is constant for all $j \in C(i)$ and similarly $\nu_j$ is constant, though we do not know the status of the difference $\mu_j - \nu_j$, except that it is constant. Recognizing the shared parameter states, I sum the measurements within each sample and across the clique to obtain clique-level data, which may be organized over the samples by a matrix operation: $\mathbf{x}'$ and $\mathbf{y}'$ of dimension $K$ by $n_1$ and $K$ by $n_2$, where $\mathbf{x}' = (C)^T\mathbf{x}$ and $\mathbf{y}' = (C)^T\mathbf{y}$. There may be dependence between units, but that has no negative impact on the procedure, as the variance of the sum will be accounted for in the test-statistic construction. Here $K$ counts the maximal cliques and $C$ is an $m \times K$ incidence matrix with $C_{i,C(i)} = 1$ and 0 elsewhere. The clique-level sums are now amenable to an unpaired, between-group t-test, since $\mathrm{H}_{0,j}$ is constant for all $j \in C(i)$, and therefore equal to $\mathrm{H}_{0,i}$. Next, `GraphicalT` computes the clique-specific t-statistic, and from the system-wide collection of statistics derives a clique-level local FDR:

$$l_{C(i)} \quad = \quad \mathbb{P}\left(\mathrm{H}_{0,i} | \mathcal{G}, \mathbf{x}, \mathbf{y}\right) \tag{5.3}$$

I compute maximal cliques of each $\mathcal{G}$ by applying `igraph::max_cliques` [Csardi and Nepusz, 2006]. A node may be in more than one maximal clique. Presently I

average the $t-$statistics computed on each one; alternatively, a random one would meet the requirements of 5.1.

### Step 3: locFDR averaging

The final local false discovery rates $l_{\text{GraphicalT},i}$ of unit $i$ is the average of all $l_{C^t(i)}$ over simulated random graphs, $\mathcal{G}^1, \mathcal{G}^2, \cdots, \mathcal{G}^{\text{Nsim}}$ where Nsim is the number of random-graph realizations:

$$\begin{aligned} l_{\text{GraphicalT},i} &= \mathbb{P}(\text{H}_{0,i}|\mathbf{x},\mathbf{y}) \\ &= \mathbb{E}\left(\mathbb{P}(\text{H}_{0,i}|\mathcal{G},\mathbf{x},\mathbf{y})|\mathbf{x},\mathbf{y}\right) \\ &\approx \frac{1}{\text{Nsim}}\sum_{t=1}^{\text{Nsim}} l_{C^t(i)} \end{aligned} \tag{5.4}$$

where the expectation is taken with respect to random graph simulation.

The algorithm of `GraphicalT` is summarized as the following:

---

**Algorithm 1** GraphicalT

---

**Input:** Units by samples data matrices $\mathbf{x}, \mathbf{y}$ and input graph $G_{\text{aux}}$

**Output:** Local false discovery rates: $l_{\text{GraphicalT},i} = \mathbb{P}(\text{H}_{0,i}|\mathbf{x},\mathbf{y})$

1: **Step 1, node binding.**

1. For every edge $e = \{j_1, j_2\}$ of the input graph $G_{\text{aux}}$, identify data vectors $\mathbf{x}_{j_1}, \mathbf{x}_{j_2}, \mathbf{y}_{j_1}, \mathbf{y}_{j_2}$
2. Do one paired t-test between $\mathbf{x}_{j_1}, \mathbf{x}_{j_2}$ and a second one between $\mathbf{y}_{j_1}, \mathbf{y}_{j_2}$. Calculate the local false discovery rate:

$$\tilde{l}_e = \mathbb{P}(\tilde{\text{H}}^{\mathbf{x}}_{0,e}|\mathbf{x}_{j_1},\mathbf{x}_{j_2})\ \mathbb{P}(\tilde{\text{H}}^{\mathbf{y}}_{0,e}|\mathbf{y}_{j_1},\mathbf{y}_{j_2})$$

2: **Step 2, graph simulation and clique statistics.** Repeat for simulation trials:

1. Simulate random graph $\mathcal{G}$ based on the collection of edge probabilities $\{\tilde{l}_e\}$
2. Find a maximal clique for each node and form a clique-index matrix $C$.
3. Reduce to clique-level data: $\mathbf{x}' = (C)^T\mathbf{x}$ and $\mathbf{y}' = (C)^T\mathbf{y}$
4. For each clique, do an unpaired t-test on clique-level data.
5. Calculate the clique-level local false discovery rate:

$$l_{C(i)} = \mathbb{P}\left(\text{H}_{0,i}|\mathcal{G},\mathbf{x},\mathbf{y}\right)$$

3: **Step 3, locFDR averaging.** Average local false discovery rates over simulated random graphs, $\mathcal{G}^1, \mathcal{G}^2, \cdots, \mathcal{G}^{\text{Nsim}}$,

$$l_{\text{GraphicalT},i} = \mathbb{E}\left[\mathbb{P}(\text{H}_{0,i}|\mathcal{G},\mathbf{x},\mathbf{y})|\mathbf{x},\mathbf{y}\right] \approx \frac{1}{\text{Nsim}}\sum_{t=1}^{\text{Nsim}} l_{C^t(i)}$$

---

## 5.3 Simulation study

FDR control and power are important operating characteristics in evaluating large-scale hypothesis testing. Meanwhile, compared to the classic Student's-t approach (e.g., BH, Storey's `qvalue`, etc.), `GraphicalT` requires the auxiliary data graph $G_{\text{aux}}$ as the input of the algorithm. As the underlying graph $\mathcal{G}$ is only a subgraph of this input data graph $G_{\text{aux}}$, there might be edges in $G_{\text{aux}}$ not in $\mathcal{G}$; $G_{\text{aux}}$ might connect units that do not have the same underlying expected means.

In this numerical study, I evaluate the FDR control and power of `GraphicalT`. I also seek to understand the performance under scenarios where the input graph $G_{\text{aux}}$ is similar to the underlying $\mathcal{G}$ and scenarios where there are much more edges. Specifically, I simulate 400 nodes where half of them have difference between groups ($\mu_i \neq \nu_i$), i.e., $\pi_0 = 0.5$. This is demonstrated using different colors in Figure 5.2. Also, among those 400 nodes, half of them are contained in one of 20 cliques, each with 10 nodes. For example, if node $k_1, k_2, \cdots, k_{10}$ are connected through a clique $k, k \in (1, 2, \cdots, 20)$, then $\mu_{k_1} = \mu_{k_2} = \cdots = \mu_{k_{10}}$ and $\nu_{k_1} = \nu_{k_2} = \cdots = \nu_{k_{10}}$. The rest half of the nodes are isolated. The degrees of those connected nodes are 9 while 0 for those isolated ones.

There are in total 900 edges in $\mathcal{G}$. Through this numerical experiment, I expand this $\mathcal{G}$ to get different simulated input graphs by randomly connecting some units that are not connected in the underlying $\mathcal{G}$. As an illustration, Figure 5.3 shows the adjacency matrix of input graph $G_{\text{aux}}$ (left column) and the output graph constructing by $\{\tilde{l}_e\}$ after the binding step (right column). As shown in Figure 5.3, the adjacency matrix after the binding step is almost the same as the original one if $G_{\text{aux}}$ is similar to $\mathcal{G}$ (panel A). When $G_{\text{aux}}$ connects much more nodes than $\mathcal{G}$, the graph after the node-binding step is less clear, but could still differentiate the pattern between cliques and isolated nodes.

Figure 5.4 summarizes empirical FDR and empirical true positive rate under FDR control 0.1 using 50 time randomly sampled data sets, each with a different simulated $G_{\text{aux}}$. The x-axis records the number of edges in $G_{\text{aux}}$ but not in $\mathcal{G}$. This experiment concludes that the proposed `GraphicalT` has increased power property without inflating FDR control, and is robust even though there is large difference between the input data graph and the underlying one.

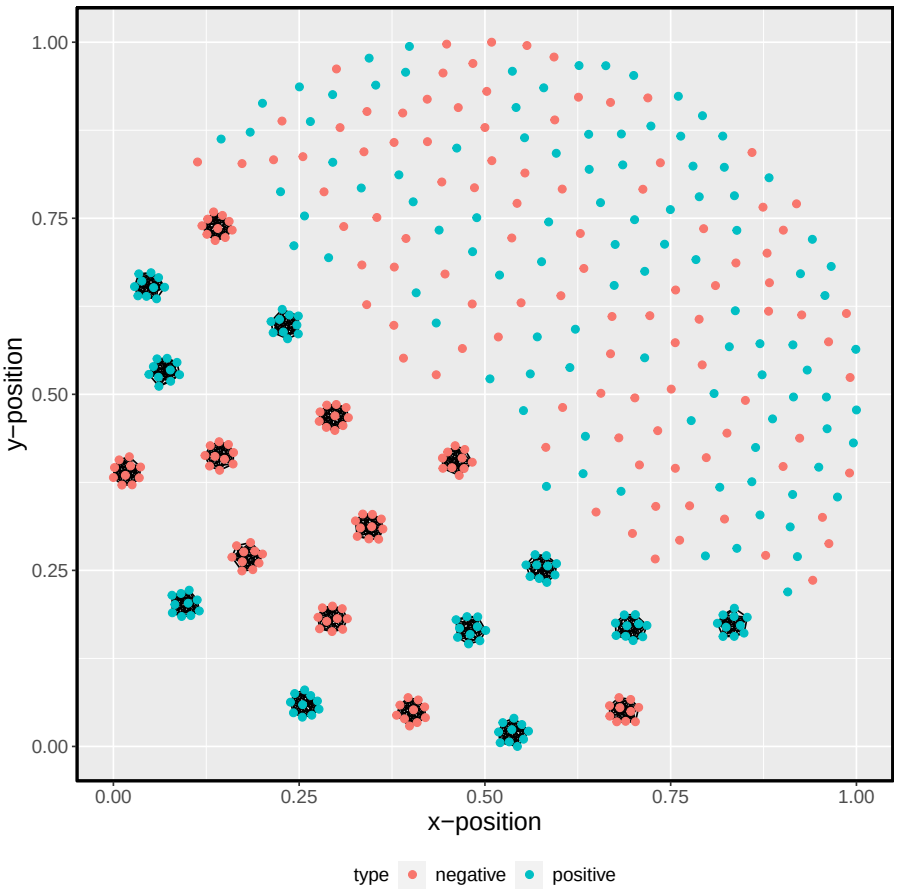


Figure 5.2: **Underlying graph $\mathcal{G}$ in simulation setting:** this graph shows the connection structure of 400 testing units. Half of the units (200) were connected within 20 fully-connected cliques while the rest of the units are isolated. Units with a non-zero latent signal effect (under alternative hypothesis $H_1$) are colored in red, while those without a signal effect (under $H_0$) are colored in blue.

