UC Berkeley
UC Berkeley Electronic Theses and Dissertations

Title

Adaptive Multivariate Multiple Testing and Bayesian Nonparametric Clustering Analysis

Permalink

https://escholarship.org/uc/item/8v4360wq

Author

Yang, Qiaoyu

Publication Date

2021

Peer reviewed|Thesis/dissertation

eScholarship.org Powered by the California Digital Library
University of California

Adaptive Multivariate Multiple Testing and Bayesian Nonparametric Clustering Analysis

by

Chiao-Yu Yang

A dissertation submitted in partial satisfaction of the

requirements for the degree of

Doctor of Philosophy

in

Statistics

in the

Graduate Division

of the

University of California, Berkeley

Committee in charge:

Professor Michael I. Jordan, Co-chair
Professor Will Fithian, Co-chair
Professor Aditya Guntuboyina

Spring 2021

Adaptive Multivariate Multiple Testing and Bayesian Nonparametric Clustering Analysis

Copyright 2021
by
Chiao-Yu Yang

Abstract

Adaptive Multivariate Multiple Testing and Bayesian Nonparametric Clustering Analysis

by

Chiao-Yu Yang

Doctor of Philosophy in Statistics

University of California, Berkeley

Professor Michael I. Jordan, Co-chair

Professor Will Fithian, Co-chair

In the modern era, statistics has evolved to embrace the emergence of large scale data sets, which brought both challenges and opportunities. We present in this dissertation two topics on this theme. In the first chapter, we present an adaptive framework for multiple hypotheses testing where each of the hypothesis testing problems is multivariate or nonparametric. Our work gives a solution when the analyst lacks prior knowledge of an optimal test statistic, a scenario common in many scientific and engineering problems including large scale A/B testing, multi-phenotype genome-wide association studies (GWAS), and response curve analysis in high-throughput toxicology experiments. In the second chapter, we study the problem of Bayesian nonparametric (BNP) clustering, a methodology known for its flexibility and expressive power. We focus on the theoretical analysis of the clustering behavior of certain BNP models and show under a general setting how such methods behave in clustering, where we give quantitative lower bounds on the probability distribution of the number of clusters in the asymptotic regime.

Contents

Contents **i**

1 Methods in Multiple Hypotheses Testing **1**
1.1 Introduction 2
1.2 The Bag of Null Statistics Procedure 7
1.3 Data Experiment 17
1.4 Discussion 23

2 Clustering Inference in Bayesian Nonparametric Models **24**
2.1 Introduction 25
2.2 Preliminaries 26
2.3 A Quantitative Result for the Posterior of the DPMM 29
2.4 Extension to Pitman-Yor Process Mixture Models 37
2.5 Experiments 42
2.6 Discussion 45

Bibliography **48**

A Appendix **53**
A.1 FDR Control 53
A.2 Linear Multivariate Gaussian Testing 54

Acknowledgments

Throughout my time as a Ph.D. student, I have met many people who generously offered me help and guidance. First and foremost, I am deeply grateful to my advisors, Profs. Michael Jordan and Will Fithian.

As an advisor, Mike is extremely knowledgeable, not only in machine learning but also in many other fields, and he often quickly pointed out the link between my research problem and some other papers to a surprisingly remarkable degree. For example, there was once in my second year when I presented my plan to attack a problem to him, and upon hearing only a high-level description, he retrieved from memory a result from some probability paper published in the 1970s, where the paper has studied a very similar problem in one hidden corner. Besides his achievement and reputation, it is surprising how down-to-earth and caring he is as a mentor. He is always very cheering and supportive, by which I mean I really cannot recall even once where he does not behave in this way.

I am also very fortunate to have Will as my advisor. Will has taught me many important lessons, often through his personal behavior. As a researcher he is both creative and rigorous. As a person he is polite and considerate. It has generally been relaxing to meet him, after which we often have lunch together (prior to the pandemic) and would chat about a wide range of topics, which form some precious memories in graduate school. Besides those warm fireside chats, his consistent pursuit of being rational and precise has a deep impact on me, teaching me to avoid the easy path and do the right thing.

During my time in graduate school, I have met many fellow Ph.D. students, postdocs, and undergraduate students, and these interactions have become an essential part of graduate school that gives me both intellectual inspiration and mental support. In particular, I want to thank Lihua Lei, a true academic hero and supportive person among my peers. I also owe a lot to my collaborators, including Nhat Ho, Tamara Broderick, Trevor Campbell, Zitong Yang, Eric Xia, Hongyuan Cao, Yixin Wang, and Kevin Wang. The content of this dissertation is based on the joint work with Lihua Lei, Nhat Ho, Eric Xia, and both of my advisors.

The staff in the department and especially our graduate student advisor La Shana Porlaris have helped a lot in dealing with the logistics, and I am sincerely grateful for their timely help that makes our lives much easier.

I did not really know where life would take me when I first arrived in Berkeley in the summer of 2016. Compared to now, I was more ignorant at that point, so I am grateful to all my teachers and friends for making me a less ignorant person, and for their patience and tolerance along the journey. Life is such an interconnected Indra's net that if I have accomplished anything, however small, I believe that it would not

be possible without the support of my family, my teachers, and my friends. Thanks for making it happen, and I wish I will be helpful to others in the future.

Chapter 1

Methods in Multiple Hypotheses Testing

Our era has seen a massive explosion in data scale, where one gets to see data in the order of millions to billions in biology, medicine, and business, among others. Not surprisingly, just like hypothesis testing has been a central topic in classical statistical inference, multiple hypotheses testing has become an important topic in modern inferential problems.

Since the seminal work of [6], false discovery rate (FDR) control, the control of the expected proportion of false positives in multiple testing, has become popular in research and industry. In many fields where discoveries are more important or no strict constraint on having a single error is present, the FDR control criterion has been adopted by practitioners as the standard.

In the following, we propose a new adaptive framework, the Bag-Of-NUll-Statistic (BONuS) procedure, for multiple testing where each hypothesis testing problem is itself multivariate or nonparametric and we want to control the FDR. This procedure is an adaptive and interactive knockoff-type [5] method that helps improve the testing power while always controlling the false discovery rate (FDR) in finite samples. Contrary to procedures that start with a set of p-values, our method analyzes the entire data set to estimate an optimal likelihood ratio test statistic based on an estimated mixture over the alternative hypothesis but controls FDR in finite samples despite the adaptivity. We also propose the Double BONuS procedure as an extension that gives a robust gain in power and guards against bad choices of modeling assumption. This is a joint work with Lihua Lei, Nhat Ho, and Will Fithian.

1.1 Introduction

Multiple multivariate testing

In the literature of multiple testing, it is customary to begin with one p-value for each of n null hypotheses as the primitive input and then focus on designing or analyzing methods for processing them. In many scientific problems, however, each of the n experiments yields multivariate data, and it is unclear *a priori* how best to summarize each one with a univariate p-value. As a result, the ultimate power of the full procedure may depend much more on how the p-values are calculated than on what procedure we apply after calculating them. Typical examples of multivariate or nonparametric testing problems that we may encounter include large scale A/B testing, GWAS with multivariate phenotypes, and analysis of dose-response curves in high-throughput toxicology experiments. In such problems, as the dimension of each problem grows, an agnostic p-value transformation may yield little power unless an exceptionally strong signal is present.

In most multivariate hypothesis testing problems, there is no uniformly most powerful (UMP) test that is efficient against all alternatives. For example, the generalized likelihood ratio test (GLRT) searches over all possible directions in which the true parameter $\boldsymbol{\theta} \in \mathbb{R}^d$ might differ from some hypothesized $\boldsymbol{\theta}_0$, but is not asymptotically efficient against local alternatives in any given direction. In high-dimensional or infinite-dimensional settings, the power tradeoff between different possible alteratives becomes especially pressing: [28] shows that for a Gaussian shift experiment in a real Hilbert space, for any test there exists a finite-dimension subspace outside of which the power curve is essentially flat. The same problem exists in nonparametric goodness-of-fit testing, where methods like Pearson's χ^2 test, Neyman's smooth test, and the Kolmogorov–Smirnoff test all represent different compromises across the many different ways that the true distribution might differ from the hypothesized distribution [28].

In a single multivariate testing problem, we cannot avoid paying the price of agnosticism without prior knowledge of which alternatives are more likely to actually occur. By contrast, when testing many multivariate hypotheses at once, we can pool information across hypotheses to *learn* the requisite prior knowledge to craft a more powerful test for each hypothesis. This article proposes an interactive empirical Bayes testing framework that uses a partially masked version of the entire data set to jointly estimate a prior distribution over the alternative. Our method, which we call the *Bag of Null Statistics* (BONuS) procedure, controls the *false discovery rate* (FDR) criterion proposed by [6]: R is the number of rejections and V is the number of false rejections, the false discovery proportion (FDP) is defined as $V/(1 \vee R)$ and

the FDR is defined as its expectation, FDR $= \mathbb{E}\,[\text{FDP}]$. The BONuS procedure adaptively estimates an optimal sequence of nested rejection regions, selecting the largest region for which an estimator of the FDP is below a prespecified significance level α. It achieves robust finite-sample control of the FDR at level α whether or not the underlying empirical Bayes model is correctly specified.

To illustrate the cost of an inefficient agnostic test, we consider a rudimentary multivariate Gaussian simulation with

$$\boldsymbol{X}^{(i)} \overset{\text{ind.}}{\sim} \mathcal{N}_{10}(\boldsymbol{\theta}^{(i)}, I_{10}), \qquad \text{for } i = 1, \ldots, n = 10,000,$$

where we wish to test $\boldsymbol{\theta}^{(i)} = \mathbf{0}$ against $\boldsymbol{\theta}^{(i)} \neq \mathbf{0}$ for each i. We generate $n_1 = 500$ non-null statistics with mean parameters drawn independently from $\boldsymbol{\theta}^{(i)} \sim \mathcal{N}_{10}(0, 4\boldsymbol{v}\boldsymbol{v}')$, and the remaining $n_0 = n - n_1$ parameters are set to $\mathbf{0}$. In this problem the GLRT statistic is (equivalent to) $T_{\text{GLRT}}(\boldsymbol{X}^{(i)}) = \|\boldsymbol{X}^{(i)}\|_2^2$, while the Bayes-optimal test statistic is $T(\boldsymbol{X}^{(i)}) = (\boldsymbol{v}'\boldsymbol{X}^{(i)})^2$, which focuses all of its power in a single dimension of $\mathbb{R}^{10}$.

Figure 1.1 compares the single- and multiple-hypothesis testing power of three test statistics: the GLRT test, the oracle test, and an adaptive test statistic using an estimator $\widehat{\boldsymbol{v}}$ fitted on the full data set using EM-PCA [51]. Figure 1.1a shows the average power for a level-α hypothesis test on a new problem with parameter $\boldsymbol{\theta}^{(n+1)}$ drawn at random from the alternative. Even in a relatively low-dimensional setting with $d = 10$, we see that there are substantial power gains to be had by substituting the oracle test for the agnostic test, especially for small values of α. The adaptive estimate of the oracle test statistic, obtained by plugging in $\widehat{\boldsymbol{v}}$ for $\boldsymbol{v}$, nearly recovers the power of the oracle test. These differences are magnified in multiple testing, as shown in Figure 1.1b, where we compare the true discovery proportion of the Benjamini-Hochberg (BH) procedure [6] with the GLRT and oracle test statistics, as well as our BONuS procedure which also uses the plug-in estimator $\widehat{\boldsymbol{v}}$.

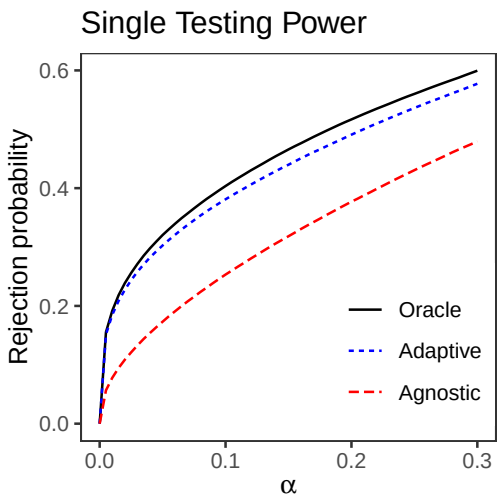
Multiple testing and the two-groups model

We consider testing the null hypothesis $H_0^{(i)} : \boldsymbol{\theta}^{(i)} = \mathbf{0}$ against $H_1^{(i)} : \boldsymbol{\theta}^{(i)} \neq \mathbf{0}$ in n independent experiments

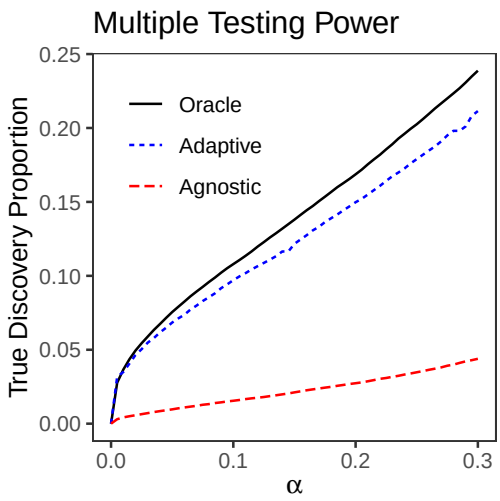
$$\boldsymbol{X}^{(i)} \overset{\text{ind.}}{\sim} f_{\boldsymbol{\theta}^{(i)}}(\boldsymbol{X}) \text{ for } i = 1, \ldots, n, \tag{1.1}$$

with possibly infinite-dimensional parameter $\boldsymbol{\theta}^{(i)} \in \Theta$. Let $\mathcal{H}_0 = \{i : H_0^{(i)} \text{ is true}\}$, and $n_0 = \#\mathcal{H}_0$, the number of true null hypotheses.

We assume throughout that we are testing a simple null against a composite alternative, but it is possible to extend the analysis to some problems with nuisance



(a) Average power for a single level-α hypothesis test using each of three test statistics: The GLRT statistic $T_{\text{GLRT}}(\boldsymbol{X}^{(i)}) = \|\boldsymbol{X}^{(i)}\|_2^2$, the oracle test statistic $T(\boldsymbol{X}^{(i)}) = (v'\boldsymbol{X}^{(i)})^2$, and an adaptive test statistic that estimates v.



(b) True discovery proportion for three FDR-controlling multiple testing procedures: the BH(α) procedure using the GLRT statistic, the BH(α) procedure using the oracle test statistic, and our BONuS procedure, which estimates v.

Figure 1.1: Motivating Gaussian example, where $\boldsymbol{\theta}^{(i)} \sim \mathcal{N}_{10}(0, 4vv')$ under the alternative. The Bayes-optimal oracle test dramatically outperforms the agnostic GLRT test, but requires prior knowledge of v. Pooling information across all n hypotheses allows estimation of v using EM-PCA on the full data set, nearly recovering the oracle performance.

parameters; for example we could take $\boldsymbol{X}^{(i)}$ to be a multivariate score statistic for the parameter of interest, calculated at a maximum likelihood estimate for the nuisance parameters. The data $\boldsymbol{X}^{(i)}$ may represent the entire data set for the ith experiment or a d-variate sufficient statistic summarizing it; we let $\mathcal{X}$ represent the sample space for a generic experiment and assume all distributions under consideration have densities with respect to a common measure μ on $\mathcal{X}$. Because the null hypothesis is simple, we can define a valid hypothesis test and calculate p-values using any (fixed) univariate transformation $T : \mathcal{X} \to \mathbb{R}$ as our test statistic, rejecting for large values of $T(\boldsymbol{X}^{(i)})$; likewise, we could test all $H_0^{(i)}$ using the BH, Storey-BH [54], or other multiple testing procedure that accepts independent p-values as input. We say another test statistic

$\widetilde{T}(\boldsymbol{x})$ is *monotonically equivalent* to $T(\boldsymbol{x})$ if it can be written as a strictly increasing function of $T(\boldsymbol{x})$; if $\widetilde{T}$ and T are monotonically equivalent then they yield identical p-values.

If we take a Bayesian perspective and assume that $\boldsymbol{\theta} \sim \Lambda$ under the alternative, then the test with highest average power rejects for large values of $\mathrm{LR}_\Lambda(\boldsymbol{X}) = \bar{f}_\Lambda(\boldsymbol{X})/f_{\mathbf{0}}(\boldsymbol{X})$, where the mixture density $\bar{f}_\Lambda(\boldsymbol{X}) = \int_\Theta f_{\boldsymbol{\theta}}(\boldsymbol{X}) d\,\Lambda(\boldsymbol{\theta})$ represents the marginal distribution of $\boldsymbol{X}$ under the alternative. By contrast, the GLRT rejects for large values of $T_{\mathrm{GLRT}}(\boldsymbol{X}) = \sup_{\boldsymbol{\theta} \in \Theta \backslash \{0\}} f_{\boldsymbol{\theta}}(\boldsymbol{X})/f_{\mathbf{0}}(\boldsymbol{X})$. If the prior Λ is relatively concentrated around a lower-dimensional region of $\Theta\backslash\{0\}$ then the test based on LR_Λ may have much higher power, as illustrated in Figure 1.1, but we can only use it if we know Λ.

In real applications we typically have no access to Λ, but when we test many hypotheses simultaneously, we can hope to reap many of the gains by jointly estimating Λ in a hierarchical Bayesian working model. Defining the Bernoulli indicator $H^{(i)} = 0$ if $H_0^{(i)}$ is true and 1 otherwise, we may introduce a version of the *two-groups model* of [9] as follows:

$$\begin{aligned} H^{(i)} &\overset{\text{i.i.d.}}{\sim} \mathrm{Bern}(1-\pi_0) \\ \boldsymbol{\theta}^{(i)} \mid H^{(i)} = 1 &\overset{\text{i.i.d.}}{\sim} \Lambda \\ \boldsymbol{X}^{(i)} \mid \boldsymbol{\theta}^{(1)}, \ldots, \boldsymbol{\theta}^{(n)} &\overset{\text{ind.}}{\sim} f_{\boldsymbol{\theta}^{(i)}}(\boldsymbol{x}). \end{aligned} \tag{1.2}$$

We emphasize here that (1.2) is merely a "working model" in the sense that Λ need not exist at all for our methods to control FDR: finite-sample control is guaranteed under the *fixed effects model* (1.1) where $\boldsymbol{\theta}^{(1)}, \ldots, \boldsymbol{\theta}^{(n)}$ are arbitrary fixed values. Because (1.1) can be obtained by conditioning on the latent parameters in (1.2), the tower rule implies that FDR control is also marginally guaranteed under the working two-groups model for any π_0 and Λ; in particular, FDR is controlled regardless of whether the analyst specifies a correct model for Λ.

Under the working model, the data follow a closely related mixture density with

$$\boldsymbol{X}^{(i)} \overset{\text{i.i.d.}}{\sim} f_{\mathrm{mix}}(\boldsymbol{x}) = \pi_0 f_{\mathbf{0}}(\boldsymbol{x}) + (1-\pi_0)\bar{f}_\Lambda(\boldsymbol{x}), \tag{1.3}$$

and the posterior probability that $H_0^{(i)}$ is true, called the *local FDR* or lfdr, is given by

$$\mathrm{lfdr}_{\pi_0,\Lambda}(\boldsymbol{x}) \;=\; \mathbb{P}_{\pi_0,\Lambda}\left(H^{(i)} = 0 \mid \boldsymbol{X}^{(i)} = \boldsymbol{x}\right) \;=\; \frac{\pi_0 f_{\mathbf{0}}(\boldsymbol{x})}{f_{\mathrm{mix}}(\boldsymbol{x})} \;=\; \left(1 + \frac{1-\pi_0}{\pi_0} \cdot \mathrm{LR}_\Lambda(\boldsymbol{x})\right)^{-1}.$$

Thus, from either a Bayesian or frequentist perspective, any optimal decision rule should reject for large values of $f_{\mathrm{mix}}(\boldsymbol{x})/f_{\mathbf{0}}(\boldsymbol{x})$, the ratio of the observable mixture

density to the null density, which is always identifiable and monotonically equivalent to the likelihood ratio $\mathrm{LR}_\Lambda(\boldsymbol{x})$. In other words, optimal rejection regions are super-level sets of $f_{\mathrm{mix}}(\boldsymbol{x})/f_{\mathbf{0}}(\boldsymbol{x})$. Calculating LR_Λ or lfdr does require knowing π_0, but is not strictly necessary for testing.

A natural empirical Bayes idea is to estimate either Λ or f_{mix} directly from the data, calculate p-values with respect to the plug-in test statistic $\mathrm{LR}_{\widehat{\Lambda}}(\boldsymbol{x})$ or $\hat{f}_{\mathrm{mix}}(\boldsymbol{x})/f_0(\boldsymbol{x})$, and then use a method like BH to control the FDR. The main difficulty with this plan is how to account properly for its using the same data twice. If we implement it with no safeguards, we could very easily arrive at an anticonservative procedure, for example by overestimating $f_{\mathrm{mix}}(\boldsymbol{x})$ at the observed values of $\boldsymbol{X}^{(i)}$. Furthermore, expecting consistent estimation of Λ is highly dubious for several reasons: first, the space of priors over the alternative is very large, and the number of clearly discernible observations from the alternative is most often relatively small. Second, without strong assumptions it is very difficult to disentangle the proportion π_0 of exact nulls from the proportion of non-nulls with parameter values very close to $\mathbf{0}$. Finally, density de-convolution is a hard statistical problem even without these challenges. As a result, we should demand that any adaptive procedure robustly account for its own adaptivity without relying on consistent estimation or even correct specification of the prior Λ.

Related Work

As an adaptive procedure for multiple testing, BONuS is motivated by several papers on adaptive inference. In particular, the idea of creating synthetic controls is inspired by the knockoff procedure [5], where one constructs synthetic nulls mimicking the original covariance structure and use the synthetic controls as a natural way to provide FDR control. Both AdaPT [31] and STAR [32] perform adaptive inference in multiple testing and are closely related to the knockoff methods as well.

BONuS attempts to improve the power by using a better test statistic and the motivation comes from the empirical Bayes model introduced by [8]. [58] studied a synthetic control method with the same martingale structure as BONuS with a view toward analyzing the power of knockoffs for LASSO solution path in regression settings with i.i.d. Gaussian predictor variables; their algorithm amounts to a non-interactive version of BONuS for regression with i.i.d. design.

In BONuS, the objective is to adaptively learn the structure of the problem from the data and use the srtucture to construct a more powerful test statistics. Similar in spirit, many recent methodology papers in post-selective inference have explored the use of structural information to improve testing power when certain prior information is available. For example, [33, 31, 15] studied a common type of structure that comes

up in dosage response experiment and LASSO solution path, where the hypotheses are ranked in a order such that a hypothesis can be rejected only if its preceding hypotheses have been rejected. In gene expression data, [25, 48] studied another structure represented by a directed acyclic graph (DAG). In [34], a generalization of utilizing prior information is proposed. Finally, [30, 27] studied how to exploit covariates independent of p-values when they are available.

In applications of genome wide association studies (GWAS), there are many situations where researchers are interested in diseases related to multiple endophenotypes, which naturally motivates the study of quantitative trait loci (QTL) that have a joint impact on these endophenotypes. Following this motivation, practitioners proposed various methods for solving multivariate GWAS problems. For example, in [12], the authors used canonical correlation analysis to extract linear combinations of traits that explains the most correlation with the markers. Another approach is given by [43], where in testing the regression coefficients of genotypes for some quantitative phenotypes, the authors proposed to use multiple phenotypes jointly to test the coefficients, different from the traditional approach that adopts a T-test for each genotype-phenotype pair. There is also a principle component based dimension reduction method in multivariate GWAS [35]. However, the aforementioned methods are often nonadaptive and rely on strong modelling assumptions. There has also been some study of optimizing multivariate test statistics without much modeling assumption, such as [1] in high dimensional setting and [13] in the setting of nonparametric permutation testing. However, neither method achieves finite-sample FDR control.

1.2 The Bag of Null Statistics Procedure

Definition of the procedure

The BONuS procedure begins by generating a set of $\widetilde{n}$ synthetic controls drawn from the null distribution,

$$\widetilde{\boldsymbol{X}}^{(1)}, \ldots, \widetilde{\boldsymbol{X}}^{(\widetilde{n})} \overset{\text{i.i.d.}}{\sim} f_{\boldsymbol{0}}(\boldsymbol{x}),$$

and then hiding them among the real statistics $\boldsymbol{X}^{(1)}, \ldots, \boldsymbol{X}^{(n)}$, without revealing to the analyst which test statistics are real samples and which are synthetic nulls. Formally, the analyst observes a permuted version of the data set,

$$\boldsymbol{Z} \;=\; \left(\boldsymbol{Z}^{(1)}, \ldots, \boldsymbol{Z}^{(n+\widetilde{n})}\right) \;=\; \Pi\left(\boldsymbol{X}^{(1)}, \ldots, \boldsymbol{X}^{(n)}, \widetilde{\boldsymbol{X}}^{(1)}, \ldots, \widetilde{\boldsymbol{X}}^{(\widetilde{n})}\right),$$

where Π is a uniformly random permutation on $n_+ = n + \widetilde{n}$ elements. Equivalently, the analyst observes the pooled empirical distribution of synthetic null and real test

statistics. We will use the variable j to refer to indices of the permuted vector, so that if $j = \Pi(i)$ for $i \in [n]$, then $\boldsymbol{Z}^{(j)} = \boldsymbol{X}^{(i)}$, and likewise $\boldsymbol{Z}^{(j)} = \widetilde{\boldsymbol{X}}^{(i)}$ if $j = \Pi(i+n)$.

Under the working model (1.2), the permuted values $\boldsymbol{Z}^{(j)}$ are exchangeable (but not quite independent) with marginal distribution

$$\boldsymbol{Z}^{(j)} \sim \widetilde{f}_{\text{mix}}(\boldsymbol{x}) = \frac{\widetilde{n}}{n_+} f_{\boldsymbol{0}}(\boldsymbol{x}) + \frac{n}{n_+} f_{\text{mix}}(\boldsymbol{x}) = \frac{\widetilde{n} + \pi_0 n}{n_+} f_{\boldsymbol{0}}(\boldsymbol{x}) + \frac{(1-\pi_0)n}{n_+} \bar{f}_{\Lambda}(\boldsymbol{x}).$$

The BONuS method proceeds iteratively, gradually revealing more information to the analyst, who continually updates an estimator $\widehat{T} : \mathcal{X} \to \mathbb{R}$ of either $\text{LR}_{\Lambda}(\boldsymbol{x})$ or $f_{\text{mix}}(\boldsymbol{x})/f_{\boldsymbol{0}}(\boldsymbol{x})$. This estimator may be based on plugging in a parametric estimate for the prior Λ, or on estimating f_{mix} directly; our notation is meant to capture either. The analyst uses the evolving estimator $\widehat{T}$ to construct a shrinking sequence of candidate rejection regions $\mathcal{X} \supseteq \mathcal{R}_0 \supsetneq \mathcal{R}_1 \supsetneq \mathcal{R}_2 \supsetneq \cdots$. As soon as an estimator of FDP falls below a pre-specified significance level α, the analyst halts the procedure and rejects all null hypotheses $H^{(i)}$ for which $\boldsymbol{X}^{(i)}$ is in the current rejection region.

To formally define the procedure, it will be convenient to define the binary indicator $B^{(j)} = 1$ if $\boldsymbol{Z}^{(j)}$ is real (i.e., if $\Pi^{-1}(j) \leq n$) and $B^{(j)} = 0$ otherwise, and let $B(\mathcal{A}) = (B^{(j)} : \boldsymbol{Z}^{(j)} \in \mathcal{A})$ denote the real/synthetic identities for all observations in $\mathcal{A} \subseteq \mathcal{X}$. In addition define the counting processes

$$N(\mathcal{A}) = \#\{i : \boldsymbol{X}^{(i)} \in \mathcal{A}\}, \quad \text{and} \quad \widetilde{N}(\mathcal{A}) = \#\{i : \widetilde{\boldsymbol{X}}^{(i)} \in \mathcal{A}\},$$

representing respectively the number of real and synthetic observations in $\mathcal{A} \subseteq \mathcal{X}$.

At step $t = 0$, the analyst uses the permuted data to calculate an initial estimator $\widehat{T}_0(\boldsymbol{x}; \boldsymbol{Z})$, and an initial rejection region $\mathcal{R}_1 \subseteq \mathcal{X}$. $\mathcal{R}_1$ will typically be a super-level set of $\widehat{T}_0(\boldsymbol{x})$, which we should think of as a (random) function defined on $\mathcal{X}$. In the Storey-BONuS version of our procedure, the analyst also selects a *correction set* $\mathcal{A} \subseteq \mathcal{R}_1^{\mathsf{c}}$, typically a sub-level set of $\widehat{T}_0(\boldsymbol{x})$. At step $t = 1, \ldots, 1+n_+$, the analyst is allowed to observe $B(\mathcal{R}_t^{\mathsf{c}})$, "unmasking" the real/synthetic identities of all observations excluded from the current rejection region, and then calculates an estimator $\widehat{\text{FDP}}_t$ (defined below) for the rejection region $\mathcal{R}_t$. The analyst either halts the procedure or proposes a new candidate rejection region $\mathcal{R}_{t+1} \subsetneq \mathcal{R}_t$, typically $\mathcal{R}_t$ intersected with a super-level set of an updated estimator $\widehat{T}_t(\boldsymbol{x}; \boldsymbol{Z}, B(\mathcal{R}_t^{\mathsf{c}}))$. Figure 1.2 illustrates how information gradually accrues to the analyst as more observations are unmasked. We finally reject all $H_0^{(i)}$ with $\boldsymbol{X}^{(i)} \in \mathcal{R}_{\hat{t}}$, where

$$\hat{t} = \min \left\{ t \geq 0 : \widehat{\text{FDP}}_t \leq \alpha, \text{ or } N(\mathcal{R}_t) = 0 \right\}.$$

To ensure that the method terminates after at most $1 + n_+$ steps, we require that at least one new observation be excluded from the rejection region at every step.

This requirement is without loss of generality because the analyst observes no new information unless a new $B^{(i)}$ is revealed. Otherwise, there are no restrictions at all on how the analyst may choose $\mathcal{R}_{t+1}$, provided it depends only on $\boldsymbol{Z}$ and $B(\mathcal{R}_t^c)$.

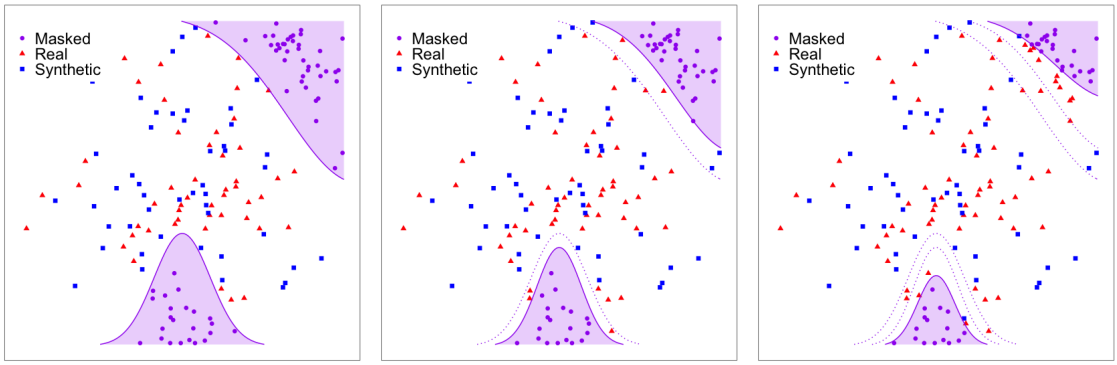


Figure 1.2: Schematic illustration of the BONuS procedure as the candidate rejection region shrinks. Each point represents either a two-dimensional test statistic for one of the real hypotheses (red triangles), or a synthetic test statistic generated from the null distribution (blue squares). The points in the rejection region are "masked" in the sense that the analyst does not observe whether each one is real or synthetic.