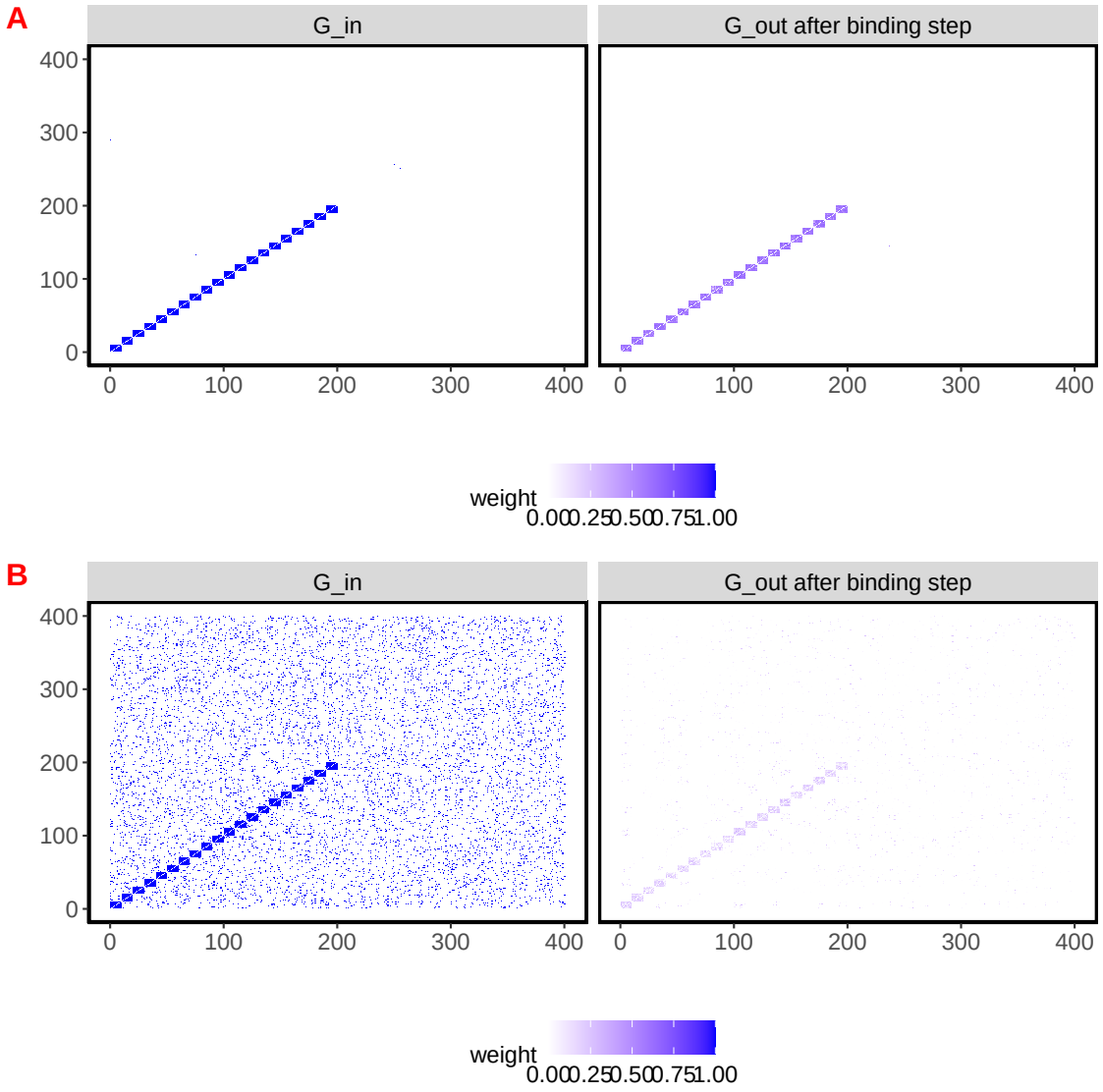


Figure 5.3: **Adjacency matrix of auxiliary graph $G_{\text{aux}}$ and the output graph after node binding:** The left panel shows the adjacency matrix of auxiliary graph $G_{\text{aux}}$ and the right panel shows the adjacency matrix of the graph constructed by $\{\tilde{l}_e\}$ in the node-binding step. Panel A shows a scenario where $G_{\text{aux}}$ only has 10 more edges than $\mathcal{G}$, while panel B shows a scenario where this number is 10,000. Weight of adjacency matrix is coded using different colors.

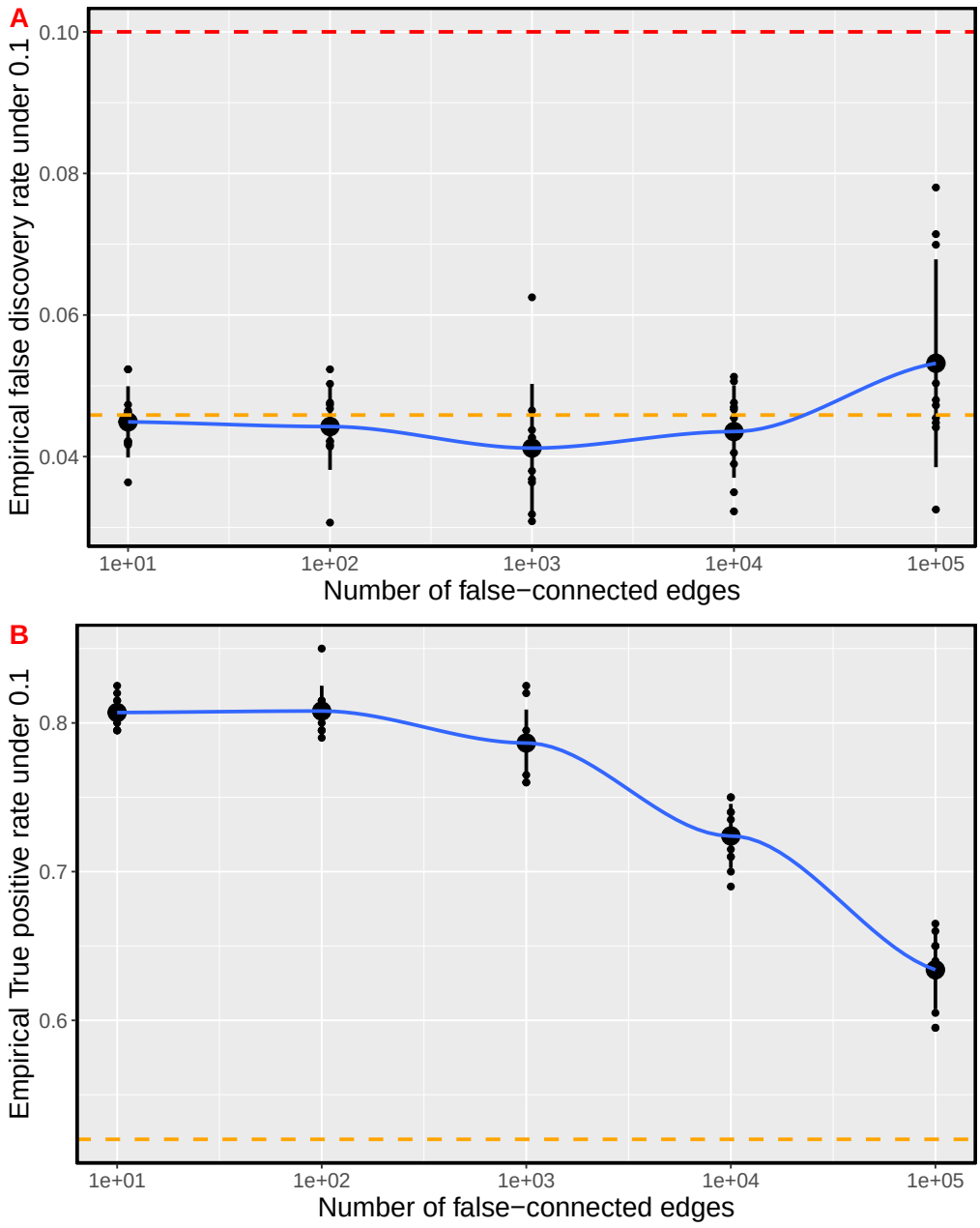


Figure 5.4: **FDR and true positive rate comparing GraphicalT and Student's t:** Each dot in the figure is a randomly simulated data set with different $G_{\text{aux}}$. The x-axes records the number of edges in $G_{\text{aux}}$ but not in $\mathcal{G}$ and y-axis summarizes empirical false discovery rate (FDR, panel A) and true positive rate (panel B) under the nominated significance level 0.1. Orange dashed lines in both panels indicate the performance using Student's t approach (specifically, q-value) and the red dashed line in panel A refers to the nominated FDR level (0.1).

## 5.4 Brain image example with 3D lattice graph

One example of graph-associated data is the structural magnetic resonance imaging data measured in studies of brain structure, as part of the Alzheimer's Disease Neuroimaging Initiative (ADNI-2) [Weiner and Veitch, 2015, Vo et al., 2021]. Three dimensional brain image data were collected from 123 cognitively normal control subjects (CN) and a second group of 148 subjects suffering from late-stage mild cognitive impairment (MCI), a precursor of Alzheimer's disease (AD). Gray matter tissue probability maps derived from the co-registered T1-weighted magnetic resonance imaging (MRI) data were pre-processed so that specific spatial coordinates (voxels) could be compared between the two samples of brain images. I used the data prepared and reported in [Vo et al., 2021]. The auxiliary data graph is simply a three-dimensional lattice where neighboring voxels are related by spatial coordinates. Preprocessing via low marginal deviation filtering [Bourgon et al., 2010] gives the final graph with 464,441 voxels.

Among all 464,441 voxels, BH [Benjamini and Hochberg, 1995] yields 5130 voxels under 5% FDR control. The `qvalue` method [Storey, 2002] estimates the null proportion $\hat{\pi}_0 = 0.82$ and yields 5817 voxels. The `ASH` procedure [Stephens, 2017], with the same estimated null proportion, provides a slightly larger list (6057). The 5% FDR-controlled list by `GraphicalT` contains 7934 significant voxels under the same FDR control level and with the same estimated null proportion. As shown in the Venn Diagram of Figure 5.5, among those 7934 voxels found by `GraphicalT`, 5813 of them are also on the discovery list of `qvalue` and `ASH` while there are 1940 voxels not found by other tools. Examples of those voxels on the discovery list are discussed in Supplementary section D.2 The lower panel of Figure 5.5 compares the locFDR's reported using `GraphicalT` with locFDRs using Student's t.