We consider two versions of the above procedure, the *BH-BONuS procedure* and the *Storey-BONuS procedure*, which respectively use the FDP estimators

$$\widehat{\mathrm{FDP}}_t^{\mathrm{BH}} = \frac{n}{\widetilde{n}+1} \cdot \frac{\widetilde{N}(\mathcal{R}_t)+1}{1 \vee N(\mathcal{R}_t)}, \quad \text{and} \quad \widehat{\mathrm{FDP}}_t^{\mathrm{St}} = \frac{N(\mathcal{A})+1}{\widetilde{N}(\mathcal{A})} \cdot \frac{\widetilde{N}(\mathcal{R}_t)+1}{1 \vee N(\mathcal{R}_t)},$$

If $\widetilde{N}(\mathcal{A}) = 0$ then $\widehat{\mathrm{FDP}}_t^{\mathrm{St}} = +\infty$; to avoid this occurring we should be sure to choose $\widetilde{n}$ and $\mathcal{A}$ large enough that $\mathbb{P}(\widetilde{N}(A) = 0) = \left(1 - \int_{\mathcal{A}} f_{\mathbf{0}}(\boldsymbol{x})\, d\boldsymbol{x}\right)^{\widetilde{n}} \approx 0$.

To understand the motivation for these estimators, note that $\widetilde{N}(\mathcal{R}_t)/\widetilde{n}$ may be viewed as an estimator of $N_0(\mathcal{R}_t)/n_0$, where $N_0(\mathcal{R}) = \{i \in \mathcal{H}_0 : \boldsymbol{X}^{(i)} \in \mathcal{R}_t\}$ is the number of false rejections we would make if we used $\mathcal{R}_t$ as our rejection set. As a result, for large rejection regions, we have

$$\widehat{\mathrm{FDP}}_t^{\mathrm{BH}} \approx \frac{n}{n_0} \cdot \frac{N_0(\mathcal{R}_t)}{1 \vee N(\mathcal{R}_t)} = \frac{n}{n_0} \mathrm{FDP}(\mathcal{R}_t).$$

The extra factor of n/n_0 makes the BH-BONuS procedure conservative; the Storey-BONuS procedure attempts to adjust for this using the correction set. If $\mathcal{A}$ is a

region where the test statistic is small, we might hope that $N(\mathcal{A}) \approx N_0(\mathcal{A})$, in which case $\widehat{\mathrm{FDP}}_t^{\mathrm{St}} \approx \mathrm{FDP}(\mathcal{R}_t)$.

We summarize the Storey-BONuS procedure with Algorithm 1.

Algorithm 1: The Storey-BONuS procedure

Input : the permuted data set $\boldsymbol{Z}$ of n real and $\widetilde{n}$ synthetic statistics
Output: The rejection region
select a correction set $\mathcal{A}$;
select $\mathcal{R}_1 \subset \mathcal{A}^{\mathsf{c}}$, let $t = 1$ and compute $\widehat{\mathrm{FDP}}_t(\mathcal{R}_t; \mathcal{A})$;
while $\widehat{\mathrm{FDP}}(\mathcal{R}_t; \mathcal{A}) > \alpha$ *and* $N(\mathcal{R}_t) \neq 0$ **do**
  increase t by 1;
  update $\mathcal{R}_t$ based on $\mathcal{F}_{t-1}$ such that $\mathcal{R}_t \subsetneq \mathcal{R}_{t-1}$;
  compute $\widehat{\mathrm{FDP}}(\mathcal{R}_t; \mathcal{A})$ and reveal $B(\mathcal{R}_t^{\mathsf{c}})$;
end
return $\mathcal{R}_t$.

We note that for the BH-BONuS procedure we simply skip the first step of specifying $\mathcal{A}$ and its use in computing $\widehat{\mathrm{FDP}}$.

To prove finite-sample control we rely on an optional stopping argument formalizing the preceding logic. Both $\widehat{\mathrm{FDP}}_t^{\mathrm{BH}}$ and $\widehat{\mathrm{FDP}}_t^{\mathrm{St}}$ can be computed from $\boldsymbol{Z}$ and $B(\mathcal{R}_t^{\mathsf{c}})$, since

$$N(\mathcal{R}_t) \;= n \;- \sum_{\boldsymbol{Z}^{(j)} \in \mathcal{R}_t^{\mathsf{c}}} B^{(j)}, \quad \text{and} \quad N(\mathcal{A}) \;= \sum_{\boldsymbol{Z}^{(j)} \in \mathcal{A} \subseteq \mathcal{R}_t^{\mathsf{c}}} B^{(j)},$$

and likewise for $\widetilde{N}(\mathcal{R}_t)$ and $\widetilde{N}(\mathcal{A})$ replacing n with $\widetilde{n}$ and $B^{(j)}$ with $1 - B^{(j)}$. As a result, if we use either estimator then $\hat{t}$ is a stopping time with respect to the filtration defined by $\mathcal{F}_t = \sigma\left(\boldsymbol{Z}, B(\mathcal{R}_t^{\mathsf{c}})\right)$, the information available to the analyst at step t. We show next that both variants of our method control FDR in finite samples. Our results rely on a lemma regarding the expectations of two functions of a hypergeometric random variable:

Lemma 1. *Let* $V \sim \mathrm{Hypergeom}(a + b, a, k)$*, and define* $U = k - V$*. Then*

$$\mathbb{E}\left[\frac{V}{1+U}\right] \;\leq\; \frac{a}{1+b}, \quad \textit{and} \quad \mathbb{E}\left[\frac{V}{1+U} \cdot \frac{b-U}{1+a-V}\right] \;\leq\; 1. \tag{1.4}$$

The proof of Lemma 1 is given in Appendix A.1. The inequalities in (1.4) is a standard result in the literature used in [54], [5], [58] and [30], but we reproduce the proof here for completeness.

Theorem 2. *Assume that the null test statistics* $(\boldsymbol{X}^{(i)} : i \in \mathcal{H}_0)$ *are drawn i.i.d. from* $f_{\mathbf{0}}$ *conditional on the non-null test statistics* $(\boldsymbol{X}^{(i)} : i \in \mathcal{H}_0^{\mathsf{c}})$. *Then the BH-BONuS procedure controls FDR at level* $\alpha n_0/n$ *and the Storey-BONuS procedure controls FDR at level* α.

Proof. For our optional stopping arguments, we will use the augmented filtration that also unmasks the identities of all real, non-null observations:

$$\mathcal{F}_t^+ = \sigma\left(\boldsymbol{Z}, B(\mathcal{R}_t^{\mathsf{c}}), (B^{(\Pi(i))} : i \in \mathcal{H}_0^{\mathsf{c}})\right).$$

We also define the σ-fields $\mathcal{F}_0 = \sigma(\boldsymbol{Z})$ and $\mathcal{F}_0^+ = \sigma\left(\boldsymbol{Z}, (B^{(\Pi(i))} : i \in \mathcal{H}_0^{\mathsf{c}})\right)$. Conditional on $\mathcal{F}_0^+$, the $n_0 + n_+$ unmasked observations are exchangeable, so that each of the $\binom{n_0+n_+}{n_0}$ configurations of their $B^{(j)}$ values is equally likely. Recall that $\mathcal{A}$ and $\mathcal{R}_1$ are data dependent subsets chosen by the analyst based on $\mathcal{F}_0 \subseteq \mathcal{F}_0^+$.

As the procedure unfolds, each time more $B^{(j)}$ values are unmasked, the remaining unmasked values remain conditionally exchangeable. As a result, defining $V_t = N_0(\mathcal{R}_t)$ and $U_t = \widetilde{N}(\mathcal{R}_t)$, and $V_0 = n_0, U_0 = \widetilde{n}$, we have for $t \geq 1$

$$V_t \mid \mathcal{F}_{t-1}^+ \sim \text{Hypergeom}(V_{t-1} + U_{t-1},\, V_t + U_t,\, V_{t-1}).$$

Note that $V_t + U_t$ is $\mathcal{F}_{t-1}^+$-measurable since the analyst chooses $\mathcal{R}_t$ knowing how many total observations are in $\mathcal{R}_{t-1} \setminus \mathcal{R}_t$ (or in $\mathcal{X} \setminus \mathcal{R}_1$, for $t = 1$). As a result, by the first inequality in Lemma 1 the quotient $V_t/(1 + U_t)$ is a super-martingale with respect to the filtration $\mathcal{F}^+ = (\mathcal{F}_t^+)_{t=0}^{1+n_+}$. Moreover, because both estimators $\widehat{\text{FDP}}_t^{\text{BH}}$ and $\widehat{\text{FDP}}_t^{\text{St}}$ are measurable with respect to $\mathcal{F}_t \subseteq \mathcal{F}_t^+$, $\hat{t}$ is a stopping time with respect to $\mathcal{F}^+$.

We are now ready to prove the result for the BH-BONuS method:

$$\begin{aligned}
\text{FDR} &= \mathbb{E}\Big[\frac{N_0(\mathcal{R}_{\hat{t}})}{1 \vee N(\mathcal{R}_{\hat{t}})}\Big] && (1.5)\\
&= \mathbb{E}\Big[\widehat{\text{FDP}}_{\hat{t}}^{\text{BH}} \cdot \frac{\tilde{n}+1}{n} \cdot \frac{N_0(\mathcal{R}_{\hat{t}})}{1 + \widetilde{N}(\mathcal{R}_{\hat{t}})}\Big] && (1.6)\\
&\leq \alpha \cdot \frac{\tilde{n}+1}{n} \cdot \mathbb{E}\Big[\frac{N_0(\mathcal{R}_{\hat{t}})}{1 + \widetilde{N}(\mathcal{R}_{\hat{t}})}\Big] && (1.7)\\
&\leq \alpha \cdot \frac{\tilde{n}+1}{n} \cdot \mathbb{E}\Big[\frac{N_0(\mathcal{X})}{1 + \widetilde{N}(\mathcal{X})}\Big] && (1.8)\\
&= \alpha \cdot \frac{n_0}{n}. && (1.9)
\end{aligned}$$

The inequality in (1.7) follows from the fact that either $\widehat{\mathrm{FDP}}_{\hat{t}}^{\mathrm{BH}} \leq \alpha$ or $N_0(\mathcal{R}_{\hat{t}}) = N(\mathcal{R}_{\hat{t}}) = 0$. The inequality in (1.8) follows from the optional stopping theorem.

Next we consider the Storey-BONuS method:

$$\mathrm{FDR} = \mathbb{E}\Big[\frac{N_0(\mathcal{R}_{\hat{t}})}{1 \vee N(\mathcal{R}_{\hat{t}})}\Big] \tag{1.10}$$

$$= \mathbb{E}\Big[\widehat{\mathrm{FDP}}_{\hat{t}}^{\mathrm{St}} \cdot \frac{N_0(\mathcal{R}_{\hat{t}})}{1 + \widetilde{N}(\mathcal{R}_{\hat{t}})} \cdot \frac{\widetilde{N}(\mathcal{A})}{1 + N(\mathcal{A})}\Big] \tag{1.11}$$

$$\leq \alpha \cdot \mathbb{E}\Big[\frac{N_0(\mathcal{R}_{\hat{t}})}{1 + \widetilde{N}(\mathcal{R}_{\hat{t}})} \cdot \frac{\widetilde{N}(\mathcal{A})}{1 + N_0(\mathcal{A})}\Big] \tag{1.12}$$

$$\leq \alpha \cdot \mathbb{E}\Big[\frac{N_0(\mathcal{R}_1)}{1 + \widetilde{N}(\mathcal{R}_1)} \cdot \frac{\widetilde{N}(\mathcal{A})}{1 + N_0(\mathcal{A})}\Big] \tag{1.13}$$

$$\leq \alpha. \tag{1.14}$$

For the expectation in (1.11), we define the integrand as 0 if $\widetilde{N}(\mathcal{A}) = 0$; in that case it coincides with the integrand in (1.10) because $\widehat{\mathrm{FDP}}_t^{\mathrm{St}} = +\infty$ for all t so the method makes no rejections. The inequality in (1.13) follows from the optional stopping theorem, applied conditional on $\mathcal{F}_1^+$ since $\widetilde{N}(\mathcal{A})$ and $N_0(\mathcal{A})$ are $\mathcal{F}_1^+$-measurable (but not $\mathcal{F}_0^+$-measurable).

To justify step (1.14), define $A = N_0(\mathcal{A} \cup \mathcal{R}_1)$ and $B = \widetilde{N}(\mathcal{A} \cup \mathcal{R}_1)$. Since the masked $B^{(j)}$ values in $\mathcal{A} \cup \mathcal{R}_1$ are exchangeable, we have

$$V_1 \mid \mathcal{F}_0^+, A, B \sim \mathrm{Hypergeom}(A + B, V_1 + U_1, A),$$

and we apply the second inequality in Lemma 1.

□

The martingales that appear in our method are very similar to those in Section 3 of [58]. With a different focus, their paper studies the power-FDR tradeoff of a knockoff procedure for linear models with i.i.d. Gaussian design and this martingale structure is used to calibrate the FDR for their knockoff procedure. Although our procedure has a very similar martingale structure and shares the use of FDR calibration with the null statistics, we use this martingale for a different purpose, namely, to adaptively use the data to design a better test statistic for multiple testing problems with a multivariate structure.

We emphasize once again that Theorem 2 is valid with respect to the fixed-effects model (1.1), which can be obtained by conditioning on $\boldsymbol{\theta}^{(1)}, \ldots, \boldsymbol{\theta}^{(n)}$ in the working

two-groups model (1.2). Hence, by the tower rule, FDR is controlled under (1.2) for any π_0 and Λ, whether or not they match whatever models for π_0 and Λ that may be implicitly or explicitly specified by the analyst.

Asymptotic Power

In the Bayesian two-groups model (1.2), the optimal test statistic is the likelihood ratio statistic $\bar{f}_\Lambda(\boldsymbol{x})/f_{\mathbf{0}}(\boldsymbol{x})$, or equivalently its more easily estimated proxy $T(\boldsymbol{x}) = f_{\text{mix}}(\boldsymbol{x})/f_0(\boldsymbol{x})$. We use $\hat{T}_n$ to denote the BONuS estimator of T when there are n real hypotheses and $\tilde{n}$ synthetic controls, where the $\widetilde{n}$ is omitted in the notation as one may always assume $\tilde{n}$ to be a fixed proportion of n. We define:

$$\epsilon_n(\boldsymbol{x}) := T(\boldsymbol{x}) - \hat{T}_n(\boldsymbol{x}),$$

and show in the following that asymptotically as $\epsilon_n(\boldsymbol{Z}^{(i)})$ shrinks, the BH-BONuS procedure can reproduce the same power as that of the oracle BH procedure. The same proof applies to the Storey-BH procedure as well.