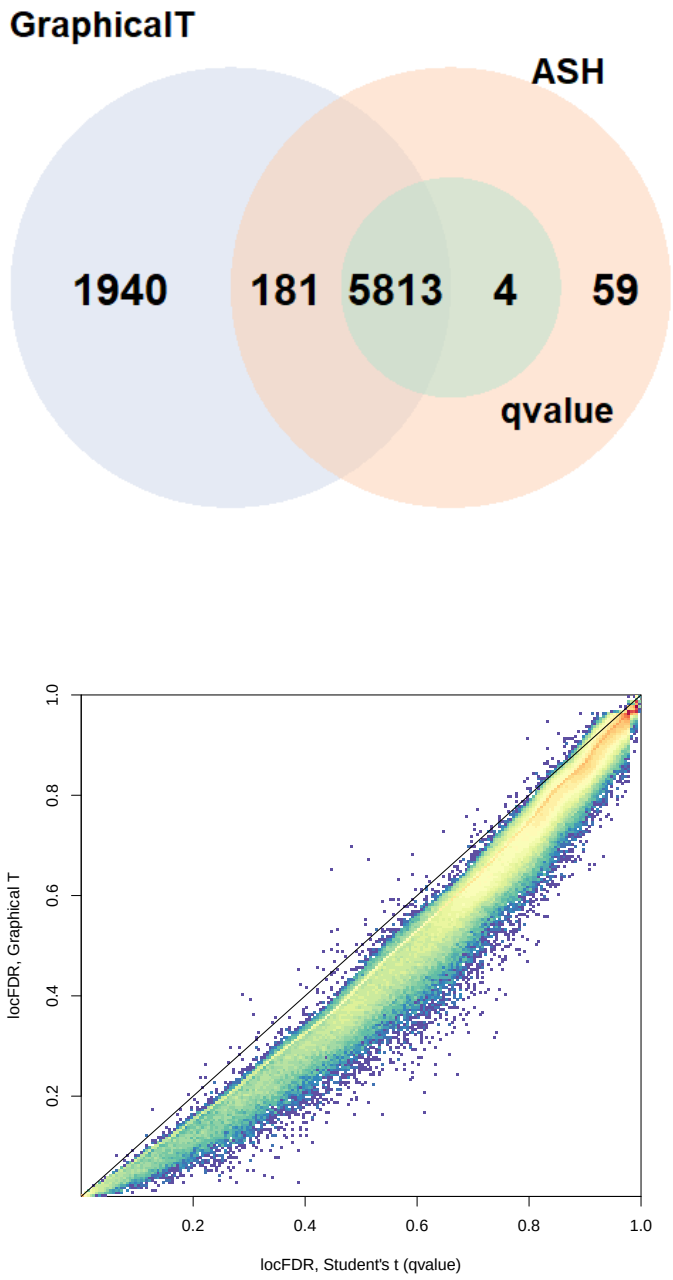


Figure 5.5: **Venn diagram and two-dimensional histogram comparing different testing tools:** The upper panel shows the Venn Diagram of discovery list comparing different methods under FDR 0.05 level, and the lower panel shows two-dimensional histogram of local FDRs using Student's t procedure (x-axis) and GraphicalT procedure (y-axis).

# Chapter 6

# Conclusion

The large-scale hypothesis testing problem has been widely studied and strongly motivated in a wide variety of subject-matter domains. In responding to practical challenges that I encountered in the high density peptide array data and other bioinformatics experiments, I made several contributions in my dissertation.

Many conventional large-scale testing tools control the false discovery rate (FDR), however performance improvements are possible. One of the key challenges is to better handle the variance that fluctuates over the testing units, especially with limited sample sizes. Motivated by this, I proposed an empirical-Bayes computational tool, `MixTwice`, that intervenes on estimated effect and estimated standard error to calculate the local false discovery rates for each testing unit. The proposed method not only involves a shape-constrained mixture distribution for latent effects, but also a separate nonparametric mixture for variance parameters. `MixTwice` was shown to have better operating characteristics through a variety of numerical experiments compared to conventional testing approaches and also achieved better power and reproducibility in several applications.

Large-scale testing problems with multiple-group comparison are less studied compared to scenarios with only two conditions, even though they also have a wide range of applications. Conventional computational tools that make multiplicity adjustment based on F test p-values can be applied in such settings, but sometimes they also have limited power. `MixTwice-ANOVA` is proposed in this dissertation in response to this challenge, by applying empirical Bayes mixing to sum-of-square components of the F statistic. This statistical methodology is shown to have good operating characteristics in simulations and in an example comparing three groups of Sjogren disease (SjD) patients.

In the development of statistical methodology, computational aspects are essential to solve the high throughput experimental problem efficiently. Such topics involve nonparametric MLE and high-dimensional optimization problems with shape constraints. Motivated from theoretical guarantees and applications of isotonic regression, I proposed the `EM-PAVA` algorithm within `MixTwice` for optimization subject to a unimodal or monotonic shape constraint. Compared to conventional algorithms for the shape-constrained optimization problem, `EM-PAVA` substantially improves computational efficiency.

Graph-associated data is common in applications. Such graph-associated information, as additional data incorporating in large-scale testing, can improve the local false discovery statistics. However, most of the computational tools incorporating such information proceed through parametric assumptions and elaborate model specifications. To construct a simpler approach for the analysis of graph-associated data, I presented `GraphicalT` to compute local false discovery rates semiparametrically using the available graph information. The method showed good performance in synthetic examples and in a brain-imaging problem from the study of Alzheimer's disease.

My contributions to large-scale data inference are outlined in the preceding paragraphs. Those contributions aim to standardize modern statistical inference and computational tools, such as empirical-Bayes inference, nonparametric and semiparametric estimation, constrained optimization, that could be used in data analysis across a wide range of applications. There are also several unanswered questions that merit more investigation. First, in the large-scale hypothesis testing problems, different researchers focused on different perspectives to improve power property. For example, `MixTwice` concentrates on better incorporating the information associated with the variance parameter. Many computational tools suggest the unit-specific covariates could also gain additional power. It is not surprised that the combination of these two insights could improve the operating characteristics of large-scale hypothesis testing. However, estimating the mixing distribution, especially the mixing distribution of effect and variance, might be complicated and challenging with the additional unit-specific covariates. Second, besides hypothesis testing, ranking is also one of the most important but challenging tasks of large-scale data inference. Also, the combination of testing and ranking is widely applied in many data problems. Recent progress shows the empirical-Bayes mixture model can have improved operating characteristics for the large-scale ranking problem (e.g., [Henderson and Newton, 2016]). That procedure might also be improved with a separate mixing

distribution estimation on variance and a proper regularization on effect. Last, numerical experiments and many data examples suggest the improved properties of `GraphicalT`. However, it might be good to learn more about the mechanism that underpins it, or even the theory that underpins it. This study, which will most likely use a simple case as an example, may be useful in better understanding and interpreting the methodology. Besides those three specific examples, there are certainly many other interesting questions to advance data analysis in large-scale data inference and I encourage researchers to dig further.

# Appendix A

# Appendix to Chapter 2

*The material in this chapter was reported previously in Supplementary material of Zheng et al. [2021]*

## A.1 Gradient and Hessian of optimization objective

We derive the gradient and Hessian of the log-likelihood equation, $l(g,h)$. Recall the definition of $l(g,h)$:

$$\begin{aligned} l(g,h) &= \sum_{i=1}^{m} \log p(x_i, s_i^2|g,h) \\ p(x_i, s_i^2|g,h) &= \sum_k \sum_l g_k h_l \frac{1}{\sqrt{b_l}} \phi\left(\frac{x_i - a_k}{\sqrt{b_l}}\right) \frac{\nu}{b_l} \chi_{2,\nu}\left(\frac{\nu s_i^2}{b_l}\right). \end{aligned}$$

To simplify notation we use $c_{i,k,l}$ to denote the prior, mixture density of sample $i$ on the grid $a_k, b_l$, and we let $d_i$ denote the observation density:

$$\begin{aligned} c_{i,k,l} &:= \frac{1}{\sqrt{b_l}} \phi\left(\frac{x_i - a_k}{\sqrt{b_l}}\right) \frac{\nu}{b_l} \chi_{2,\nu}\left(\frac{\nu s_i^2}{b_l}\right) \\ d_i &:= p(x_i, s_i^2|g,h) = \sum_k \sum_l g_k h_l \; c_{i,k,l}. \end{aligned}$$

Consider the parameter vector $(g,h)$ of length $(2K+1)+L$, where the first $2K+1$ components are for the effect mixing probabilities, $g = (g_k)$, and the remaining $L$ components are for variance mixing probabilities $h = (h_l)$. The quantities $c_{i,k,l}$ depend on the data, the support points but not the probabilities $g$ and $h$.

Constraints are critical to the optimization; of course all elements of $g$ and $h$ must be positive and sum to unity. We also impose a unimodality constraint on $g$. But in deploying the augmented Lagrangian method, these constraints act on the differentiable function $l(g,h)$, which we consider initially as varying freely over $2K+L+1$ Euclidean space. The gradient of $l(g,h)$ is a column vector of length $(2K+1)+L$ with the following format:

$$\nabla l(g,h) \;=\; \left(\left(\frac{\partial l(g,h)}{\partial g}\right)', \left(\frac{\partial l(g,h)}{\partial h}\right)'\right)'$$

where each component has the explicit form:

$$\begin{aligned}\frac{\partial l(g,h)}{\partial g_k} &= \sum_{i=1}^{m}\frac{1}{d_i}\sum_{l} h_l\ c_{i,k,l}\\ \frac{\partial l(g,h)}{\partial h_l} &= \sum_{i=1}^{m}\frac{1}{d_i}\sum_{k} g_k\ c_{i,k,l}.\end{aligned}$$

The Hessian of $l(g,h)$ is a $(2K+1)+L$ by $(2K+1)+L$ matrix:

$$\nabla^2 l(g,h) = \begin{pmatrix} A & B \\ B' & C \end{pmatrix}$$

where matrix $A$ ($2K+1$ by $2K+1$) contains second derivative with respect to $g$, matrix $C$ ($L$ by $L$) contains second derivative with respect to $h$ and matrix $B$ ($2K+1$ by $L$) contains second derivative with respect to $g$ and $h$.

For entries of matrix $A$:

$$\begin{aligned}\frac{\partial^2 l(g,h)}{\partial g_k^2} &= -\sum_{i=1}^{m}\frac{1}{d_i^2}\left(\sum_{l} h_l\ c_{i,k,l}\right)^2\\ \frac{\partial^2 l(g,h)}{\partial g_{k_1}\partial g_{k_2}} &= -\sum_{i=1}^{m}\frac{1}{d_i^2}\left(\sum_{l} h_l\ c_{i,k_1,l}\right)\left(\sum_{l} h_l\ c_{i,k_2,l}\right).\end{aligned}$$

For entries of matrix $C$:

$$\begin{aligned}\frac{\partial^2 l(g,h)}{\partial h_l^2} &= -\sum_{i=1}^{m}\frac{1}{d_i^2}\left(\sum_{k} g_k\ c_{i,k,l}\right)^2\\ \frac{\partial^2 l(g,h)}{\partial h_{l_1}\partial h_{l_2}} &= -\sum_{i=1}^{m}\frac{1}{d_i^2}\left(\sum_{k} g_k\ c_{i,k,l_1}\right)\left(\sum_{k} g_k\ c_{k,l_2}\right).\end{aligned}$$

For entries of matrix $B$:

$$\frac{\partial^2 l(g,h)}{\partial g_k\partial h_l} \;=\; \sum_{i=1}^{m}\frac{1}{d_i^2}\left(c_{i,k,l}\ d_i - \sum_{l} h_l\ c_{i,k,l}\sum_{k} g_k\ c_{i,k,l}\right).$$

## A.2 Random subsampling

The optimization to compute $\hat{g}$ and $\hat{h}$ becomes computationally challenging as the number of testing units increases. `MixTwice` provides an option for users to use a randomly-selected subset of testing units to obtain the fitted distributions. Here we illustrate the compute-time improvements associated with relatively little degradation in the quality of the estimates.

We use the CCP+RF+ RA example to illustrate the random subsampling properties in terms of estimation error and computational benefit. Relative to the estimate obtained from half the units, we evaluate the discrepancy in distribution estimation and the user's CPU time (with Inter(R) Core(TM) i5-7400HQ CPU processor) when the `prop`, the proportion of testing units used to fit the distribution, changes. We use 1-Wasserstein distance between two cumulative distribution functions as the metric to evaluate the discrepancy from the case when `prop` $= 0.5$ as benchmark.