To begin with, we denote $\mathcal{U} := \{T(\boldsymbol{x}) : \boldsymbol{x} \in \mathcal{X}\}$ and define the idealized BH procedure as one with the rejection threshold:

$$\eta^* = \min\Big\{\eta \in \mathcal{U} : \frac{F_{\mathbf{0}}(T(\boldsymbol{X}) > \eta)}{F_{\text{mix}}(T(\boldsymbol{X}) > \eta)} \le \alpha\Big\}, \tag{1.15}$$

where for simplicity we define $\min(\emptyset) = \infty$. The difference between the idealized BH procedure and the actual BH procedure is that in the actual BH procedure we need to estimate $F_{\text{mix}}(T(\boldsymbol{X}) > \eta)$, which give rises to an empirical rejection threshold:

$$\hat{\eta}_n^{\text{BH}} = \min_\eta \Big\{\eta \in \mathcal{U} : \lim_{n\to\infty} \frac{n\, F_{\mathbf{0}}(T(\boldsymbol{X}) > \eta)}{\sum_{i=1}^n 1\{T(\boldsymbol{X}^{(i)}) > \eta\}} \le \alpha\Big\}.$$

We observe that the term $\frac{\sum_{i=1}^n 1\{T(\boldsymbol{X}^{(i)})>\eta\}}{n}$ converges to $F_{\text{mix}}(T(\boldsymbol{X}) > \eta)$, but in the special situation where $T(\boldsymbol{X})$ is not continuous at η^* under F_{mix} and $\frac{F_{\mathbf{0}}(T(\boldsymbol{X})>\eta^*)}{F_{\text{mix}}(T(\boldsymbol{X})>\eta^*)}$ is exactly equal to α, the empirical rejection threshold would not converge to a constant asymptotically but would instead oscillate between η^* and a value that is more stringent than η^*. In this case, it may be inappropriate to define a constant asymptotic power for the BH procedure. To avoid this inconvenience, in the following theorem, for the α we choose and η^* as defined in equation (1.15), we assume that the quantity $\frac{F_{\mathbf{0}}(T(\boldsymbol{X})>\eta)}{F_{\text{mix}}(T(\boldsymbol{X})>\eta)}$ is less than α when $\eta > \eta^*$ and greater than α when $\eta < \eta^*$.

In addition, We note that for a discrete measure, there exist at most countably many $\alpha \in [0, 1]$ that would break the existence of a constant asymptotic power in

the BH procedure so this condition effectively does not impose any real concern in practice.

Theorem 3. *For a significance level α, if the BH procedure equipped with the oracle test statistic T has a corresponding idealized rejection threshold η^* as defined in equation (1.15), where $\frac{F_{\mathbf{0}}(T(\boldsymbol{X})>\eta)}{F_{\text{mix}}(T(\boldsymbol{X})>\eta)}$ is less than α when $\eta > \eta^*$ and greater than α when $\eta < \eta^*$, and if the BH-BONuS procedure estiamtes T by $\hat{T}_n$ such that for any $\delta > 0$ the proportion of $\boldsymbol{Z}^{(i)}$ with $|\epsilon_n(\boldsymbol{Z}^{(i)})| > \delta$ goes to 0 as $n \to \infty$, then the BH-BONuS procedure produces the same asymptotic power as that of the BH procedure with the oracle test.*

Proof. By the Glivenko-Cantelli theorem, $\frac{\sum_{i=1}^n 1\{T(\boldsymbol{X}^{(i)})>\eta\}}{n} \xrightarrow{a.s.} F_{\text{mix}}(T(\boldsymbol{X}) > \eta)$. Furthermore, since $\frac{F_{\mathbf{0}}(T(\boldsymbol{X})>\eta)}{F_{\text{mix}}(T(\boldsymbol{X})>\eta)}$ is less than α when $\eta > \eta^*$ and greater than α when $\eta < \eta^*$, asymptotically $\hat{\eta}_n^{\text{BH}}$ must converge to η^* almost surely. Hence, asymptotically the BH procedure rejects an alternative hypothesis $H^{(i)}$ if and only if $T(\boldsymbol{X}^{(i)}) > \eta^*$.

The rejection threshold of the BH-BONuS procedure is:

$$\hat{\eta}_n^{\text{BONuS}} = \min\Big\{\eta \in \mathcal{U} : \ \frac{n}{\widetilde{n}+1} \cdot \frac{1+\sum_{i=1}^{\widetilde{n}} 1\{\hat{T}_n(\widetilde{\boldsymbol{X}}^{(i)}) > \eta\}}{1 \vee \sum_{i=1}^{n} 1\{\hat{T}_n(\boldsymbol{X}^{(i)}) > \eta\}} \leq \alpha\Big\}.$$

If $\eta^* = \infty$, since the proportion of $\boldsymbol{Z}^{(i)}$ such that $|\epsilon_n(\boldsymbol{Z}^{(i)})| > \delta$ goes to 0 for any $\delta > 0$, as $n, \widetilde{n} \to \infty$, by Glivenko-Cantelli theorem, for any $\eta \in \mathcal{U}$, the term $\frac{n}{\widetilde{n}+1} \cdot \frac{1+\sum_{i=1}^{\widetilde{n}} 1\{\hat{T}_n(\widetilde{\boldsymbol{X}}^{(i)})>\eta\}}{1\vee\sum_{i=1}^{n} 1\{\hat{T}_n(\boldsymbol{X}^{(i)})>\eta\}}$ is larger than α so $\hat{\eta}_n^{\text{BONuS}} \xrightarrow{a.s} \infty$. Otherwise, if $\eta^* < \infty$, since the proportion of $\boldsymbol{Z}^{(i)}$ such that $|\epsilon_n(\boldsymbol{Z}^{(i)})| > \delta$ goes to 0 for any $\delta > 0$, the probability that $\hat{\eta}_n^{\text{BONuS}} \notin (\eta^* - \delta, \eta^* + \delta)$ goes to 0 as $n \to \infty$ so we conclude that $\hat{\eta}_n^{\text{BONuS}} \xrightarrow{P} \eta^*$.

In the case $\eta^* = \infty$, asymptotically no rejection will be made for any of the above procedures. Otherwise, we assume that $\eta^* < \infty$. Choose any $\gamma > 0$ and consider a hypothesis $H^{(i)}$ such that $T(\boldsymbol{X}^{(i)}) > \eta^* + \gamma$. This hypothesis $H^{(i)}$ is not rejected by the BH-BONuS procedure only if $\hat{T}_n(\boldsymbol{X}^{(i)}) \leq \hat{\eta}_n^{\text{BONuS}}$, which then implies:

$$\big|\epsilon_n(\boldsymbol{X}^{(i)})\big| > \eta^* - \hat{\eta}_n^{\text{BONuS}} + \gamma. \tag{1.16}$$

Since $\hat{\eta}_n^{\text{BONuS}} \xrightarrow{P} \eta^*$ and the proportion of $|\epsilon_n(\boldsymbol{X}^{(i)})| > \delta$ for any $\delta > 0$ goes to 0, it follows that the proportion of $\boldsymbol{X}^{(i)}$ satisfying (1.16) goes to 0 for any $\gamma > 0$. On the other hand, suppose $H^{(i)}$ is rejected but $T(\boldsymbol{X}^{(i)}) - \eta^* \leq \delta$, and note that by right continuity, as $\delta \to 0^+$, the fraction of hypotheses $H^{(i)}$ such that:

$$0 < T(\boldsymbol{X}^{(i)}) - \eta^* < \delta$$

must go to 0 in probability. Therefore, asymptotically the fraction of alternative hypothese rejected by the BH procedure with test statistic T but not by the BH-BONuS procedure goes to 0, and as a result they share the same asymptotic power. □

Corollary 4. *With the same condition as theorem 3 and a correction set $\mathcal{A}$ where $F_{\text{mix}}(\mathcal{A}) \in (0, 1)$, the Storey-BONuS procedure produces the same asymptotic power as that of the Storey-BH procedure with the oracle test statistic.*

The proof is almost identical and thus omitted.

Remark. A sufficient condition for theorem 3 is when both T and $\hat{T}_n$ are continuous functions on $\boldsymbol{x}$, and the total variation distance between $T(\boldsymbol{X})$ and $\hat{T}_n(\boldsymbol{X})$ goes to 0 as $n \to \infty$. Another condition is when $\boldsymbol{X}$ is a discrete variable, and the total variation distance between $T(\boldsymbol{X})$ and $\hat{T}_n(\boldsymbol{X})$ goes to 0.

In the proof, we implicitly assume that in BONuS, we shrink the rejection regions slowly enough such that there is no miss caused by rejecting overly conservative regions. In Appendix A.2, we discuss a theoretical example where a reliable estimator is available. However, in general, without making a parametric assumption, the density estimation problem remains unrealistic due to its high sampling complexity requirement. The practitioner may be more interested in deriving a test statistic that does not hurt the power when the modeling assumption is inappropriate and gain power when the assumption is appropriate. In the next subsection, we introduce an extension that allows us to do so in a robust way.

Double BONuS

The BONuS method guarantees FDR control while learning an adaptive test statistic from the data, but there is no guarantee that it will improve the power in any given example. In particular, if we specify an inappropriate empirical Bayes model or overfit the data, BONuS may underperform relative to a method that uses an agnostic test. As a result, it is appealing to have a way to try out several different competing models, perhaps including the agnostic method as a competitor, and assess which yields the best power. Even if we choose an appropriate model for the prior, our estimation method may involve tuning parameters that we will need to choose in a principled way. However, if we naively run many different variants of our method on the same data set and keep the one that makes the most rejections, we will violate the FDR control guarantee.

To allow for data-driven choices that make the procedure's power more robust, this section proposes a validation scheme for assessing the power gain of different

variants of the BONuS procedure, based on introducing a second group of synthetic nulls and running an initial stage of each BONuS variant. The method, which we call *Double BONuS*, has three steps:

1. Create a set $\widetilde{\boldsymbol{X}}$ of $\tilde{n}$ number of synthetic samples, and mix them with n real statistics to get a mixed sample $\boldsymbol{Z}$ with size and $n_+ = n + \tilde{n}$. Then, generate another set $\widetilde{\boldsymbol{Z}}$ of synthetic samples with size $\tilde{n}_+$.

2. Run each variant of BONuS on $\boldsymbol{Z}$ and $\widetilde{\boldsymbol{Z}}$, treating the first group as the "real" observations and the second group as the synthetic nulls.

3. Apply BONuS to $\boldsymbol{Z}$ with the variant that makes the most rejections in Step 2 (or a mixture of several competitive options from Step 2).

The above may be applied for either the Storey-BONuS or BH-BONuS method (or to choose between the BH-BONuS and Storey-BONuS methods). Using Double BONuS does not violate the FDR guarantee of Theorem 2 because the results of Step 2 are all $\mathcal{F}_0$-measurable, i.e. they can all be calculated from $\boldsymbol{Z}$ without knowing anything about which observations in $\boldsymbol{Z}$ correspond to real observations and which correspond to (the first group of) synthetic nulls. We recommend always including a non-adaptive, agnostic test as a competitor in case there is very little structure to find; this will tend to guard against the BONuS method actively harming the power relative to the BH or Storey-BH procedures. The Double BONuS method is defined formally in Algorithm 2.

Algorithm 2: The Double BONuS procedure

Input : $\boldsymbol{X}$, $\widetilde{n}$, $\widetilde{n}_+$, updating rules $\{\mathcal{M}_1, \cdots, \mathcal{M}_k\}$
Output: The rejection region
generate $\widetilde{n}$ null statistics and mix them with $\boldsymbol{X}$ to get $\boldsymbol{Z}$, and generate another $\widetilde{n}_+$ null statistics to be mixed with $\boldsymbol{Z}$ to get $\tilde{\boldsymbol{Z}}$;
LR **for** $i \in \{1, \cdots, k\}$ **do**
    Run algorithm 1 with $\boldsymbol{Z} \cup \tilde{\boldsymbol{Z}}$ as input and $\mathcal{M}_i$ as the updating rule for the rejection region, where $\boldsymbol{Z}$ is treated as real and $\tilde{\boldsymbol{Z}}$ is treated as synthetic in the computation of FDP estimator
end
analyze the results of updating rules to finalize a updating rule $\mathcal{M}^*$ and then run algorithm 1 with $\boldsymbol{Z}$ and use $\mathcal{M}^*$ to update the rejection region. Return the final rejection region.

Besides the additional computational cost from the additional modeling assumptions, there is no extra cost in the use of double BONuS. We note that the intention of double BONuS is to help the analyst select the most reasonable modeling assumptions among the several competitors, but not to exhaust the space of all possible modeling assumptions, where she can always find one, among the astronomical number of options, that generates high power in step 2 as a result of overfitting. In Section 1.3, we demonstrate the use of the double BONuS extension, where we also discuss strategies of finalizing $\mathcal{M}^*$ in Section 1.3.

1.3 Data Experiment

In this section we discuss implementation techniques, some simulation results, and a real data experiment.

Choosing the Number of Synthetic Controls

In choosing the number $\tilde{n}$ of synthetic nulls, the main tradeoff to consider is that larger values of $\widetilde{n}$ improve the accuracy of our FDP estimates, but also make it more difficult to estimate the alternative density $\bar{f}_\Lambda$: as $\widetilde{n} \to \infty$, the signal from the alternative hypotheses is lost in the sea of synthetic nulls. The other downside of picking a large $\tilde{n}$ is the additional computational cost, which may be burdensome in extremely large scale experiments.

In particular, for problems where one expects the final number of rejections to be very small, it is important to choose a value $\widetilde{n}$ large enough to mitigate the finite sample correction that arises in the computation of $\widehat{\text{FDP}}$. For example, in the BH-BONuS procedure, unless the procedure can reject at least $n/\alpha(\widetilde{n}+1)$ hypotheses, it cannot reject any at all; the minimum number of rejections for the Storey-BONuS method is roughly the same. As a result, picking $\widetilde{n}$ in the order of n/α is a natural choice in problems where there is a chance we will make very few rejections. Iterative techniques such as the EM algorithm may be employed in the case where we need to adopt a large $\tilde{n}$.

Although the analyst is certainly free to attempt any choice or even a multiple-layer BONuS that allows her to experiment with different numbers of synthetic controls, hereby we make a general recommendation as follows. First, we choose $\tilde{n}$ in the order of n/α, unless constraint in computational power or storage forbids us from doing so, in which case we may choose $\tilde{n}$ as large as possible. Then, we choose $\tilde{n}_+$, the number of second layer synthetic controls in double BONuS, to be any sufficiently large number. Note that $\tilde{n}_+$ is only used in the double BONuS phase where one wants

to compare different candidate models so its value does not need to match $n_+ = n+\tilde{n}$ in order. If the analyst expects the problem to come with many discoveries, she can adjust $\tilde{n}$ to the order of n instead of n/α.

Ensemble Method in Double BONuS

Double BONuS attempts to cherry pick the optimal model from a pool of candidates. However, there are often situations where none of the candidate models is exactly the same as the true model and several models pick up different subsets of alternative hypotheses. The traditional wisdom from ensemble learning instructs one to use multiple models' result, and we recommend a similar approach as well.

Specifically, at the stage of double BONuS and before we decide the model to be used in BONuS, we assess the performance of different candidate models and check if, besides the winning model's discoveries, there is a significant number of discoveries from other candidate models. For example, suppose the winning model makes 200 discoveries while another candidate model only makes 100 discoveries but many of these 100 discoveries are not in the set of 200 discoveries from the winning model, we may want to pick them up as well. Essentially, we want to combine the candidate models when they happen to detect different perspectives of the problem.

There is no consensus on how to use the multiple candidate models and combine their results, so we encourage the analyst to exploit domain knowledge to make an appropriate judgement. In addition, we recommend an approach based on p-values' ranking: among the candidate models, we can compute the (empirical) p-values of each of these models and use the minimum of them as the new p-value. If this new p-value, equivalent to a test statistic, turns out to perform even better in the double BONuS step, we can adopt it as the test statistic in the final step.

Multivariate Gaussian simulation

In this simulation, we use a sample size $n = 5,000$ with the number of true positives being $n_1 = 500$ and choose the dimension d to be 50. The distribution of the summary statistic $\boldsymbol{X}$ is:

$$\boldsymbol{X}^{(i)}|H^{(i)} = 0 \sim \mathcal{N}_d(\mathbf{0}, I_d) \tag{1.17}$$

$$\boldsymbol{X}^{(i)}|H^{(i)} = 1 \sim \mathcal{N}_d(\mathbf{0}, I_d + \boldsymbol{M}) \tag{1.18}$$

where $\boldsymbol{M}$ is a fixed matrix of rank 5. In our double BONuS experiment, we use a principal component analysis (PCA) approach with k principal components, where k ranges from 1 to d.