Figure A.1 summaries the result. Panel A highlights the estimation of $\hat{g}, \hat{h}$ when `prop` $= 0.5, 0.1, 0.01$ where the estimations are quite similar. Panel B shows how the discrepancy decreases when the proportion of testing units used to fit the distribution increases. Note that even when `prop` $= 0.01$, the discrepancy is quite small (error in $\hat{g}$ less than 0.02 and error in $\hat{h}$ only $10^{-}4$). Panel C shows the computational benefits.

## A.3 On identifiability

On units $i$ with a fixed, known standard error $\sigma$, the mixing model for effects $\theta_i$ is puts point mass at 0, with probability $\pi_0$, and distributes the remaining mass according to some distribution $g_{\text{alt}}$, which in the following is treated as a density function with respect to Lebesgue measure. Ignoring mixing over $\sigma$ (according to $h$), the predictive density of estimator $\hat{\theta}_i$, at argument $x$, is

$$\pi_0 \frac{1}{\sigma}\phi\left(\frac{x}{\sigma}\right) + (1-\pi_0)\int \frac{1}{\sigma}\phi\left(\frac{x-\theta}{\sigma}\right) g_{\text{alt}}(\theta)\, d\theta$$

where $\phi$ is the standard normal density reflecting Gaussian errors of the estimators. The alternative effect density $g_{\text{alt}}$ need not be zero in neighborhoods of the null, in which case it may happen that there exists a *gap* $c = c_\sigma > 0$ for which, for all $x$,

$$c_\sigma \frac{1}{\sigma}\phi\left(\frac{x}{\sigma}\right) \leq \int \frac{1}{\sigma}\phi\left(\frac{x-\theta}{\sigma}\right) g_{\text{alt}}(\theta)\, d\theta. \tag{A.1}$$

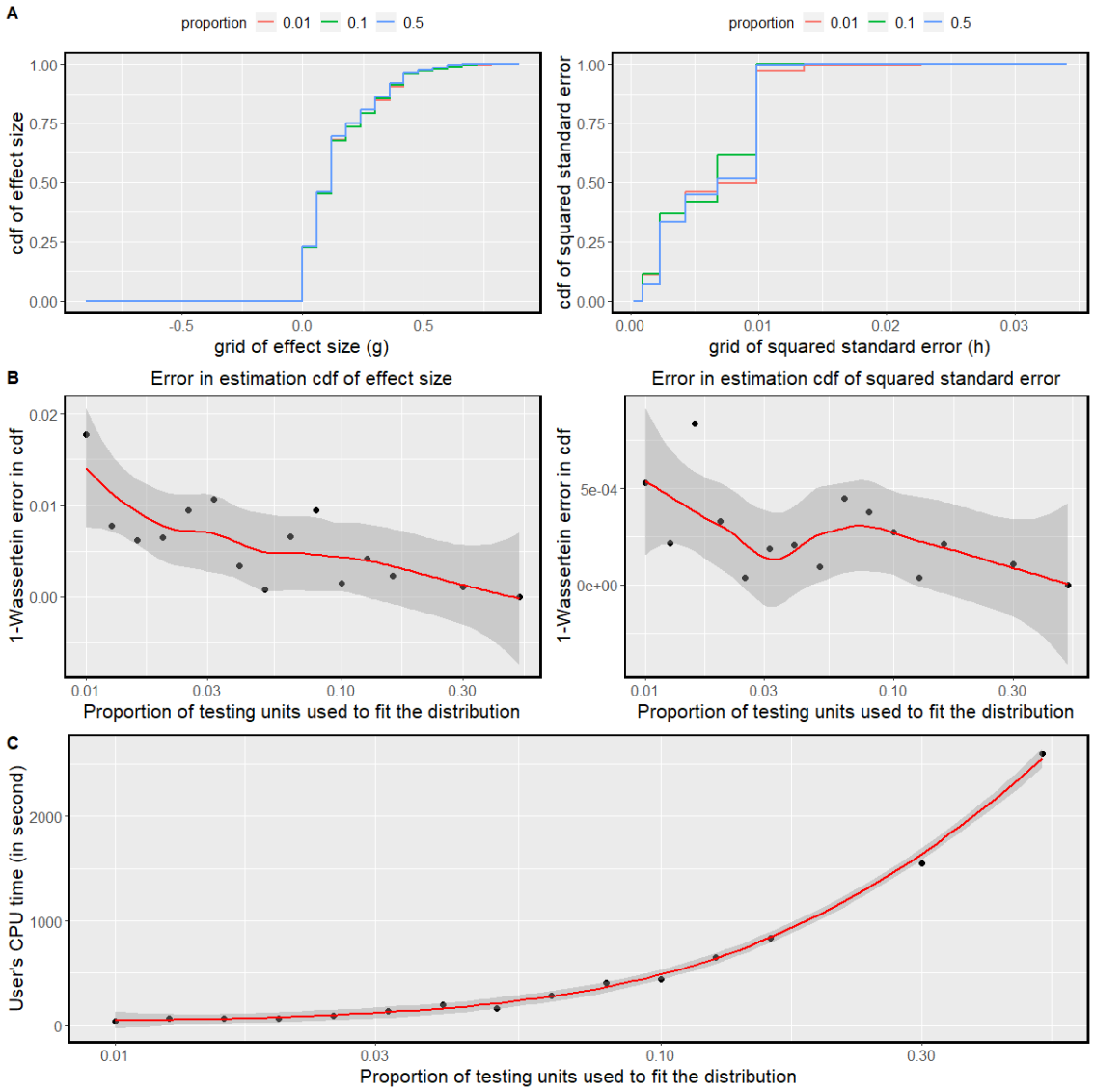


Figure A.1: **How does random subsampling influence estimation accuracy and computational efficiency?** Panel A shows the estimation in $\hat{g}, \hat{h}$ when various proportions of the units are used for estimation. Panel B shows the 1-Wasserstein discrepancy (between estimate at that proportion and estimate from half the units) as a function of subsampling proportion. Panel C shows the corresponding CPU time.

If there is a gap, the alternative predictive density contains within it a shrunken version of the null. The problem with such a gap is well known; an amount $c_\sigma(1-\pi_0)$ of mass from the alternative predictive component may be pushed into the null component, with no effect on the marginal predictive density. A small gap emerges in cases such as *spiky* (Figure 1, main) where $g_{\text{alt}}$ concentrates substantial mass near the null value. This constitutes an identifiability issue, however, we find that the gap is small or nonexistent in many cases, and anyway can be shown to converge to zero when $\sigma$ converges to zero. To see this feature, rearrange (A.1) to see that for all $x$ we require

$$c_\sigma \leq \int \exp\left\{\frac{1}{2\sigma^2}(2\theta x - \theta^2)\right\} g_{\text{alt}}(\theta)\, d\theta$$

The bound on the right depends on $x$; differentiating in $x$, under the integral gives

$$\int \frac{\theta}{\sigma^2} \exp\left\{\frac{1}{2\sigma^2}(2\theta x - \theta^2)\right\} g_{\text{alt}}(\theta)\, d\theta.$$

Notice that if $g_{\text{alt}}$ is symmetric, then at $x = 0$ this derivative is zero, and so

$$c_\sigma \leq \int \exp\left\{\frac{-\theta^2}{2\sigma^2}\right\} g_{\text{alt}}(\theta)\, d\theta.$$

Taking appropriate limits in $\sigma$ towards 0 shows that $c_\sigma$ must vanish, which will happen with increasing amounts of information per unit. The question of mixing over $\sigma$ using the second mixing distribution $h$ is not directly addressed by the above computations. However, we would predict from them that as the estimated mixing distribution $\hat{h}$ concentrates more of its mass on small standard errors, then inferences about effects $\theta_i$ will be ever more reliable.

## A.4 Compare MixTwice, ASH and two-step ASH

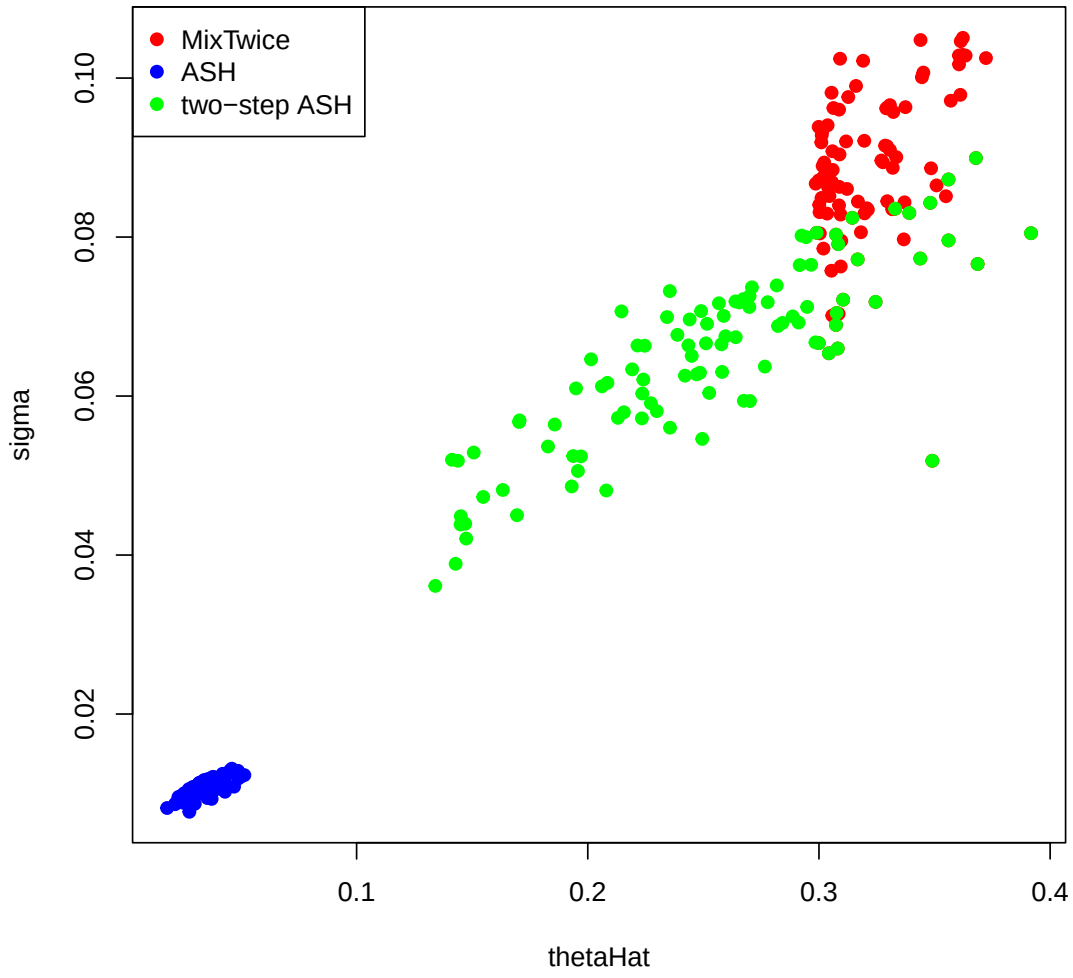


Figure A.2: **Different selection pattern among ASH, two-step ASH and MixTwice**: Scatter plot comparing estimated effect size ($x$, x-axes) and estimated standard error ($s$, y-axes) for top 100 peptides with smallest locFDR in ASH (green), two-step ASH (green) and red (MixTwice) in the CCP+RF- vs control example.

# Appendix B

# Appendix to Chapter 3

## B.1 Grid in MixTwice-ANOVA optimization

The grid of the optimization problem, including the grid spread (i.e., lower and upper range) and grid spacing (i.e., linear space or quadratic space, etc.) is essential to finite grid approximation for each distribution $g$ and $h$. Note that $\mathbb{E}\left(\frac{\text{SSE}}{\sigma^2}\right) = n-m$, so we could simply take the range of the grid of $h$ to be the range of $\frac{\text{SSE}}{n-m}$. However, it might take further effort to think of the grid of $g$, the distribution of $\lambda$. Knowing the fact that $\text{SSB}|\sigma^2$ follows non-central chi-square distribution with degree of freedom $m-1$ and non-central parameter $\frac{n\lambda}{\sigma^2}$, $\mathbb{E}(\frac{\text{SSB}}{\sigma^2}) = m-1+\frac{n\lambda}{\sigma^2}$. By taking $\sigma^2 = \frac{\mathbb{E}(\text{SSE})}{n-m}$, it follows that:

$$\lambda = \frac{\mathbb{E}(\text{SSB})}{n} - \frac{(m-1)}{n}\frac{\mathbb{E}(\text{SSE})}{n-m}. \tag{B.1}$$

On the other hand, we can also think in another way $\frac{\frac{\text{SSB}}{m-1}}{\frac{\text{SSE}}{n-m}}|(\sigma^2,\lambda) \sim F_{m-1,n-m,\frac{n\lambda}{\sigma^2}}$, the non-central F distribution with degree of freedom $m-1$ on numerator, $n-m$ on denominator and non-central parameter $\frac{\lambda}{\sigma^2}$. It is equivalently to say that $\mathbb{E}\left(\frac{\text{SSB}}{\text{SSE}}\right) = \frac{m-1+\frac{n\lambda}{\sigma^2}}{n-m-2}$. Then it follows that:

$$\begin{aligned}\lambda &= \mathbb{E}\left(\frac{\text{SSB}}{\text{SSE}}\right)\frac{(n-m-2)}{n}\frac{\mathbb{E}(\text{SSE})}{n-m} - \frac{(m-1)}{n}\frac{\mathbb{E}(\text{SSE})}{n-m} \\ &\approx \mathbb{E}(\text{SSB})\frac{n-m-2}{n-m}\frac{1}{n} - \frac{(m-1)}{n}\frac{\mathbb{E}(\text{SSE})}{n-m}.\end{aligned} \tag{B.2}$$

where we make a first order Taylor approximation $\mathbb{E}\left(\frac{\text{SSB}}{\text{SSE}}\right) \approx \frac{\mathbb{E}(\text{SSB})}{\mathbb{E}(\text{SSE})}$. Though we have two different formulation of $\lambda$, as in Equation (B.1) and Equation (B.2), they

are mathematically and numerically very close. We can take the range of grids of $g$ on either one of them.

The original implementation of `MixTwice` takes a linear grid (equally spaced from the lower range to the upper range) for either $g$ and $h$. However, the effect size distribution $g$, a function of $\lambda = \frac{1}{n}\sum_{j=1}^{m}(\mu_j - \bar{\mu})^2$ takes quadratic spacing itself. Moreover, due to the quadratic functional of effect size, the grid spread (upper range) of $g$, determined previously, could be very large. Furthermore, even with a relative dense grid, there might not be enough supports close to zero to accurately estimate those point masses, which are indeed more important to influence the follow-up $\pi_0$ and local false discovery rate estimation. Therefore, we allow the grid of mixing distribution of $\lambda$, $g$ also to be quadratic spacing.

## B.2 Connection between MixTwice and MixTwice-ANOVA

Generalized from `MixTwice`, `MixTwice-ANOVA` deals with problem with multiple group comparison. They can be both applied on analyzing the data when the number of groups is 2. Indeed, in the unit-specific setting, when the number of groups is 2, the F statistic in one-way ANOVA is simply the square of t statistic in two sample t test. It is curious to understand the relationship in performance between `MixTwice` and `MixTwice-ANOVA` under such degenerated scenarios.

In this section we discuss this problem through a toy example with 3000 testing units on a two-group comparison problem with 20 replicates at each group. This toy data is applied on both `MixTwice` and `MixTwice-ANOVA`. As shown in Figure B.1, we compare the $\pi_0$ estimation using `MixTwice`, `MixTwice-ANOVA` with linearly-spaced grid and `MixTwice-ANOVA` with quadratically-spaced grid. The example is also examined under a variety of alternative distributions. To compare both results, we make the number of support points of `MixTwice-ANOVA` 20 except at exactly 0 (in total 21) and number of support points of `MixTwice` on each side also 20 except at exactly 0 (in total 41). It is clear that `MixTwice` and `MixTwice-ANOVA` with quadratically-spaced grid yield equal $\hat{\pi}_0$ estimation while the `MixTwice-ANOVA` with linearly-spaced grid got much larger $\hat{\pi}_0$. Both methods would overestimate $\pi_0$ under the spiky alternative distribution (panel B) and this is consistent with the result reported in Zheng et al. [2021].

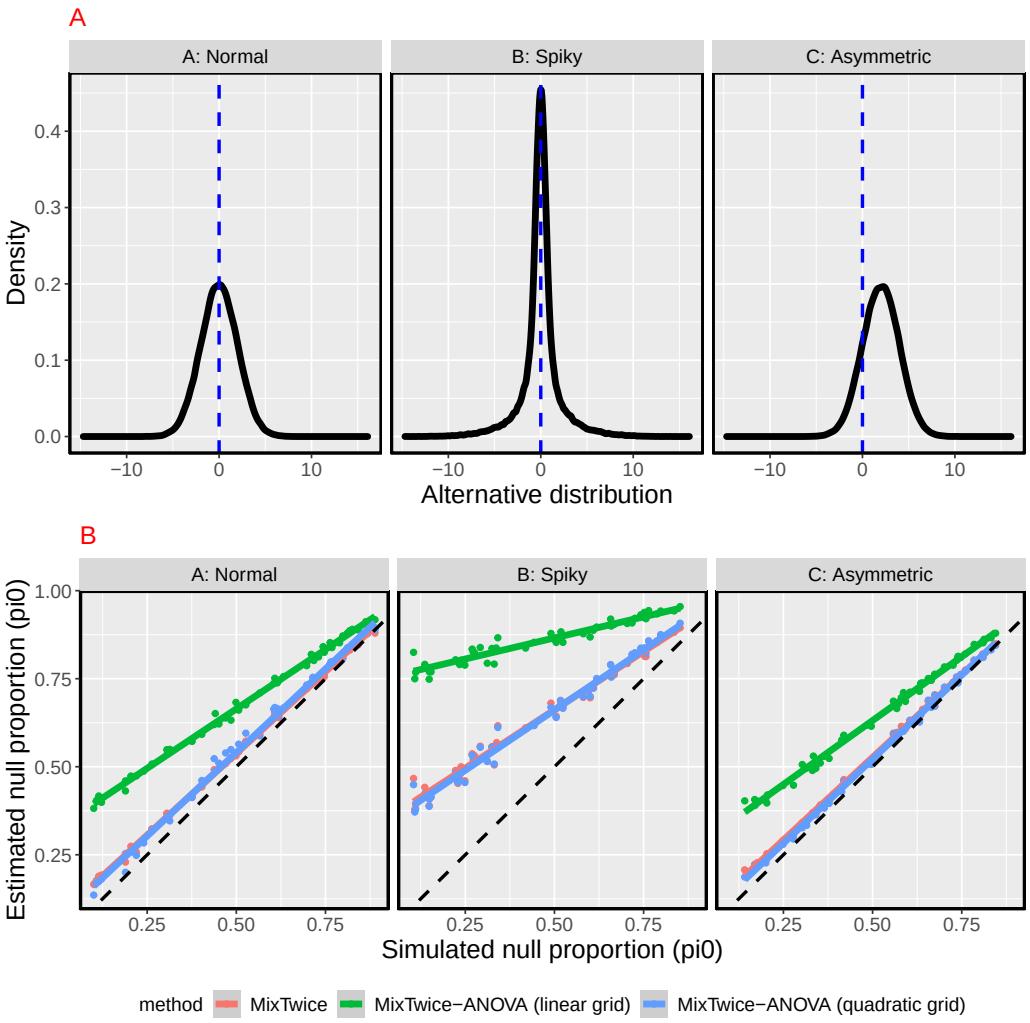


Figure B.1: $\pi_0$ **estimation for MixTwice and MixTwice-ANOVA:** the estimation of null proportion estimation ($\hat{\pi}_0$) is examined for MixTwice (red), MixTwice-ANOVA with linear grid (green) and MixTwice-ANOVA with quadratic grid (blue) under variety of settings of alternative distributions (panel A). Panel B shows scatter plots of $\hat{\pi}_0$ (y-axes) and $\pi_0$ (x-axes) with the dotted line as reference.

## B.3 Data example: CCP+RF+ RA vs CCP-RF- RA vs control

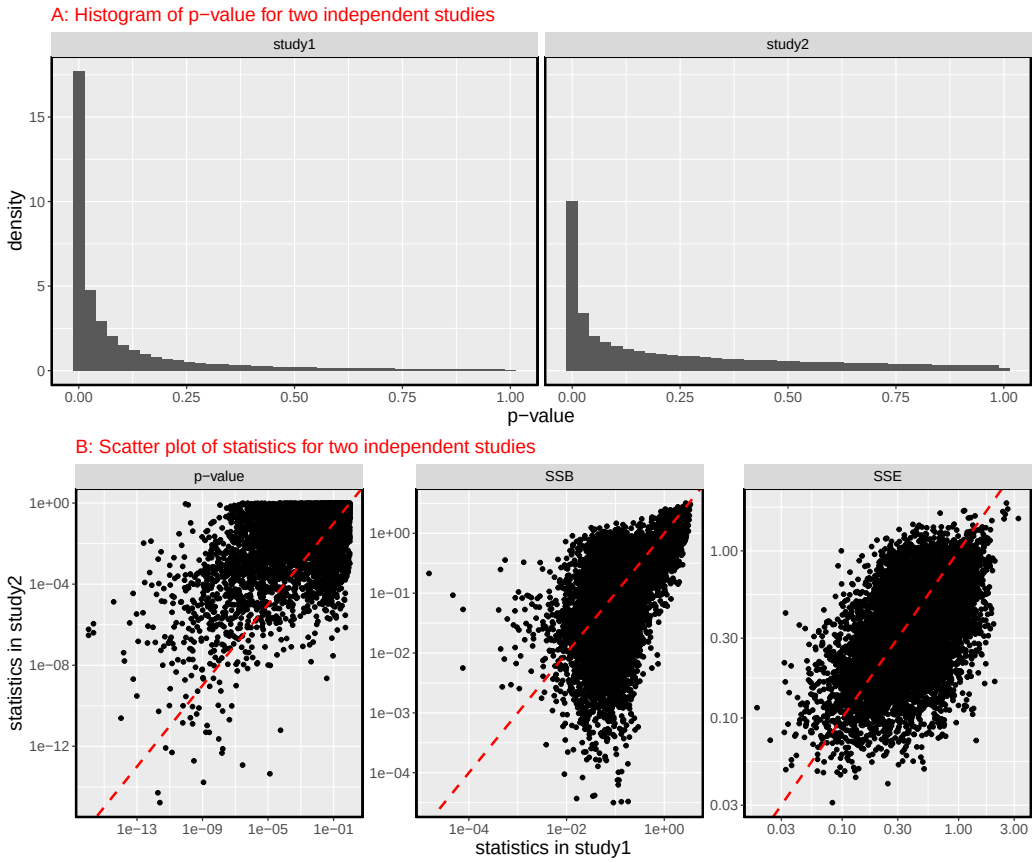


Figure B.2: **Summary and visualization figure for RA comparison in two independent peptide array studies:** Panel A shows the distribution of p-value in both studies. Panel B shows the scatter plot (and correlation) for three major statistics (from left to right, p-value, SSB and SSE) between two studies. The dashed line in red is reference $y = x$.