Besides the set of real statistics $\boldsymbol{X}$, we create a set $\widetilde{\boldsymbol{X}}$ of $\tilde{n} = n$ synthetic controls for BONuS and mix them to get $S = \boldsymbol{X} \cup \widetilde{\boldsymbol{X}}$. Then, we create another set $\widetilde{S}$ of $2n$ synthetic samples as the second layer synthetic nulls for double BONuS. We perform PCA with different choices of k on $S \cup \widetilde{S}$ to cherry pick the optimal model. The result, for a level 0.05 FDR control, is shown in Figure 1.3.

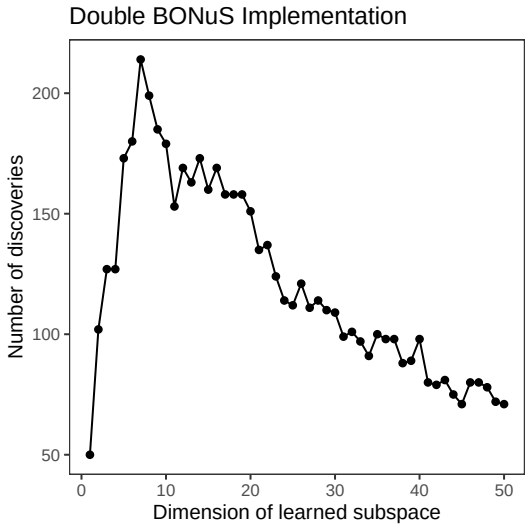


Figure 1.3: Numerical simulation for multivariate Gaussian with low rank signals, $\alpha = 0.05$

The optimal choice is the model with 7 components, where we attained 214 'discoveries' in S. After the screening, we pick this model, with which we apply BONuS to the mixture of $\boldsymbol{X}$ and $\widetilde{\boldsymbol{X}}$.

To see the effect, we also run the same set of data with different choices of α, ranging from 0.01 to 0.2. In Figure 1.4, we compare the false discovery proportion and empirical power of the three approaches: the agnostic (Chi-squared) test, the double BONuS procedure, and the oracle procedure. All three procedures used the same Storey correction, and the oracle procedure is computed with knowledge of M. As expected, the double BONuS procedure is able to capture much of the low rank structure in the problem and thus much more powerful than the agnostic test.

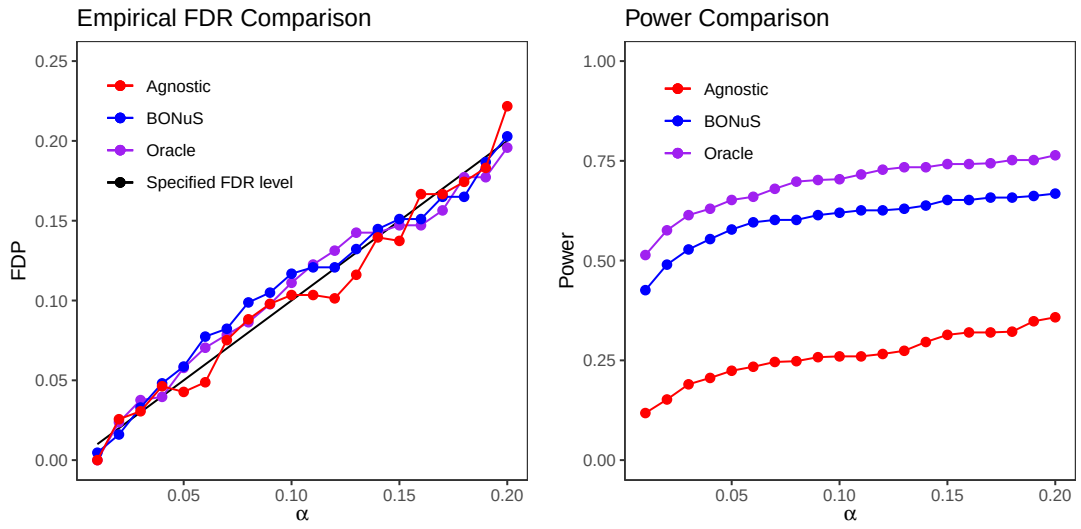


Figure 1.4: Comparison of false discovery proportions and powers of agnostic, double BONuS, and oracle procedure for the multivariate normal testing problem.

Multiple Multinomial Testing

BONuS is especially suitable for a high dimensional setting where an appropriate test statistic is typically not available and the cost of an agnostic test is huge. To illustrate this point, we run a simulation of multinomial tests with varying dimensions.

We consider the case where each $\boldsymbol{X}^{(i)}$ follows a multinomial distribution $\text{MultiNom}(N, \boldsymbol{\theta})$, with $\boldsymbol{\theta} = \boldsymbol{\theta}_0$ under the null and $\boldsymbol{\theta} = \boldsymbol{\theta}_1$ under the alternative, where we know $\boldsymbol{\theta}_0$ but not $\boldsymbol{\theta}_1$. For the simulation, we choose $\boldsymbol{\theta}_0$ to be $(1/d, \cdots, 1/d)$ and add normalized Rademacher perturbation to $\boldsymbol{\theta}_0$ to obtain $\boldsymbol{\theta}_1$, a setting from [4]. We repeat this simulation with 120 independent runs for each $d \in \{6, \cdots, 30\}$ and $N = 2000$, where in each run the number of hypotheses is $n = 3000$ and the number of alternative hypotheses is $n_1 = 300$. Note that we generate a new $\boldsymbol{\theta}_1$ for each run as well.

We compare BONuS with the oracle test and an agnostic test, where the oracle procedure assumes the knowledge of $\boldsymbol{\theta}_1$ and uses the corresponding likelihood ratio test statistic. For the agnostic test, we use the χ_2 test statistics and note that although it is suboptimal for multinomial test in general, but for the case of uniform null, the χ_2 test is equivalent to the truncated χ_2 test, which was shown to be minimax [4, 3]. We include the result in Figure 1.5.

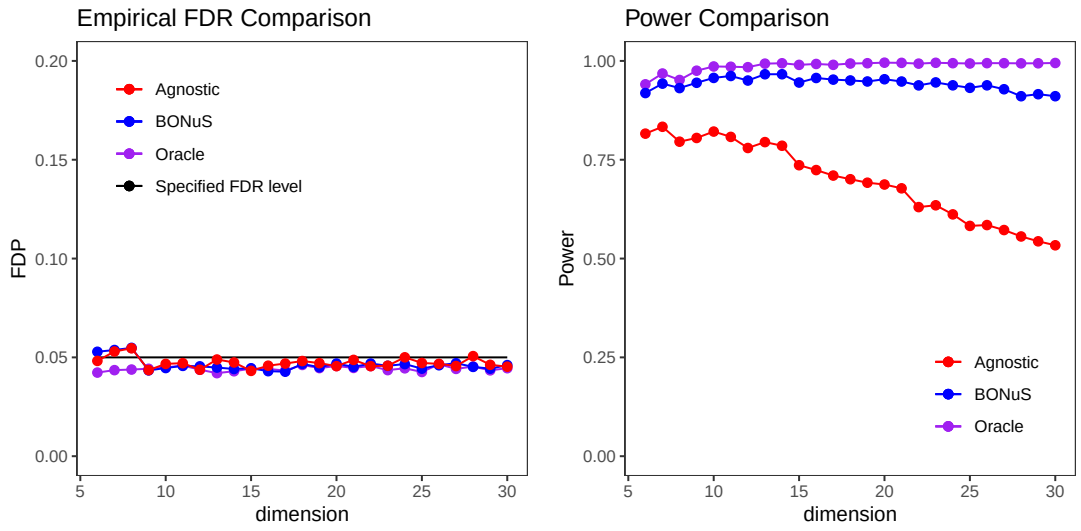


Figure 1.5: FDR and power of agnostic, double BONuS, and oracle procedures for the multinomial testing problem.

The Metabolic Syndrome GWAS

In many GWAS experiments, scientists are interested in identifying the SNPs related to certain diseases, whose severity can be characterized by multiple phenotypes. In this section, we apply BONuS to study the SNPs associated with metabolic syndrome, a problem studied by [35] with a different focus.

Metabolic syndromes refer to a medical condition found to be associated with a higher risk in cardiovascular disease and type-II diabetes. In this experiment, we want to identify the SNPs related to the metabolic syndrome, which in turn is linked to the following list of quantitative phenotypes: BMI, waist-hip ratio adjusted for BMI, high-density lipoprotein cholesterol (HDL), low-density lipoprotein cholesterol (LDL), Triglycerides (TG), fasting glucose, and fasting insulin. Although it is more likely for a SNP to be related to only one or few phenotypes, scientists speculate about detecting the SNPs with a weak effect on any single phenotype but a non-negligible joint effect.

For our experiment, we have z-scores for each of the phenotypes from several different medical research projects: [37, 53, 56, 39]. Since these projects study slightly different sets of SNPs, we choose to focus only on the intersection of SNPs in all studies.

Before running the experiment, we use LD-pruning, a method described in [47], to prune SNPs such that the remaining SNPs can be considered approximately independent under the null. After the preprocessing, there are about 1.8 million SNPs left.

One challenge is in the specification of the null. Under the null, a SNP is not associated with the metabolic syndrome so we expect its z-score vector to have mean 0. However, the z-scores of different phenotypes are correlated so the covariance matrix of the z-score vector is not the identity matrix: for example, the z-scores for SNPs in the fasting glucose study and the fasting insulin study have a 0.26 correlation coefficient. To create synthetic controls, we use a robust covariance matrix estimation on the z-score matrix $\boldsymbol{X}$, and then transform $\boldsymbol{X}$ to have a identity covariance matrix, after which we may assume that under the null, the z-score vector for a SNP follows a standard multivariate Gaussian distribution and thus create the synthetic controls.

In Figure 1.6, we show the number of discoveries as a function of α for both BONuS and agnostic approaches. One may observe that the power gain here is less than those in the previous experiments, but in this study one actually expects a large fraction of discoveries to be related to only one phenotype so it is the marginal gain that matters.

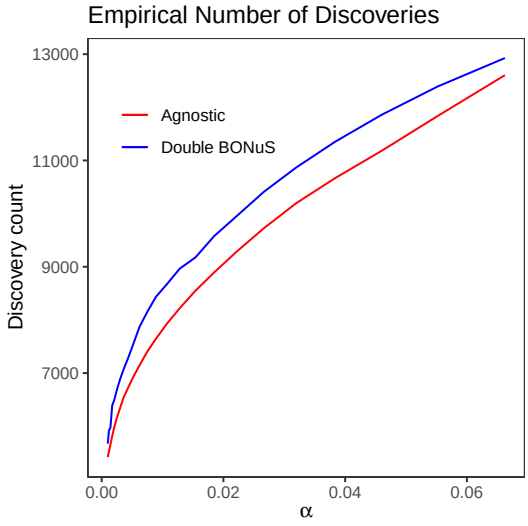


Figure 1.6: Metabolic syndrome GWAS experiment.

1.4 Discussion

We have presented the BONuS procedure as a novel method to estimate the oracle test when such a test is not known. BONuS attains this gain in power by adaptively and iteratively learning from the data in a way that secures the FDR control in finite samples. The secret that BONuS can simultaneously access the data while controlling FDR lies in the use of synthetic controls, which are mixed with the real statistics such that the analyst does not have full access to their identities before ending the procedure. This procedure is especially useful in high dimensional cases, where an agnostic test is likely to fail to capture the structure of the problem without strong prior knowledge.

Besides the BONuS procedure, we also presented its extension, the double BONuS procedure, as a way to robustly gain power by integrating multiple 'educated guesses' into the models and cherry pick the best one.

For future work, we discuss two directions that may be of interest. First, BONuS, like many other FDR controlling procedures, assumes that the nulls are independent conditioning on the alternatives. In regression problems and most economics or finance problems, such an assumption is not likely to be met. A procedure overcoming this constraint while not suffering from a major loss in power is highly desirable.

Second, in BONuS, we also assume the ability to generate independent synthetic controls. However, when the null distribution is composite, it remains unclear how to generate synthetic controls that are exchangeable with the original nulls.

Chapter 2

Clustering Inference in Bayesian Nonparametric Models

Classical parametric models have dominated statistical applications for a long time, where typically the problem is about studying a fixed set of data. However, in recent decades, we have seen new statistical problems where the underlying structure is no longer fixed but requires more dynamical modeling. Bayesian nonparametric (BNP) modeling refers to a set of nonparametric models originally designed to be more flexible and expressive, allowing a potentially infinite number of patterns in a continuously increasing dataset.

In this chapter we study the clustering behaviors of two most important BNP models, the Dirichlet process mixture model (DPMM) and the Pitman-Yor process mixture model (PYPMM). Dirichlet process mixture models and Pitman-Yor process mixture models have been widely used in density estimation, mode detection, and other nonparametric estimation problems. Due to their discrete, combinatorial nature of the underlying random measures, these models have also been suggested for inference in clustering problems, but it has been shown that these methods are inconsistent in the number of components. We contribute to this line of work by giving quantitative statements about the clustering behavior. Specifically, we provide nonasymptotic and asymptotic lower bounds on the ratio between the posterior probabilities of consecutive numbers of clusters in these models under different choices of prior for the parameters.

This is a joint work with Nhat Ho, Eric Xia, and Michael Jordan.

2.1 Introduction

In the wake of the seminal work of Ferguson [11, 7], a line of research on Bayesian nonparametric inference emerged that was based on the use of the Dirichlet process and related combinatorial stochastic processes, such as the Pitman-Yor process [46], as prior distributions. Most commonly, these priors are used in the context of mixture models, where both the prior and the posterior place probability mass on an unbounded number of mixture components [2].

Statistical applications for these models include a variety of problems in density estimation [10, 21, 17, 52, 22, 19, 18, 57, 20] and parameter estimation [55, 49, 14, 45]. Further applications can be found in engineering and scientific fields such as computer vision, information retrieval, economics, astronomy, molecular biology, and genetics [55, 49, 44, 10, 29, 26, 40]. In these latter applications, the problem is often one of clustering, and Bayesian nonparametric mixture models are viewed as a substitute for classical methods such as K-means, which generally require the number of components K to be known a priori. Particularly in the setting of dynamically growing data sets, it is difficult and even unreasonable to fix the number of clusters a priori, and Bayesian nonparametric mixture models have been considered as a natural way to circumvent this issue since they allow the number of clusters to be random and subject to posterior inference.

However, there is little theoretical support for this methodology. Unlike the case of density estimation, results on the convergence of the number of components in Bayesian nonparametric mixtures are largely absent from the literature. Indeed, it has been demonstrated that under certain parametric assumptions, these models exhibit posterior inconsistencies in the number of clusters when the underlying data-generating distribution satisfies some mild conditions [41, 42]. Moreover, in practice, it has been observed that inference based on Dirichlet process or Pitman-Yor process mixture models can generate small clusters that do not reflect the underlying data-generating process, a serious concern when the real number of components is small.

We study the posterior distribution of the number of clusters of Bayesian clustering models based on the Dirichlet process and the Pitman-Yor process. In contrast to previous work on this topic, we do not assume the underlying data-generating process to be be a specific parametric family or even a finite mixture model, so our results apply to nonparametric data-generating processes as well. In this general setting, there is no analytic form for the posterior distribution; nonetheless, we have obtained an analytical characterization of the ratio between the probability of obtaining $k+1$ clusters and that of obtaining k clusters, for any positive k. We provide novel lower bounds on these ratios, denoted $R(s|\{x_i\}_{i=1}^n)$, where $x_1, \ldots, x_n$ are exchangeable observations. We study the setting where the prior on the parameters is

either Gaussian or uniform. Our results are as follows:

- When the prior on the parameters is uniform over a bounded subset of $\mathbb{R}$, denoted Θ, the lower bound on $R(s|\{x_i\}_{i=1}^n)$ is of order $\frac{\alpha}{s|\Theta|}$, under a Dirichlet process mixture model, where α is the scale parameter of the Dirichlet process prior. In the case of a Pitman-Yor process mixture model, the lower bound becomes $\frac{\alpha}{s|\Theta|} + \frac{\xi}{|\Theta|}$ up to a universal constant, where ξ denotes the positive discount parameter. These bounds are nonasymptotic.

- When the prior on the parameters is Gaussian, $\mathcal{N}(0, \sigma^2)$, the asymptotic lower bound on $R(s|\{x_i\}_{i=1}^n)$ is of order $\frac{\alpha}{s^2(1+\sigma)}$ under a Dirichlet process mixture model and of order $\frac{\alpha}{s^2(1+\sigma)} + \frac{\xi}{s(1+\sigma)}$ under a Pitman-Yor process mixture model.

These lower bounds provide a fine-grained understanding of the posterior distribution induced by the Dirichlet process and the Pitman-Yor process on the number of clusters. We note, however, that in general there does not exist a finite upper bound. Such an upper bound, should it exist, would require additional assumptions on the data-generating process.

Besides the lower bounds we demonstrate in this paper, we note that one can use our techniques to obtain results for more specific cases where additional assumptions on the data-generating process are imposed.

The remainder of the paper is organized as follows. We begin in Section 2.2 with background on Dirichlet process and Pitman-Yor process mixture models. Section 2.3 is devoted to a study of the posterior distribution of the number of clusters from Dirichlet process mixture models under two choices of priors on the parameters: uniform and Gaussian priors. In Section 2.4, we extend these results to the case of Pitman-Yor process mixture models. Section 2.5 presents the results of experiments that explore some of the consequences of the theoretical results in previous sections. Finally, we conclude the paper with a discussion of potential future work in Section 2.6.

Notation: We use $\{x_i\}_{i=1}^n$ to denote the sample $\{x_1, \cdots, x_n\}$, $[n]$ to denote the set $\{1, \cdots, n\}$, and $A^{n,s} \in \rho_s(n)$ to denote a set $\{A_1^{n,s}, \cdots, A_s^{n,s}\}$ such that all of the elements A_i form an s-partition of $[n]$, where $\rho_s(n)$ is the set of all s-partitions on $[n]$. For example, $\{\{1, 2\}, \{3, 4\}\}$ is in $\rho_2(4)$.

2.2 Preliminaries

In this section, we provide necessary background on Dirichlet process mixture models and establish key notation for our subsequent analysis of the posterior distribution

of the number of clusters in these models.

Dirichlet process mixture models (DPMMs) [2, 36] are specified as follows:

$$p(A^{n,s}) := \frac{\alpha^s}{\alpha^{(n)}} \prod_{i=1}^{s} (|A_i^{n,s}| - 1)! \tag{2.1}$$

$$p(\theta|A^{n,s}) := \prod_{i=1}^{s} \pi(\theta_i) \tag{2.2}$$

$$p\big(\{x_i\}_{i=1}^n \big| \{\theta_j\}_{j=1}^s, A^{n,s}\big) := \prod_{j=1}^{s} \prod_{x_i \in A_j^{n,s}} f_{\theta_j}(x_i), \tag{2.3}$$

where π is a prior on the parameter θ and $\{f_\theta(\cdot)\}$ is a family of density functions.

In this paper, we focus on the application of DPMMs to clustering problems. The prior for the number of clusters with a data set of size n is as follows:

$$\mathbb{P}(K_n = s) = \sum_{A^{n,s} \in \rho_s(n)} p(A^{n,s}).$$

Given this prior distribution, the posterior distribution for the number of clusters admits the following formulation:

$$\begin{aligned}
&\mathbb{P}(K_n = s|\{x_i\}_{i=1}^n) \\
&= \frac{\mathbb{P}(\{x_i\}_{i=1}^n|K_n = s)\mathbb{P}(K_n = s)}{\mathbb{P}(\{x_i\}_{i=1}^n)} \\
&\propto \sum_{A \in \rho_s(n)} p(A^{n,s}) \cdot \int_{\{\theta_j\}_{j=1}^s} p(\{x_i\}_{i=1}^n|\{\theta_j\}_{j=1}^s) p(\{\theta_j\}_{j=1}^s|A^{n,s}) d\{\theta_j\}_{j=1}^s \\
&= \sum_{A \in \rho_s(n)} p(A^{n,s}) \cdot \int_{\{\theta_j\}_{j=1}^s \in \Theta^s} \Big(\prod_{j=1}^{s} \prod_{x_i \in A_j} f_{\theta_j}(x_i) \prod_{j=1}^{s} \pi(\theta_j) \Big) d\{\theta_j\}_{j=1}^s.
\end{aligned}$$

To ease the ensuing discussion, we use $m(x_{A_j^{n,s}})$ to denote the probability density function of a cluster of samples, which is given by:

$$m(x_{A_j^{n,s}}) = \int_{\theta_j} f_{\theta_j}(x_{j,1}) \cdots f_{\theta_j}(x_{j,a_j}) \pi(\theta_j) d\theta_j,$$

where $x_{j,1}, \cdots, x_{j,a_j} \in A_j^{n,s}$ for all $1 \leq j \leq s$. Given this definition of $m(x_{A_j^{n,s}})$, we can rewrite $\mathbb{P}(K_n = s|x_n)$ as follows:

$$\begin{aligned}\mathbb{P}(K_n = s|\{x_i\}_{i=1}^n) &\propto \sum_{A^{n,s}\in\rho_s(n)} \Big(p(A^{n,s}) \cdot \prod_{j=1}^{s} m(x_{A_j^{n,s}})\Big) \\ &= \sum_{A^{n,s}\in\rho_s(n)} \Big(\frac{\alpha^s}{\alpha^{(n)}} \prod_{j=1}^{s}(|A_j^{n,s}| - 1)! \cdot \prod_{j=1}^{s} m(x_{A_j^{n,s}})\Big). \qquad (2.4)\end{aligned}$$

Early theoretical results on these posterior probabilities can be found in [41, 42]. A general qualitative result from [42] is as follows.

Theorem 5. *For a Dirichlet process mixture model, if the component distribution is Gaussian, log normal, Gamma, exponential, Weibull, or if it is discrete and has at least some point with nonzero measure for any choice of parameter, then, provided that the data are independently and identically distributed from a K^*-component mixture model, (K^* is bounded), and with probability one we have:*

$$\limsup_{n\to\infty} \mathbb{P}(K_n = K^*|\{x_i\}_{i=1}^n) < 1.$$

Our goal is to supply a quantitative perspective on these posterior probabilities, in a general setting where we do not confine ourselves to the case of finite or parametric mixtures. In other words, under more general assumptions on the data-generating process, we want to characterize the behavior of the posterior distribution of the number of clusters.

In order to obtain this characterization, we consider the following ratio between its values at $K_n = s + 1$ and $K_n = s$:

$$\begin{aligned}R(s|\{x_i\}_{i=1}^n) &:= \frac{\mathbb{P}(K_n = s+1 \,|\, \{x_i\}_{i=1}^n)}{\mathbb{P}(K_n = s \,|\, \{x_i\}_{i=1}^n)} \\ &= \frac{\sum_{A^{n,s+1}\in\rho_{s+1}(n)} \Big(p(A^{n,s+1}) \cdot \prod_{j=1}^{s+1} m(x_{A_j^{n,s}})\Big)}{\sum_{A^{n,s}\in\rho_s(n)} \Big(p(A^{n,s}) \cdot \prod_{j=1}^{s} m(x_{A_j^{n,s}})\Big)}.\end{aligned}$$

From the formulation of a DPMM, we further obtain that

$$\begin{aligned}&R(s|\{x_i\}_{i=1}^n) \\ &= \alpha \cdot \frac{\sum_{A^{n,s+1}\in\rho_{s+1}(n)} \Big((\prod_{i=1}^{s+1}(|A_i^{n,s+1}| - 1)!) \cdot \prod_{j=1}^{s+1} m(x_{A_j^{n,s+1}})\Big)}{\sum_{A^{n,s}\in\rho_s(n)} \Big((\prod_{i=1}^{s}(|A_i^{n,s}| - 1)!) \cdot \prod_{j=1}^{s} m(x_{A_j^{n,s}})\Big)}. \qquad (2.5)\end{aligned}$$

Our focus will be to obtain lower bounds on this ratio.

2.3 A Quantitative Result for the Posterior of the DPMM

In this section, we study the posterior distribution on the number of clusters induced by a DPMM. Our first result assumes a uniform prior on a finite interval, with a boundedness assumption on the data. The second result is an asymptotic result for the Gaussian prior.

Uniform prior

We first consider the posterior distribution of the number of clusters of a DPMM when the data lie in a bounded set. Data sets of this form often arise in fields such as biology, genetics, and economics. In this case it is natural to let the parameter space Θ be a compact set [50].

To ease the complexity of the algebraic aspects of the proof, we specifically consider a simple uniform prior on the parameter space Θ, where Θ is a bounded segment of $\mathbb{R}$ with size $|\Theta|$. With this choice, we have $\pi(\theta) = 1/|\Theta|$ for all $\theta \in \Theta$. We begin with the following result that establishes a lower bound on the ratio $R(s|\{x_i\}_{i=1}^n)$.

Theorem 6. *Given the DPMM defined in Eq.* (2.3)*, with a uniform prior* $\mathrm{Unif}(\Theta)$ *on* θ*, for sufficiently large* n*, if* $\min(\{x_i\}_{i=1}^n) > \min(\Theta)$ *and* $\max(\{x_i\}_{i=1}^n) < \max(\Theta)$*, then the ratio* $R(s|\{x_i\}_{i=1}^n)$ *between consecutive terms is lower bounded by*

$$R(s|\{x_i\}_{i=1}^n) \geq C \cdot \frac{\alpha}{s|\Theta|}, \tag{2.6}$$

where $C > 0$ *is a universal constant.*

Remark. Note that in many problems when one applies a uniform prior for the parameters, one expects the uniform prior to have support that is large enough to capture the means of all the components. It is also possible to extend to the case where some samples lie outside Θ by modifying Lemma 7, but the result would then have to be dependent on the specifics of the distribution and lose generality.

Proof. Given $A^{n,s} \in \rho_s(n)$, we define its $(s+1)$-*component extension set* $\eta_{ext}(A^{n,s})$, and its $(s-1)$-*component contraction set* $\eta_{con}(A^{n,s})$ as follows:

$$\begin{aligned}
\eta_{ext}(A^{n,s}) &:= \{\tilde{A}^{n,s+1} \in \rho_{s+1}(n) : A^{n,s} \in \eta_{con}(\tilde{A}^{n,s+1})\} \\
\eta_{con}(A^{n,s}) &:= \Big\{\tilde{A}^{n,s-1} \in \rho_{s-1}(n) : !\exists\, i, j \in [s] : \\
&\qquad \{A_1^{n,s}, \cdots, A_s^{n,s}, A_i^{n,s} \cup A_j^{n,s}\} \backslash \{A_i^{n,s}, A_j^{n,s}\} = \{\tilde{A}_1^{n,s-1}, \cdots, \tilde{A}_{s-1}^{n,s-1}\}\Big\},
\end{aligned}$$

where we require $s > 1$ in the definition of $\eta_{con}(A^{n,s})$.

In words, $\eta_{con}(A^{n,s})$ is the set of partitions in $\rho_{s-1}(n)$ that can be obtained from $A^{n,s}$ by combining two elements of $A^{n,s}$ into one element while keeping everything else the same. On the other hand, $\eta_{ext}(A^{n,s})$ is the set of partitions in $\rho_{s+1}(n)$ for which we can combine two of its elements into one to get $A^{n,s}$. For example, if $s = 2$, and we have $A^{n,3} = \{A_1^{n,3}, A_2^{n,3}, A_3^{n,3}\}$, then $\eta_{con}(A^{n,3})$ consists of following three partitions: $\{A_1^{n,3}, A_2^{n,3} \cup A_3^{n,3}\}, \{A_1^{n,3} \cup A_3^{n,3}, A_2^{n,3}\}, \{A_1^{n,3} \cup A_2^{n,3}, A_3^{n,3}\}$, and $\eta_{ext}(A^{n,3})$ is the set of all 4-partitions that can contract to $A^{n,3}$.

We further define the posterior probabilities of a partition to be:

$$p(A^{n,s}, x) := p(A^{n,s}) \cdot \prod_{i=1}^{s} m(x_{A_i^{n,s}}).$$

Given these definitions, we claim that we can rewrite the ratio $R(s|\{x_i\}_{i=1}^n)$ as follows:

$$\begin{aligned} R(s|\{x_i\}_{i=1}^n) &= \frac{\mathbb{P}(K_n = s+1|\{x_i\}_{i=1}^n)}{\mathbb{P}(K_n = s|\{x_i\}_{i=1}^n)} \\ &= \frac{2}{(s+1)s} \cdot \frac{\sum_{A^{n,s} \in \rho_s(n)} \Big(\sum_{A^{n,s+1} \in \eta_{ext}(A^{n,s})} p(A^{n,s+1}, x) \Big)}{\sum_{A^{n,s} \in \rho_s(n)} p(A^{n,s}, x)}, \end{aligned} \tag{2.7}$$

for $n \geq s + 1$. The proof of this claim is deferred to the end of the proof and we proceed while assuming this result. Note that the last fractional term in the claim takes the generic form $\frac{\sum_a f(a)}{\sum_a g(a)}$, which motivates us to study $f(a)/g(a)$ in the following.

For each partition in $\eta_{ext}(A^{n,s})$, there exist exactly two sets that can be combined to get one set in $A^{n,s}$, and all the others are the same as the other clusters of $A^{n,s}$. We now let $\eta_{ext}^{i,j}(A^{n,s})$ denote the subcollection of $\eta_{ext}(A^{n,s})$ such that for any partition in this subcollection, two sets can be combined into $A_i^{n,s}$, and that one set has size j. Since the order of sets in each partition does not matter, we will assume for simplicity that the two sets are the i-th and $(s+1)$-th clusters. To distinguish the s partition and its induced $s+1$ partitions, we use $B^{n,s}$ to denote the s-partition, and $A^{n,s+1} \in \rho_{ext}(B^{n,s})$ to denote its induced partitions. To simplify the notation, we use $b_1, \ldots, b_s$ to denote the cardinalities of $B_1^{n,s}, \cdots, B_s^{n,s}$ and $a_1, \ldots, a_{s+1}$ to denote those of $A_1^{n,s+1}, \ldots, A_{s+1}^{n,s+1}$. Furthermore, in the following derivation, the value n is fixed, so will omit it in the notation unless there is the necessity to discuss different

n. We obtain the following:

$$\begin{aligned}
&\frac{\sum_{A^{s+1}\in\eta_{ext}(B^s)} p(A^{s+1}|x)}{p(B^s|x)} \\
&= \sum_{A^{s+1}\in\eta_{ext}(B^s)} \alpha\cdot\frac{\prod_{i=1}^{s+1}(|A_i^{s+1}|-1)!\, m(x_{A_i^{s+1}})}{\prod_{i=1}^{s}(|B_i^s|-1)!\, m(x_{B_i^s})} \\
&= \alpha\cdot\sum_{i=1}^{s}\sum_{j=1}^{b_i-1}\sum_{A^{s+1}\in\eta_{ext}^{i,j}(B^s)} \frac{(j-1)!(b_i-j-1)!\, m(x_{A_i^{s+1}})m(x_{A_{s+1}^{s+1}})}{(b_i-1)!\, m(x_{B_i^s})} \\
&= \alpha\cdot\sum_{i=1}^{s}\sum_{j=1}^{b_i-1}\Big\{\frac{(j-1)!(b_i-j-1)!}{(b_i-1)!}\Big(\sum_{A\in\eta_{ext}^{i,j}} \frac{m(X_{A_i^{s+1}})m(X_{A_{s+1}^{s+1}})}{m(X_{B_i^s})}\Big)\Big\}. \qquad (2.8)
\end{aligned}$$

Recall that, by definition, $A_i^{s+1}\cup A_{s+1}^{s+1}=B_i^s$.