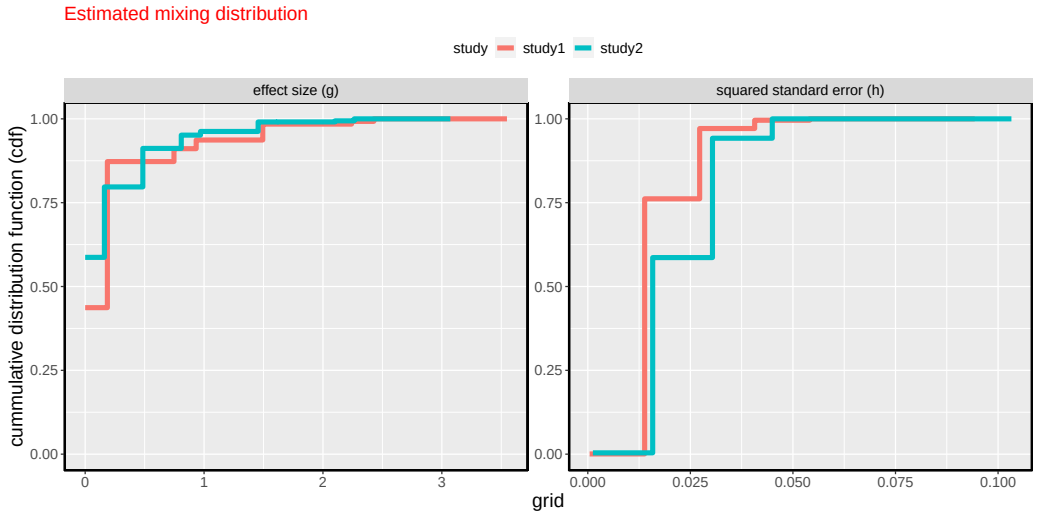


Figure B.3: **Estimated mixing distribution:** left panel shows the estimated cumulative distribution function (cdf) of effect size ($g$) and right panel shows the estimated cumulative distribution function of squared standard error ($h$). Red and blue are for different studies.

# Appendix C

# Appendix to Chapter 4

## C.1 Reproducibility toy example

Intuition behinds why methods using not only p-value but also estimated effect size and estimated error often yield higher reproducibility is also of great interest. Reproducibility measures the *correlation information* on the same testing unit but on independent replicates (for example, different subject samples in peptide array data). This *correlation information*, in a nutshell, could be reduced by the ratio calculation (from numerator statistic and denominator statistic to test statistic) and could also be further reduced by the sign calculation (from test statistic to p-value). The following simulation example provides a more concrete interpretation on why more data information per testing unit could increase the reproducibility performance. The simulation result is based on 100 repeated trials, each with a randomly simulated null proportion $\pi_0 \sim U(0,1)$. Within each trial, we simulate data with $p = 5000$ units and 20 samples. Those 20 samples are then evenly and randomly split to form two data sets $\mathbf{x}_j$, $j = 1, 2$. For each data, four statistics are calculated in pair $x_{i,j}, s_{i,j}, t_{i,j}, p_{i,j}$ where $i = 1, 2, \cdots, p$ and $j = 1, 2$. Correlation of each pair are calculated and plotted over $\pi_0$ in panel A in Figure C.1. Scatter plot for four statistics in a single trial, with $\pi_0 = 0.5$ is demonstrated as an example in panel B.

It is clear that the correlation of $x_i$ (red) is slightly higher compared to $t_i$ (orange), the ratio between $x_i$ and $s_i$ (blue) and is much higher than the sign-free statistic, $p_i$ (green). The correlation of $s_i$ is robust over $\pi_0$. However, the correlations are decreasing for $x_i, t_i, p_i$ with larger $\pi_0$ since it is less likely to differentiate among units with more nulls (signal is 0).

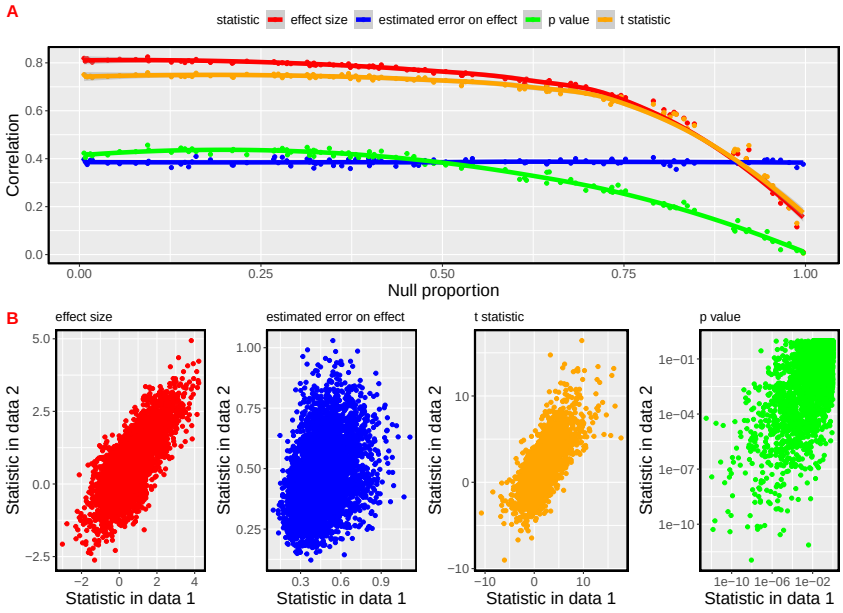


Figure C.1: **Why approaches using $x_i$ and $s_i$ are more reproducible compared to those using only $p_i$?** Panel A shows the relationship between null proportion and correlation for four different test statistics, and panel B picks a single trial with $\pi_0 = 0.5$ to demonstrate the scatter plot of each statistic in two independent data sets.

# Appendix D

# Appendix to Chapter 5

## D.1 Number of simulations of random graph

One important parameter when running `GraphicalT` is the number of simulations of random graph, specifically `Nsim` in the package. The following Table D.1 summarizes the simulation result with different number of random graph simulations (from 40 to 200, each replicated 30 times). The simulated data contains 400 nodes, half of them are isolated while the rest half construct 20 cliques, each with size 10. For all 400 nodes, 200 of them have differential binding between groups ($\pi_0 = 0.5$). Two operating characteristics: true positive rate and time complexity [1] are of interest and the following table reports both mean parameter and standard deviation of those two operating characteristics.

Table D.1: True positive rate and Time complexity with increased number of simulations of random graph

| Number of simulations | TPR (mean) | TPR (sd) | CPU time (mean) | CPU time (sd) |
|---|---|---|---|---|
| 40 | 0.802 | 0.0082 | 0.133 | 0.073 |
| 80 | 0.801 | 0.0065 | 0.252 | 0.071 |
| 120 | 0.803 | 0.0060 | 0.372 | 0.082 |
| 160 | 0.800 | 0.0046 | 0.416 | 0.075 |
| 200 | 0.802 | 0.0028 | 0.468 | 0.074 |

It is clear that the number of simulations of the random graph does not change the expected value of true positive rate, but it makes the result more robust (lower standard deviation) as the number of simulations increases. Also, it is not surprised that the number of simulations increases the CPU time. Though there is not a

[1] time complexity is CPU time evaluated with Inter® Core™ i5-7400HQ CPU processor

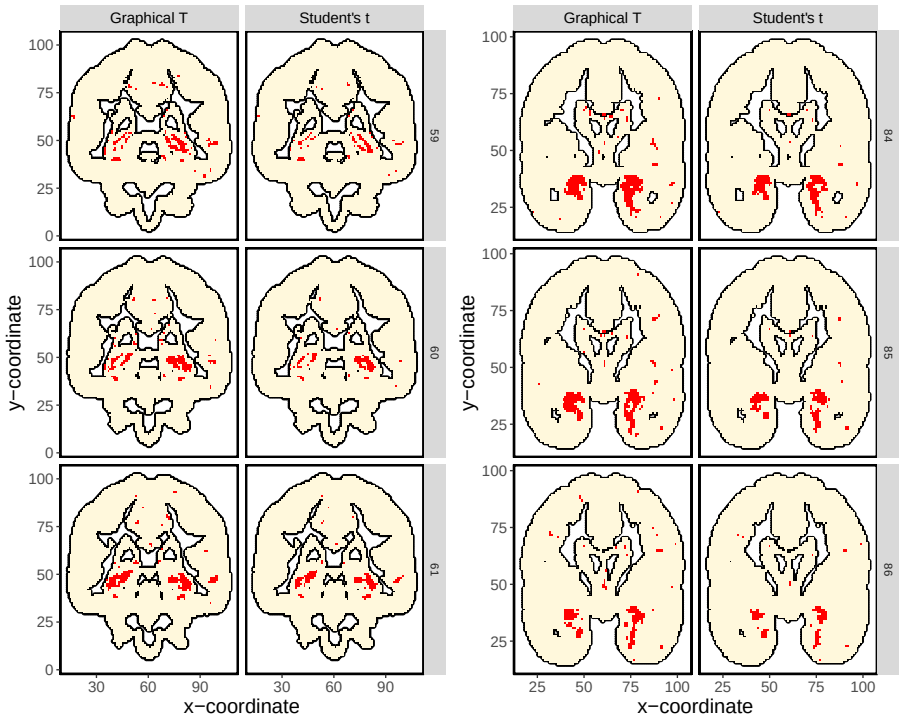


Figure D.1: **Examples of voxels on a few slices with locFDRs reported using GraphicalT and student's t:** Demonstration of the brain image on slices with spatial z-coordinates 59-61 (left panel) and 84-86 (right panel). Voxels with local false discoveries smaller than 0.05 from either GraphicalT (left columns) or student's t (right columns) were colored in red.

golden rule when determining the number of random graph simulations as it is a trade-off between the robustness of true positive rate and computational complexity, the user could make their own choice based on number of nodes and the expected complexity of the graph.

## D.2 Supplement to brain image data example

I also ran two examinations in order to check the FDR control, one on subject-sample randomization and the other on node-label randomization. The same `GraphicalT` algorithm was applied on the data with same testing nodes but the group label (MCI vs CN) was randomly permuted with the same number of subjects. It returned an empty list under the same FDR control and the locFDR's reported using the subject-randomized data set were uncorrelated with the locFDR's reported using

the original data set, as shown in Figure D.2. Another FDR control was done on the input graph $G_{\text{aux}}$. We randomly permuted the labels of testing nodes so that the input 3-D lattice, in theory, did not provide as much information about shared parameters as the original data set. By doing such node-permutation, the locFDR's in the binding step $\tilde{l}_e$ are much smaller (with mean 0.26) and the locFDR's were similar as methods without the input data graph (i.e., Student's t approach), as shown in Figure D.2.