Lemma 7. *For a set x_S with n_s elements and x_T with n_t elements, suppose $\Theta\subseteq[\theta_{\min},\theta_{\max}]$ is an interval in $\mathbb{R}$ such that $\min x_S, \min x_T > \theta_{\min}+c$ and $\max x_S, \max x_T < \theta_{\max}-c$ for some $c>0$, we have:*

$$\frac{m(x_S)m(x_T)}{m(x_S\cup x_T)} \geq \frac{c_1\sqrt{2\pi}}{|\Theta|}\cdot\frac{\sqrt{n_s+n_t}}{\sqrt{n_s\, n_t}},$$

for some positive constant c_1 that does not depend on x_S, x_T but only on c.

Given Lemma 7, we can derive the following bounds for the term in (2.8):

$$\frac{\sum_{A^{s+1}\in\eta_{ext}(B^s)} p(A^{s+1}|x)}{p(B^s x)} \geq \frac{\alpha\, c_1\sqrt{2\pi}}{|\Theta|}\cdot\sum_{i=1}^{s}\sum_{j=1}^{b_i-1}\Big(\frac{b_i}{j(b_i-j)}\Big)^{3/2}. \qquad (2.9)$$

Note that:

$$\sum_{j=1}^{b_i-1}\Big(\frac{b_i}{j(b_i-j)}\Big)^{3/2} \geq \int_{x=1}^{b_i-1}\Big(\frac{b_i}{x(b_i-x)}\Big)^{3/2}dx \geq \frac{4(b_i-2)}{\sqrt{(b_i-1)b_i}}$$

The inequality implies that the leftmost term is no less than 2 for $b_i\geq 4$. When $b_i=2,3$, simple algebra indicates that $\sum_{j=1}^{b_i-1}\left(\frac{b_i}{j(b_i-j)}\right)^{3/2}\geq 2$. Invoking these results, we have the following lower bound:

$$\frac{\sum_{A^{s+1}\in\eta_{ext}(B^s)} p(A^{s+1}|x)}{p(B^s|x)} \geq \frac{\alpha\, c_1^2\sqrt{2\pi}}{|\Theta|}\cdot\sum_{i=1}^{s}2\cdot I_{b_i\geq 2} \gtrsim \frac{\alpha s}{|\Theta|}.$$

Combining this lower bound with Eq. (2.7), we obtain the following evaluation of the ratio between consecutive terms $R(s)$:

$$R(s|\{x_i\}_{i=1}^n) = \frac{\mathbb{P}(K_n = s+1|\{x_i\}_{i=1}^n)}{\mathbb{P}(K_n = s|\{x_i\}_{i=1}^n)} \gtrsim \frac{\alpha}{s|\Theta|}.$$

As a consequence, we reach the conclusion of the theorem.

Proof of claim (2.7)*:* Using equation (2.4), we can rewrite the ratio between the posterior probability of $s+1$ components and that of s components as follows:

$$\frac{\mathbb{P}(K_n = s+1|\{x_i\}_{i=1}^n)}{\mathbb{P}(K_n = s|\{x_i\}_{i=1}^n)} = \frac{\sum_{A^{s+1}\in\rho_{s+1}(n)} p(A^{s+1}, x)}{\sum_{B^s\in\rho_s(n)} p(B^s, x)}.$$

Note that for each $A^{s+1} \in \rho_{s+1}(n)$, we can merge any two of its $s+1$ parts to obtain some $B^s \in \rho_s(n)$. The number of distinct ways to do so is exactly $\binom{s+1}{2}$. Also, for each $B^s \in \rho_s(n)$, the set $\eta_{ext}(B^s)$ contains all $A^{s+1} \in \rho_{s+1}(n)$ such that they can merge some parts to get B^s. Thus, the index of the numerator on the right-hand side counts each $A^{s+1} \in \rho_{s+1}(n)$ exactly $\binom{s+1}{2}$ times, from which the equation follows. □

Proof of Lemma 7: For a set x_S with n_s elements, we use the following shorthand:

$$\overline{x_S} = \frac{\sum_{X\in x_S} X}{n_s}; \quad S^2_{x_S} = \sum_{X\in x_S} X^2.$$

After some algebra, we can verify that

$$\begin{aligned}
&\frac{m(x_S)m(x_T)}{m(x_S\cup x_T)} \\
&= \frac{\displaystyle\int_{\theta\in\Theta} \exp\Big(-\sum_{X\in x_S}\frac{(X-\theta)^2}{2}\Big)d\theta \cdot \int_{\theta\in\Theta}\exp\Big(-\sum_{X\in x_T}\frac{(X-\theta)^2}{2}\Big)d\theta}{|\Theta|\displaystyle\int_{\theta\in\Theta}\exp\Big(-\sum_{X\in x_S\cup x_T}\frac{(X-\theta)^2}{2}\Big)d\theta} \\
&= \frac{\frac{2\pi}{\sqrt{n_s n_t}}\exp\Big(-\frac{(S^2_{x_S}+n_1\overline{x_S}^2)}{2}-\frac{(S^2_{x_T}+n_t\overline{x_T}^2)}{2}\Big)P_{x_S}(\Theta)P_{x_S\cup x_T}(\Theta)}{|\Theta|\sqrt{\frac{2\pi}{n_s+n_t}}\exp\Big(-\frac{(S^2_{x_S\cup x_T}+(n_s+n_t)\overline{x_S\cup x_T}^2)}{2}\Big)} \\
&= \frac{\sqrt{2\pi}}{|\Theta|}\cdot\frac{\sqrt{n_s+n_t}}{\sqrt{n_s\,n_t}}\cdot\exp\Big(\frac{n_s\overline{x_S}^2+n_t\overline{x_T}^2-(n_s+n_t)\overline{x_S\cup x_T}^2}{2}\Big) \\
&\qquad\qquad \times\frac{P_{x_S}(\Theta)\,P_{x_T}(\Theta)}{P_{x_S\cup x_T}(\Theta)},
\end{aligned}$$

where $P_x(\Theta) := P(\theta \in \Theta | \theta \sim N(\overline{x}, \frac{1}{|x|}))$, which can be shown to be at least $\text{erf}(c|x|/\sqrt{2})$, where erf is the error function. Hence:

$$(\text{erf}(c/\sqrt{2}))^2 < \frac{P_{x_1}(\Theta)\, P_{x_2}(\Theta)}{P_{x_1 \cup x_2}(\Theta)} < \frac{1}{\text{erf}(c/\sqrt{2})}.$$

Note that this range in general is very small (around 1) for c that is not too small. For example, the range above is $(0.997^2, 1/0.997)$ for $c = 3$. Now, we write $c_1 = (\text{erf}(c/\sqrt{2}))^2$.

On the other hand,

$$\frac{n_s \overline{x_S}^2 + n_t \overline{x_T}^2 - (n_s + n_t)\overline{x_S \cup x_T}^2}{2} = \frac{n_s\, n_t (\overline{x_S} - \overline{x_T})^2}{2(n_s + n_t)} \geq 0.$$

Combining these results, we obtain the desired inequality. □

The bound in the result of Theorem 6 does not require the original distribution to be a mixture distribution. Instead, it holds provided that the true underlying distribution has finite and nonzero variance.

In particular, note that by the law of large numbers, if the observations in x_S, x_T are i.i.d., we may consider the empirical average of the exponential term in Lemma 7 and Eq. (2.8):

$$\exp\Big(n_s n_t (\overline{x_S} - \overline{x_T})^2 / 2(n_s + n_t)\Big).$$

The term inside the exponential is approximately $\frac{\sigma\sqrt{n_s n_t}}{2\sqrt{n_s+n_t}} Z$-distributed as n_s and n_t go to infinity, where Z is the standard normal distribution and σ^2 is the variance of the original distribution. Using the moment-generating function of the χ_2 distribution, we can see that this term grows unbounded as n_s and n_t grow, in contrast to the constant value one we derived in the proof of the Theorem 6. Therefore, given the result of Theorem 6, we obtain the following corollary:

Corollary 8. *If the conditions in Theorem 6 hold as $n \to \infty$, then for any s fixed, we have:*

$$\lim_{n \to \infty} R(s | \{x_i\}_{i=1}^n) \to \infty.$$

We note that the rate at which $R(s|\{x_i\}_{i=1}^n)$ grows is unknown and may be very slow when σ^2 is small, due to the complicated combinatorial behavior. Combining the results from Theorem 6 and Corollary 8, we can see that for any distribution with nonzero variance, the posterior probability of obtaining $s+1$ clusters will eventually exceed that of obtaining s clusters, and their ratio grows in an unbounded way. An important implication of this result is that, with a larger sample size we will tend to

fit a larger number of clusters, and this model ultimately leads to an infinite number of clusters almost surely. However, in the finite sample-size regime, the model's behavior depends more on the variance of the original distribution. In practice, one should rely on the finite bound in Theorem 6 unless additional information about the distribution is available.

Gaussian Prior

We now consider the more commonly used Gaussian prior on the parameter θ [10, 38]. In particular, we choose the prior density to be that of a univariate Gaussian distribution, $\mathcal{N}(0, \sigma^2)$, with fixed variance $\sigma > 0$. Given this choice of prior, we have the following asymptotic lower bound on $R(s|\{x_i\}_{i=1}^n)$:

Theorem 9. *For the DPMM defined in Eq.* (2.3)*, with a Gaussian prior* $\mathcal{N}(0, \sigma^2)$ *on* θ*, as* n *goes to infinity, the ratio* $R(s|\{x_i\}_{i=1}^n)$ *satisfies the following asymptotic lower bound:*

$$\lim_{n\to\infty} R(s|\{x_i\}_{i=1}^n) \geq C \cdot \frac{\alpha}{s^2(1+\sigma)}, \tag{2.10}$$

where $C > 0$ *is a universal constant.*

Remark. The result of Theorem 9 holds asymptotically. Its performance in finite samples is unknown and is complex, depending heavily on the original distribution. It would be of interest to characterize the finite-sample behavior of distributions satisfying certain conditions on variance and the true number of components or the true rate of growth in the number of components for an infinite-component distribution induced by processes such as the Dirichlet process.

Proof. We use the notation from the proof of Theorem 6 in this proof. As in Eq. 2.8, we study the following term: $(m(x_{A_i^{s+1}})m(x_{A_{s+1}^{s+1}}))/m(x_{B_i^s})$, where as in Theorem 6 we assume B^s to be some s-cluster formed by merging two clusters in A^{s+1}.

Lemma 10. *For sets* x_S *with* n_s *elements and* x_T *with* n_t *elements, suppose* $\Theta = \mathbb{R}$ *and* $\pi(\theta) = \frac{1}{\sqrt{2\pi}} \exp(-\frac{\theta^2}{2})$*. We have:*

$$\frac{m(x_S)m(x_T)}{m(x_S \cup x_T)} \geq \sqrt{\frac{1}{2} \cdot \frac{\tau}{1+\tau} \cdot \frac{n_s + n_t}{n_s n_t}},$$

with probability approaching one as $n_s + n_t$ *goes to infinity.*

Returning to the computation of the ratio $p(A|x)/p(B|x)$, for fixed s and a partition $B \in \rho_s(n)$, we define

$$U(B) := \{i \in [s] : b_i \geq \frac{n}{s^2}\}.$$

For any $i \in U(B)$, since b_i increases as n increases, we may assume that the aforementioned condition that $F(\tau; x_{B_1}, x_{B_2}) \geq 0$ holds asymptotically for any fixed proportion (less than one) for all the partitions of B_i. Then, for sufficiently large n we have:

$$\begin{aligned}
&\sum_{A^{s+1} \in \eta_{ext}(B^s)} \frac{p(A^{s+1}|x)}{p(B^s|x)} \\
&\quad = \alpha \cdot \sum_{i=1}^{s} \sum_{j=1}^{b_i - 1} \frac{(j-1)!(b_i - j - 1)!}{(b_i - 1)!} \Big(\sum_{A^{s+1} \in \eta_{ext}^{i,j}(B^s)} \frac{m(X_{A^{s+1}s_i}) m(X_{s+1 A_{s+1}})}{m(X_{B_i^s})} \Big) \\
&\quad \overset{w.h.p.}{\geq} \frac{\alpha}{2} \frac{\sqrt{\tau}}{\sqrt{1+\tau}} \cdot \sum_{i \in U(B)} \sum_{j=1}^{b_i - 1} \frac{(j-1)!(b_i - j - 1)!}{(b_i - 1)!} \Big(\sum_{A \in \eta_{ext}^{i,j}(B^s)} \sqrt{\frac{b_i}{j(b_i - j)}} \Big) \\
&\quad \gtrsim \frac{\alpha \sqrt{\tau}}{\sqrt{1+\tau}}.
\end{aligned}$$

Here, the second step follows with high probability by our previous argument, and the last step follows by a similar argument as in the case of uniform prior, except that we have only a constant term instead of an s in the case of uniform prior. This is true because it is possible to have $|U(A)| \ll s$, so the result can only be bounded by a constant multiple of $\alpha \cdot \frac{\sqrt{\tau}}{\sqrt{1+\tau}}$ without the s factor in the uniform case.

Finally, note that as n goes to infinity, the result holds with probability one. Therefore, we obtain that

$$\lim_{n \to \infty} R(s|\{x_i\}_{i=1}^n) = \lim_{n \to \infty} \frac{\mathbb{P}(K_n = s+1|\{x_i\}_{i=1}^n)}{\mathbb{P}(K_n = s|\{x_i\}_{i=1}^n)} \gtrsim \frac{\alpha}{s^2} \cdot \frac{\sqrt{\tau}}{\sqrt{1+\tau}}.$$

As a consequence, we reach the conclusion of the theorem. □

Proof of Lemma 10: We have:

$$\begin{aligned}&\frac{m(x_S)m(x_T)}{m(x_S \cup x_T)}\\ &= \frac{1}{\sqrt{\sigma^2}}\sqrt{\frac{(n_s+n_t)+\frac{1}{\sigma^2}}{(n_s+\frac{1}{\sigma^2})(n_t+\frac{1}{\sigma^2})}}\\ &\qquad\qquad \times \exp\left(\frac{1}{2}\left(\frac{n_s^2\overline{x_S}^2}{n_s+\frac{1}{\sigma^2}}+\frac{n_t^2\overline{x_T}^2}{n_t+\frac{1}{\sigma^2}}-\frac{(n_s+n_t)^2\overline{x_S\cup x_T}^2}{(n_s+n_t)+\frac{1}{\sigma^2}}\right)\right).\end{aligned}$$

To simplify the notation, we let $\tau := \frac{1}{\sigma^2}$ be the precision, and rewrite this expression as:

$$\sqrt{\tau}\cdot\sqrt{\frac{n_s+n_t+\tau}{(n_s+\tau)(n_t+\tau)}}\exp\Big(\frac{1}{2}\underbrace{\Big(\frac{n_s^2\overline{x_S}^2}{n_s+\tau}+\frac{n_t^2\overline{x_T}^2}{n_t+\tau}-\frac{(n_s+n_t)^2\,\overline{x_S\cup x_T}^2}{(n_s+n_t)+\tau}\Big)}_{F(\tau;x_S,x_T)}\Big).$$

Note that the term $F(\tau; x_S, x_T)$ is nonnegative at zero, since

$$F(0;x_S,x_T)=\frac{n_sn_t}{n_s+n_t}(\overline{x_S}-\overline{x_T})^2\geq 0.$$

Solving a quadratic function shows that the equation $F(\tau; x_1, x_2) = 0$ has its positive root at:

$$\begin{cases}\frac{1}{4}\Big[\sqrt{8n_sn_t\frac{(\overline{x_S}-\overline{x_T})^2}{\overline{x_S}\,\overline{x_T}}+Q^2}+Q\Big], & \text{if } \overline{x_S}\,\overline{x_T}>0\\ \frac{1}{4}\Big[\sqrt{8n_sn_t\frac{(\overline{x_S}-\overline{x_T})^2}{\overline{x_S}\,\overline{x_T}}+Q^2}+Q\Big] \text{ or } \emptyset, & \text{if } \overline{x_S}\,\overline{x_T}<0\\ \emptyset & \text{if } \overline{x_S}=0, \overline{x_T}\neq 0 \text{ or } \overline{x_S}\neq 0, \overline{x_T}=0\\ \mathbb{R}^+ & \text{if } \overline{x_S}=\overline{x_T}=0,\end{cases}$$

where $Q := n_t\frac{\overline{x_T}}{n_s\overline{x_S}}+n_s\frac{n_s\overline{x_S}}{n_t\overline{x_T}}-2(n_s+n_t)$.