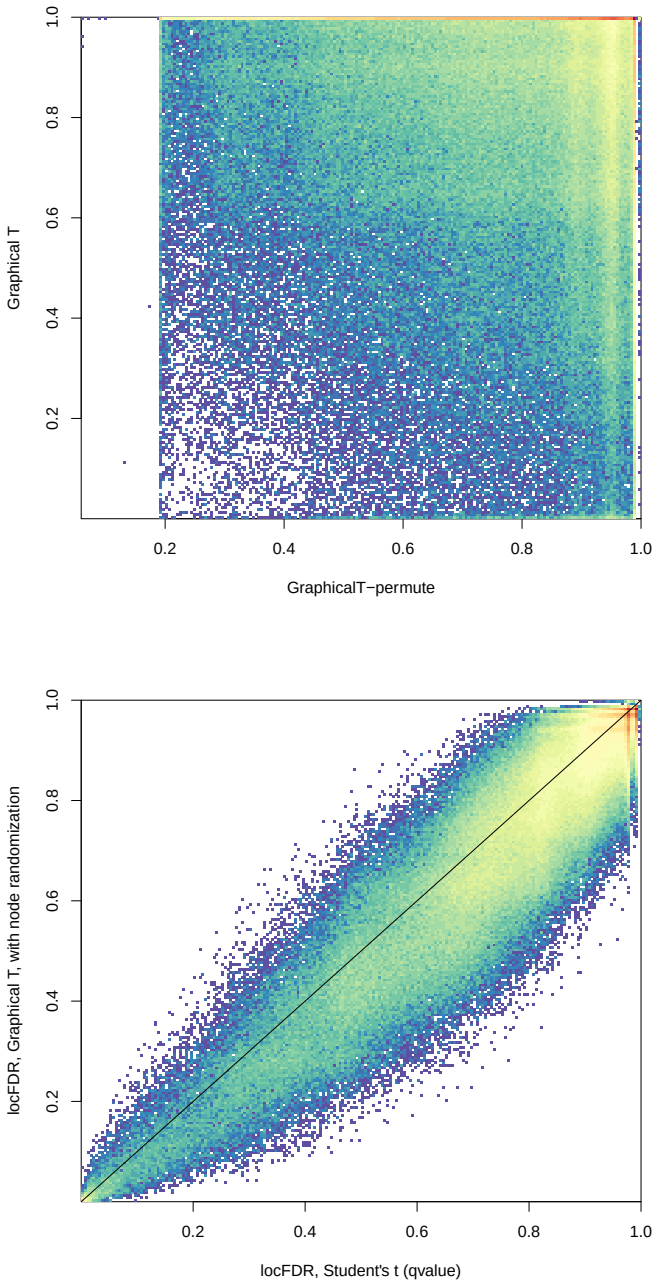


Figure D.2: **FDR control for brain image data:** Upper panel shows two dimensional histogram comparing locFDR's using GraphicalT on the original data set (y-axes) and on the data set with subject randomization (x-axes). The lower panel shows two dimensional histogram comparing locFDR's using Student's t without the graph-associated information (x-axes) and locFDR's using GraphicalT but with randomly permutated testing nodes (y-axes).

# Bibliography

B. A. Alberton, T. E. Nichols, H. R. Gamba, and A. M. Winkler. Multiple testing correction over contrasts for brain imaging. *NeuroImage*, 216:116760, 2020.

D. Aletaha, T. Neogi, A. J. Silman, J. Funovits, D. T. Felson, C. O. Bingham III, N. S. Birnbaum, G. R. Burmester, V. P. Bykerk, M. D. Cohen, et al. 2010 rheumatoid arthritis classification criteria: an american college of rheumatology/european league against rheumatism collaborative initiative. *Arthritis & rheumatism*, 62 (9):2569–2581, 2010.

B. Aragam, C. Dan, E. P. Xing, P. Ravikumar, et al. Identifiability of nonparametric mixture models and bayes optimal clustering. *Annals of Statistics*, 48(4):2277–2302, 2020.

M. Ayer, H. D. Brunk, G. M. Ewing, W. T. Reid, and E. Silverman. An empirical distribution function for sampling with incomplete information. *The annals of mathematical statistics*, pages 641–647, 1955.

A. L. Bailey, C. R. Buechler, D. R. Matson, E. J. Peterson, K. G. Brunner, M. S. Mohns, M. Breitbach, L. M. Stewart, A. J. Ericsen, C. M. Newman, et al. Pegivirus avoids immune recognition but does not attenuate acute-phase disease in a macaque model of hiv infection. *PLoS pathogens*, 13(10):e1006692, 2017.

J. A. Bailey, A. A. Berry, M. A. Travassos, A. Ouattara, S. Boudova, E. Y. Dotsey, A. Pike, C. G. Jacob, M. Adams, J. C. Tan, et al. Microarray analyses reveal strain-specific antibody responses to plasmodium falciparum apical membrane antigen 1 variants following natural infection and vaccination. *Scientific reports*, 10(1):1–12, 2020.

T. L. Bailey, M. Boden, F. A. Buske, M. Frith, C. E. Grant, L. Clementi, J. Ren, W. W. Li, and W. S. Noble. Meme suite: tools for motif discovery and searching. *Nucleic acids research*, 37(suppl_2):W202–W208, 2009.

P. Bajgrowicz and O. Scaillet. Technical trading revisited: False discoveries, persistence tests, and transaction costs. *Journal of Financial Economics*, 106(3): 473–491, 2012.

U. Båve, G. Nordmark, T. Lövgren, J. Rönnelid, S. Cajander, M.-L. Eloranta, G. V. Alm, and L. Rönnblom. Activation of the type i interferon system in primary sjögren's syndrome: a possible etiopathogenic mechanism. *Arthritis & Rheumatism*, 52(4):1185–1195, 2005.

Y. Benjamini and Y. Hochberg. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal statistical society: series B (Methodological)*, 57(1):289–300, 1995.

S. M. Boca and J. T. Leek. A direct approach to estimating false discovery rates conditional on covariates. *PeerJ*, 6:e6035, 2018.

C. Bonferroni. Teoria statistica delle classi e calcolo delle probabilita. *Pubblicazioni del R Istituto Superiore di Scienze Economiche e Commericiali di Firenze*, 8:3–62, 1936.

R. Bourgon, R. Gentleman, and W. Huber. Independent filtering increases detection power for high-throughput experiments. *Proceedings of the National Academy of Sciences*, 107(21):9546–9551, 2010.

P. Brito Zerón, N. Acar Denizli, W. F. Ng, M. Zeher, A. Rasmussen, T. Mandl, R. Seror, L. Xiaolin, C. Baldini, J. Gottenberg, et al. How immunological profle drives clinical phenotype of primary sjögren's syndrome at diagnosis: analysis of 10,500 patients (sjögren big data project). *Clinical & Exper Rheumatology*, 2018.

H. Burkhardt, T. Koller, Å. Engström, K. S. Nandakumar, J. Turnay, H. G. Kraetsch, J. R. Kalden, and R. Holmdahl. Epitope-specific recognition of type ii collagen by rheumatoid arthritis antibodies is shared with recognition by antibodies that are arthritogenic in collagen-induced arthritis in the mouse. *Arthritis & Rheumatism*, 46(9):2339–2348, 2002.

T. T. Cai and W. Sun. Simultaneous testing of grouped hypotheses: Finding needles in multiple haystacks. *Journal of the American Statistical Association*, 104(488): 1467–1481, 2009.

R. Callaghan, A. Prabu, R. Allan, A. Clarke, N. Sutcliffe, Y. S. Pierre, C. Gordon, S. Bowman, and U. S. I. Group*. Direct healthcare costs and predictors of costs in patients with primary sjögren's syndrome. *Rheumatology*, 46(1):105–111, 2007.

G. Casella and R. L. Berger. *Statistical inference*. Cengage Learning, 2021.

G. Csardi and T. Nepusz. The igraph software package for complex network research. *InterJournal*, Complex Systems:1695, 2006. URL `https://igraph.org`.

C. S. de Paiva, C. M. Trujillo-Vargas, L. Schaefer, Z. Yu, R. A. Britton, and S. C. Pflugfelder. Differentially expressed gene pathways in the conjunctiva of sjögren syndrome keratoconjunctivitis sicca. *Frontiers in Immunology*, page 2862, 2021.

S. Dudoit, J. P. Shaffer, and J. C. Boldrick. Multiple hypothesis testing in microarray experiments. *Statistical Science*, 18(1):71–103, 2003.

B. Efron. *Large-scale inference: empirical Bayes methods for estimation, testing, and prediction*, volume 1. Cambridge University Press, 2012.

B. Efron, R. Tibshirani, J. D. Storey, and V. Tusher. Empirical bayes analysis of a microarray experiment. *Journal of the American statistical association*, 96(456): 1151–1160, 2001.

M. Gierliński, C. Cole, P. Schofield, N. J. Schurch, A. Sherstnev, V. Singh, N. Wrobel, K. Gharbi, G. Simpson, T. Owen-Hughes, et al. Statistical models for rna-seq data derived from a two-condition 48-replicate experiment. *Bioinformatics*, 31 (22):3625–3630, 2015.

P. Groeneboom and G. Jongbloed. Some developments in the theory of shape constrained inference. *Statistical Science*, 33(4):473–492, 2018.

J. He, Y. Jin, X. Zhang, Y. Zhou, R. Li, Y. Dai, X. Sun, J. Zhao, J. Guo, and Z. Li. Characteristics of germinal center-like structures in patients with sjögren's syndrome. *International Journal of Rheumatic Diseases*, 20(2):245–251, 2017.

N. C. Henderson and M. A. Newton. Making the cut: improved ranking and selection for large-scale inference. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 78(4):781–804, 2016.

A. Hoefges, A. K. Erbe-Gurel, S. J. McIlwain, A. S. Melby, A. Xu, N. Mathers, A. L. Rakhmilevich, J. A. Hank, C. Baniel, R. Pinapati, et al. Thousands of new

antigens are recognized in mice via endogenous antibodies after being cured of a b78 melanoma via immunotherapy, 2020.

D. W. Huang, B. T. Sherman, and R. A. Lempicki. Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists. *Nucleic acids research*, 37(1):1–13, 2009a.

D. W. Huang, B. T. Sherman, and R. A. Lempicki. Systematic and integrative analysis of large gene lists using david bioinformatics resources. *Nature protocols*, 4(1):44–57, 2009b.

Y.-T. Huang, T.-H. Lu, P.-L. Chou, and M.-Y. Weng. Diagnostic delay in patients with primary sjögren's syndrome: A population-based cohort study in taiwan. In *Healthcare*, page 363. Multidisciplinary Digital Publishing Institute, 2021.

N. Ignatiadis, B. Klaus, J. B. Zaugg, and W. Huber. Data-driven hypothesis weighting increases detection power in genome-scale multiple testing. *Nature methods*, 13(7):577–580, 2016.

K. Korthauer, P. K. Kimes, C. Duvallet, A. Reyes, A. Subramanian, M. Teng, C. Shukla, E. J. Alm, and S. C. Hicks. A practical guide to methods controlling false discoveries in computational biology. *Genome biology*, 20(1):1–21, 2019.

E. Larssen, C. Brede, A. Hjelle, A. B. Tjensvoll, K. B. Norheim, K. Bårdsen, K. Jonsdottir, P. Ruoff, R. Omdal, and M. M. Nilsen. Fatigue in primary sjögren's syndrome: a proteomic pilot study of cerebrospinal fluid. *SAGE open medicine*, 7: 2050312119850390, 2019.

L. Lei and W. Fithian. Adapt: an interactive procedure for multiple testing with side information. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 80(4):649–679, 2018.

D. Lendrem, S. Mitchell, P. McMeekin, S. Bowman, E. Price, C. T. Pease, P. Emery, J. Andrews, P. Lanyon, J. Hunter, et al. Health-related utility values of patients with primary sjögren's syndrome and its predictors. *Annals of the rheumatic diseases*, 73(7):1362–1368, 2014.

M. Li, Y. Qi, G. Wang, S. Bu, M. Chen, J. Yu, T. Luo, L. Meng, A. Dai, Y. Zhou, et al. Proteomic profiling of saliva reveals association of complement system with

primary sjögren's syndrome. *Immunity, Inflammation and Disease*, 9(4):1724–1739, 2021.

J. Liu, P. Peissig, C. Zhang, E. Burnside, C. McCarty, and D. Page. Graphical-model based multiple testing under dependence, with applications to genome-wide association studies. In *Uncertainty in artificial intelligence: proceedings of the... conference. Conference on Uncertainty in Artificial Intelligence*, volume 2012, page 511. NIH Public Access, 2012.

M. Lu and M. Stephens. Variance adaptive shrinkage (vash): flexible empirical bayes estimation of variances. *Bioinformatics*, 32(22):3428–3434, 2016.

M. Lu and M. Stephens. Empirical bayes estimation of normal means, accounting for uncertainty in estimated standard errors. *arXiv preprint arXiv:1901.10679*, 2019.

A. Marshall. Discussion on barlow and van zwet's paper. *Nonparametric Techniques in Statistical Inference*, 1969:174–176, 1970.

Y. Masaki and S. Sugai. Lymphoproliferative disorders in sjögren's syndrome. *Autoimmunity reviews*, 3(3):175–182, 2004.

A. M. Mergaert, Z. Zheng, M. F. Denny, M. F. Amjadi, S. J. Bashar, M. A. Newton, V. Malmström, C. Grönwall, S. S. McCoy, and M. A. Shelef. Rheumatoid factor and anti-modified protein antibody reactivities converge on igg epitopes. *Arthritis & Rheumatology*, 2022.

N. Mishra, A. Caciula, A. Price, R. Thakkar, J. Ng, L. V. Chauhan, K. Jain, X. Che, D. A. Espinosa, M. M. Cruz, et al. Diagnosis of zika virus infection by peptide array and enzyme-linked immunosorbent assay. *MBio*, 9(2), 2018.

M. Newton, P. Wang, and C. Kendziorski. Hierarchical mixture models for expression profiles. In *Bayesian inference for gene expression and proteomics*, chapter 2, pages 40–52. Cambridge University Press New York, 2006.

T. Nichols and S. Hayasaka. Controlling the familywise error rate in functional neuroimaging: a comparative review. *Statistical methods in medical research*, 12 (5):419–446, 2003.

B. O'Neill. Some useful moment results in sampling problems. *The American Statistician*, 68(4):282–296, 2014.

Y. Park, J. Lee, J. H. Koh, Y.-K. Sung, S. S. Lee, J. Y. Choe, S. C. Shim, J. M. Kim, S. R. Kwon, H. O. Kim, et al. Distinct clinical characteristics of anti-ro/ssa-negative primary sjögren's syndrome: data from a nationwide cohort for sjögren's syndrome in korea. *Clin Exp Rheumatol*, 37(Suppl 118):107–113, 2019.

R. Patel and A. Shahane. The epidemiology of sjögren's syndrome. *Clinical epidemiology*, 6:247, 2014.

E. S. Pearson. The analysis of variance in cases of non-normal variation. *Biometrika*, pages 114–133, 1931.

L. Qiao, C. Deng, Q. Wang, W. Zhang, Y. Fei, Y. Xu, Y. Zhao, and Y. Li. Serum clusterin and complement factor h may be biomarkers differentiate primary sjögren's syndrome with and without neuromyelitis optica spectrum disorder. *Frontiers in Immunology*, page 2527, 2019.

H. S. K. Relangi, G. Naidu, V. Sharma, M. Kumar, V. Dhir, S. K. Sharma, A. Sharma, R. W. Minz, and S. Jain. Association of immunological features with clinical manifestations in primary sjogren's syndrome: a single-center cross-sectional study. *Clinical and experimental medicine*, pages 1–8, 2021.

A. Ritchie, R. A. Vandermeulen, and C. Scott. Consistent estimation of identifiable nonparametric mixture models from grouped observations. *arXiv preprint arXiv:2006.07459*, 2020.

T. Robertson, F.T.Wright, and R.L.Dykstra. *Order Restricted Statistical Inference*. Wiley, 1988.

G. A. Schellekens, B. De Jong, F. Van den Hoogen, L. Van de Putte, W. J. van Venrooij, et al. Citrulline is an essential constituent of antigenic determinants recognized by rheumatoid arthritis-specific autoantibodies. *The Journal of clinical investigation*, 101(1):273–281, 1998.

J. G. Scott, R. C. Kelly, M. A. Smith, P. Zhou, and R. E. Kass. False discovery rate regression: an application to neural synchrony detection in primary visual cortex. *Journal of the American Statistical Association*, 110(510):459–471, 2015.

M. L. Sembler-Møller, D. Belstrøm, H. Locht, and A. M. L. Pedersen. Proteomics of saliva, plasma, and salivary gland tissue in sjögren's syndrome and non-sjögren

patients identify novel biomarker candidates. *Journal of Proteomics*, 225:103877, 2020.

J. P. Shaffer. Multiple hypothesis testing. *Annual review of psychology*, 46(1):561–584, 1995.

C. H. Shiboski, S. C. Shiboski, R. Seror, L. A. Criswell, M. Labetoulle, T. M. Lietman, A. Rasmussen, H. Scofield, C. Vitali, S. J. Bowman, et al. 2016 american college of rheumatology/european league against rheumatism classification criteria for primary sjögren's syndrome: a consensus and data-driven methodology involving three international patient cohorts. *Annals of the rheumatic diseases*, 76(1):9–16, 2017.

A. G. Singh, S. Singh, and E. L. Matteson. Rate, risk factors and causes of mortality in patients with sjögren's syndrome: a systematic review and meta-analysis of cohort studies. *Rheumatology*, 55(3):450–460, 2016.

J. Sokolove, D. S. Johnson, L. J. Lahey, C. A. Wagner, D. Cheng, G. M. Thiele, K. Michaud, H. Sayles, A. M. Reimold, L. Caplan, et al. Rheumatoid factor as a potentiator of anti–citrullinated protein antibody–mediated inflammation in rheumatoid arthritis. *Arthritis & rheumatology*, 66(4):813–821, 2014.

J. Steen, B. Forsström, P. Sahlström, V. Odowd, L. Israelsson, A. Krishnamurthy, S. Badreh, L. Mathsson Alm, J. Compson, D. Ramsköld, et al. Recognition of amino acid motifs, rather than specific proteins, by human plasma cell–derived monoclonal antibodies to posttranslationally modified proteins in rheumatoid arthritis. *Arthritis & Rheumatology*, 71(2):196–209, 2019.

M. Stephens. False discovery rates: a new deal. *Biostatistics*, 18(2):275–294, 2017.

J. D. Storey. A direct approach to false discovery rates. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 64(3):479–498, 2002.

J. D. Storey et al. The positive false discovery rate: a bayesian interpretation and the q-value. *The Annals of Statistics*, 31(6):2013–2035, 2003.

K. Strimmer. fdrtool: a versatile r package for estimating local and tail area-based false discovery rates. *Bioinformatics*, 24(12):1461–1462, 2008.

W. Sun and T. Cai. Large-scale multiple testing under dependence. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 71(2):393–424, 2009.

E. Szarka, P. Aradi, K. Huber, J. Pozsgay, L. Végh, A. Magyar, G. Gyulai, G. Nagy, B. Rojkovich, É. Kiss, et al. Affinity purification and comparative biosensor analysis of citrulline-peptide-specific antibodies in rheumatoid arthritis. *International Journal of Molecular Sciences*, 19(1):326, 2018.

R. Tokarz, T. Tagliafierro, A. Caciula, N. Mishra, R. Thakkar, L. V. Chauhan, S. Sameroff, S. Delaney, G. P. Wormser, A. Marques, et al. Identification of immunoreactive linear epitopes of borrelia miyamotoi. *Ticks and tick-borne diseases*, 11(1):101314, 2020.

R. Turner. Iso: Functions to perform isotonic regression. *R package version 0.0-18.1*, 2020a. URL `https://CRAN.R-project.org/package=Iso`.

R. Turner. The algorithm for calculating unimodal isotonic regression in iso. *For Iso version 0.0-18*, 2020b.

K. Van den Berge, C. Soneson, M. D. Robinson, and L. Clement. stager: a general stage-wise method for controlling the gene-level false discovery rate in differential expression and differential transcript usage. *Genome biology*, 18(1):1–14, 2017.

R. Varadhan. *alabama: Constrained Nonlinear Optimization*, 2015. URL `https://CRAN.R-project.org/package=alabama`. R package version 2015.3-1.

T. Vo, V. Ithapu, V. Singh, and M. A. Newton. Dimension constraints improve hypothesis testing for large-scale, graph-associated, brain-image data. *Biostatistics*, 2021.

E. Waaler. On the occurrence of a factor in human serum activating the specific agglutination of sheep blood corpuscles. *Acta Pathologica Microbiologica Scandinavica*, 17(2):172–188, 1940.

M. W. Weiner and D. P. Veitch. Introduction to special issue: overview of alzheimer's disease neuroimaging initiative. *Alzheimer's & Dementia*, 11(7):730–733, 2015.

Y. Yan, N. Sun, H. Wang, M. Kobayashi, J. J. Ladd, J. P. Long, K. C. Lo, J. Patel, E. Sullivan, T. Albert, et al. Whole genome–derived tiled peptide arrays detect prediagnostic autoantibody signatures in non–small-cell lung cancer. *Cancer research*, 79(7):1549–1557, 2019.

J. J. Yang, J. Li, L. Williams, and A. Buu. An efficient genome-wide association test for multivariate phenotypes based on the fisher combination function. *BMC bioinformatics*, 17(1):1–11, 2016.

V. Yazisiz, B. Aslan, F. Erbasan, İ. Uçar, T. S. Öğüt, and M. E. Terzioğlu. Clinical and serological characteristics of seronegative primary sjögren's syndrome: a comparative study. *Clinical Rheumatology*, 40(1):221–229, 2021.

Z. Zheng and M. Newton. *MixTwice: MixTwice–a Package for Large-Scale Hypothesis Testing*, 2022. URL `https://CRAN.R-project.org/package=MixTwice`. R package version 2.0.

Z. Zheng, A. M. Mergaert, L. M. Fahmy, M. Bawadekar, C. L. Holmes, I. M. Ong, A. J. Bridges, M. A. Newton, and M. A. Shelef. Disordered antigens and epitope overlap between anti–citrullinated protein antibodies and rheumatoid factor in rheumatoid arthritis. *Arthritis & Rheumatology*, 72(2):262–272, 2020.

Z. Zheng, A. M. Mergaert, I. M. Ong, M. A. Shelef, and M. A. Newton. Mixtwice: large-scale hypothesis testing for peptide arrays by variance mixing. *Bioinformatics*, 37(17):2637–2643, 2021.