If there is no positive root or every positive number is a root, then $F(\tau; x_S, x_T) \geq 0$ for any $\tau > 0$. Otherwise, there exists a unique positive root $r_+(x_S, x_T) := \frac{1}{4}\Big[\sqrt{8n_sn_t\frac{(\overline{x}_S-\overline{x}_T)^2}{\overline{x}_S\overline{x}_T}+Q^2}+Q\Big]$, a random variable depending on n_s, n_t, whose probability density function concentrates on increasingly larger values as long as one of n_s, n_t goes to infinity. That is, the root grows larger stochastically as n_s and n_t increase. For any fixed τ, as $n_s + n_t$ goes to infinity, it follows that for any fixed

$\tau > 0$, the probability that $[0, r_+(x_S, x_T)]$ contains τ approaches one asymptotically, so we have that $F(\tau; x_S, x_T) \geq 0$.

Note that for any positive integers n_s, n_t and nonnegative number τ, we have

$$\frac{n_s + n_t + \tau}{(n_s + \tau)(n_t + \tau)} > \frac{1}{2} \cdot \frac{1}{1 + \tau} \cdot \frac{n_s + n_t}{n_s n_t}.$$

The result follows. □

2.4 Extension to Pitman-Yor Process Mixture Models

In this section, we extend our results to the Pitman-Yor process mixture model, a model that is similar to the Dirichlet process mixture model but allows a faster rate of creation of clusters. Accordingly, we will see that the ratio between the posterior probabilities of consecutive number of clusters is larger than that of the DPMM.

We retain most of the notation used in the previous section unless otherwise specified. We also use the standard Chinese restaurant process nomenclature to describe the dynamics associated with the partitions formed by the Pitman-Yor process, where the items to be partitioned are referred to as "customers," and partions are referred to as "tables" [46]. Letting (α, ξ) denote the parameters of the Pitman-Yor process prior, we obtain:

$$\mathbb{P}(\text{customer } n+1 \text{ joins table } c \,|\, A^{n,s}) = \begin{cases} \frac{|c| - \xi}{\alpha + n} & \text{if } c \leq s, \\ \frac{\alpha + \xi s}{\alpha + n} & \text{otherwise.} \end{cases} \tag{2.11}$$

The idea is that joining an existing table c has a probability of $\frac{|c| - \xi}{\alpha + n}$, a probability proportional to $|c| - \xi$, the number of customers already at the table, discounted by the value ξ, and the probability of joining new table is proportional to $\alpha + \xi K_n$. Then, the probability of obtaining $A^{n,s}$ is:

$$\mathbb{P}(A^{n,s}) = \frac{\alpha(\alpha + \xi) \cdots (\alpha + \xi(s-1))}{\alpha^{(n)}} \prod_{i=1}^{s} (1 - \xi)(2 - \xi) \cdots (|A_i^{n,s}| - 1 - \xi),$$

where $\alpha^{(n)} = \alpha(\alpha+1) \cdots (\alpha+n-1)$. Note that if $|A_i^{n,s}| = 1$, the corresponding term in the product is 1. We denote

$$|c \ominus \xi|! = (1 - \xi)(2 - \xi) \cdots (c - \xi),$$

and define $|0 \ominus \xi|! = 1$.

Overall, we write:

$$(A^{n,s}, s) \sim \mathrm{PYP}(\alpha, \xi),$$

and define the *Pitman-Yor Process Mixture Model* (PYPMM) as follows:

$$(A^{n,s}, s) \sim \mathrm{PYP}(\alpha, \xi) \tag{2.12}$$

$$p(\theta \,|\, A^{n,s}, s) = \prod_{i=1}^{s} \pi(\theta_i) \tag{2.13}$$

$$p(\{x_i\}_{i=1}^n \,|\, \{\theta_j\}_{j=1}^s, A^{n,s}, s) = \prod_{j=1}^{s} \prod_{x \in A_j^{n,s}} f_{\theta_j}(x_i), \tag{2.14}$$

where π is a prior on θ and f_{θ_j} belongs to a family of (known) densities. Moreover, analogously with the case of the DPMM, we obtain the following form for the posterior:

$$\begin{aligned}
&\mathbb{P}(K_n = s | \{x_i\}_{i=1}^n) \\
&= \frac{\mathbb{P}(\{x_i\}_{i=1}^n \,|\, K_n = s)\mathbb{P}(K_n = s)}{\mathbb{P}(\{x_i\}_{i=1}^n)} \\
&\propto \sum_{A^{n,s} \in \rho_s(n)} \mathbb{P}(A^{n,s}) \cdot \int_{\{\theta_j\}_{j=1}^n} p(\{x_i\}_{i=1}^n \,|\{\theta_j\}_{j=1}^s) p(\{\theta_j\} \,|A^{n,s})\, d\{\theta_j\}_{j=1}^s \\
&= \sum_{A^{n,s} \in \rho_s(n)} \mathbb{P}(A^{n,s}) \int_{\{\theta_j\}_{j=1}^s} \left(\prod_{j=1}^{s} \prod_{x_i \in A_j^{n,s}} f_{\theta_j}(x_i) \right) \prod_{j=1}^{s} \pi(\theta_j)\, d\{\theta_j\}_{j=1}^s \\
&= \sum_{A^{n,s} \in \rho_s(n)} \mathbb{P}(A^{n,s}) \prod_{j=1}^{s} \int_{\theta_j} \left(\prod_{x_i \in A_j^{n,s}} f_{\theta_j}(x_i) \right) \pi(\theta_j)\, d\theta_j \\
&= \sum_{A^{n,s} \in \rho_s(n)} \mathbb{P}(A^{n,s}) \prod_{j=1}^{s} m(x_{A_j^{n,s}}).
\end{aligned}$$

Thus, we have

$$\begin{aligned}
&\mathbb{P}(K_n = s | \{x_i\}_{i=1}^n)\mathbb{P}(\{x_i\}_{i=1}^n) \\
&= \sum_{A^{n,s} \in \rho_s(n)} \frac{\alpha(\alpha + \xi) \cdots (\alpha + \xi(s-1))}{\alpha^{(n)}} \prod_{i=1}^{s} |(|A_i^{n,s}| - 1) \ominus \xi|! \cdot m(x_{A_i^{n,s}}).
\end{aligned}$$

Recall the definitions

$$\eta_{ext}(A^{n,s}) := \{\tilde{A}^{n,s+1} \in \rho_{s+1}(n) : A^{n,s} \in \eta_{con}(\tilde{A}^{n,s+1})\},$$
$$\eta_{con}(A^{n,s}) := \Big\{\tilde{A}^{n,s-1} \in \rho_{s-1}(n) : !\exists\, i,j \in [s] : \{A_1^{n,s}, \cdots, A_s^{n,s}, A_i^{n,s} \cup A_j^{n,s}\} \backslash \{A_i^{n,s}, A_j^{n,s}\} = \{\tilde{A}_1^{n,s-1}, \cdots, \tilde{A}_{s-1}^{n,s-1}\}\Big\},$$

and note that $\eta_{s+1}(A)$ is the set of partitions formed by combining A_i and A_{s+1}, and $\tilde{\eta}_s(B)$ is all sets in $\rho_{s+1}(n)$ that can be formed by splitting one of B into two clusters (and then making one of them the "last cluster"). This yields:

$$\begin{aligned} R(s|\{x_i\}_{i=1}^n) &= \frac{\mathbb{P}(K_n = s+1 \,|\, \{x_i\}_{i=1}^n)}{\mathbb{P}(K_n = s \,|\, \{x_i\}_{i=1}^n)} \\ &= \frac{2}{(s+1)s} \cdot \frac{\sum_{B^{n,s} \in \rho_s(n)} \Big(\sum_{A^{n,s-1} \in \eta_{ext}(B^{n,s})} p(A^{n,s-1}, x)\Big)}{\sum_{B^{n,s} \in \rho_s(n)} p(B^{n,s}, x)}. \end{aligned} \tag{2.15}$$

Thus, continuing with the proof, we have:

$$\begin{aligned} &\frac{\sum_{A^{n,s+1} \in \eta_{ext}(B^{n,s})} p(A^{n,s+1}|x)}{p(B^{n,s}|x)} \\ &= \sum_{A^{n,s+1} \in \eta_{ext}(B^{n,s})} (\alpha + \xi s) \frac{\prod_{i=1}^{s+1} |(|A_i^{n,s+1}| - 1) \ominus \xi|! \cdot \prod_{i=1}^{s+1} m(x_{A_i^{n,s+1}})}{\prod_{i=1}^{s} |(|B_i^{n,s}| - 1) \ominus \xi|! \cdot \prod_{i=1}^{s} m(x_{B_i^{n,s}})} \\ &= (\alpha + \xi s) \sum_{i=1}^{s} \sum_{j=1}^{b_i - 1} \sum_{A^{n,s+1} \in \eta_{ext}^{i,j}(B^{n,s})} \frac{|(j-1) \ominus \xi|! \cdot |(b_i - j - 1) \ominus \xi|!}{|(b_i - 1) \ominus \xi|!} \\ &\qquad \times \frac{m(x_{A_i^{n,s+1}}) m(x_{A_{s+1}^{n,s+1}})}{m(x_{B_i^{n,s}})}. \end{aligned} \tag{2.16}$$

Note that we let $b_i = |B_i|$, and let $\eta_{ext}^{i,j}(B^{n,s})$ be defined to be all $A^{n,s+1} \in \eta_{ext}(B^{n,s})$ obtained by splitting $B_i^{n,s}$ into two clusters of size j and $b_i - j - 1$. We now separate into two cases, where the prior is uniform and Gaussian respectively.

Uniform prior

Similar to our analysis with the DPMM, we first consider the behavior of the posterior distribution for the number of clusters when the prior is a uniform distribution.

Theorem 11. *Given the PYPMM defined in Eq.* (2.14) *with a uniform prior* $\mathrm{Unif}(\Theta)$ *on* θ*, when* n *is sufficiently large, if* $\min(\{x_i\}_{i=1}^n) > \min(\Theta) + c$ *and* $\max(\{x_i\}_{i=1}^n) < \max(\Theta) - c$ *for some* $c > 0$*, then the ratio* $R(s|\{x_i\}_{i=1}^n)$ *between consecutive terms is lower bounded by*

$$R(s|\{x_i\}_{i=1}^n) \gtrsim \frac{(\alpha + \xi s)}{s|\Theta|}. \tag{2.17}$$

Proof. By Lemma (7) and under the same condition, we have that when $B_i^{n,s} = A_i^{n,s+1} \cup A_{s+1}^{n,s+1}$ and letting $a_k = |A_k^{n,s+1}|$ for $k \in [s+1]$,

$$\frac{m(x_{A_i^{n,s+1}})m(x_{A_{s+1}^{n,s+1}})}{m(x_{B_i^{n,s}})} \geq \frac{c_1\sqrt{2\pi}}{|\Theta|}\frac{\sqrt{a_i + a_{s+1}}}{\sqrt{a_i a_{s+1}}}.$$

Now, we obtain that

$$\begin{aligned}
&\frac{\sum_{A^{n,s+1}\in\eta_{ext}(B^{n,s})} p(A^{n,s+1}|x)}{p(B^{n,s}|x)} \\
&\geq (\alpha+\xi s)\sum_{i=1}^{s}\sum_{j=1}^{b_i-1}\frac{|(j-1)\ominus\xi|!|(b_i-j-1)\ominus\xi|!}{|(b_i-1)\ominus\xi|!} \\
&\qquad\qquad \times \sum_{A^{s+1}\in\eta_{ext}^{i,j}(B^s)} \frac{c_1\sqrt{2\pi}}{|\Theta|}\cdot\frac{\sqrt{b_i}}{\sqrt{j(b_i-j)}} \\
&= (\alpha+\xi s)\frac{c_1\sqrt{2\pi}}{|\Theta|}\sum_{i=1}^{s}\sum_{j=1}^{b_i-1}\frac{|(j-1)\ominus\xi|!|(b_i-j-1)\ominus\xi|!}{|(b_i-1)\ominus\xi|!}\cdot\binom{b_i}{j}\sqrt{\frac{b_i}{j(b_i-j)}}.
\end{aligned}$$

Note for $j = 1$ and $j = b_i - 1$ we have that for $b_i \geq 2$

$$\frac{|(j-1)\ominus\xi|!|(b_i-j-1)\ominus\xi|!}{|(b_i-1)\ominus\xi|!}\cdot\binom{b_i}{j}\sqrt{\frac{b_i}{j(b_i-j)}} = \frac{b_i}{b_i-1-\xi}\sqrt{\frac{b_i}{b_i-1}} \geq 1.$$

Plugging this back into the expression above

$$\frac{\sum_{A^{n,s+1}\in\eta_{ext}(B^{n,s})} p(A^{n,s+1}|x)}{p(B^{n,s}|x)} \geq (\alpha+\xi s)\frac{(0.997)^2\sqrt{2\pi}}{|\Theta|}\sum_{i=1}^{s} 2\cdot I_{b_i\geq 2} \gtrsim \frac{(\alpha+\xi s)s}{|\Theta|}.$$

Then, using Eq. 2.16 we get that

$$\begin{aligned} R(s|\{x_i\}_{i=1}^n) &= \frac{2}{(s+1)s} \cdot \frac{\sum_{B^{n,s}\in\rho_s(n)} \left(\sum_{A^{n,s+1}\in\eta_{ext}(B^{n,s})} p(A^{n,s+1}|x)\right)}{\sum_{B^{n,s}\in\rho_s(n)} p(B^{n,s}|x)} \\ &\geq \frac{2}{(s+1)s} \min_{B^{n,s}\in\rho_s(n)} \sum_{A^{n,s+1}\in\eta_{ext}(B^{n,s})} \frac{p(A^{n,s+1}|x)}{p(B^{n,s}|x)} \\ &\gtrsim \frac{2}{(s+1)s} \frac{(\alpha+\xi s)s}{|\Theta|} \gtrsim \frac{(\alpha+\xi s)}{s|\Theta|} = \frac{\alpha}{s|\Theta|} + \frac{\xi}{|\Theta|}. \end{aligned}$$

As a consequence, we obtain the conclusion of the theorem. □

Gaussian Prior

We now turn to the case where the prior is a Gaussian distribution. We make similar assumptions on the prior as we made in Section 2.3 with the DPMM.

Theorem 12. *For the PYPMM defined in Eq.* (2.14)*, with a Gaussian prior* $\mathcal{N}(0,\sigma^2)$ *on* θ*, as* n *goes to infinity, the ratio* $R(s|\{x_i\}_{i=1}^n)$ *satisfies the following asymptotic lower bound:*

$$\lim_{n\to\infty} R(s|\{x_i\}_{i=1}^n) \geq \frac{C(\alpha+\xi s)}{s^2} \cdot \frac{1}{1+\sqrt{\sigma^2}}, \tag{2.18}$$

where $C > 0$ *is a universal constant.*

Proof. Most steps are the same as in the derivation for the DPMM. We denote

$$D_{j,b_i} = \frac{|(j-1)\ominus\xi|!|(b_i-j-1)\ominus\xi|!}{|(b_i-1)\ominus\xi|!}.$$

Then, by equation (2.16) we have that

$$\begin{aligned} &\sum_{A^{n,s+1}\in\eta_{ext}(B^{n,s})} \frac{p(A^{n,s+1}|x)}{p(B^{n,s}|x)} \\ &= (\alpha+\xi s)\sum_{i=1}^{s}\sum_{j=1}^{b_i-1} D_{j,b_i} \sum_{A^{n,s+1}\in\eta_{ext}^{i,j}(B^{n,s})} \frac{m(x_{A_i^{n,s+1}})m(x_{A_{s+1}^{n,s+1}})}{m(x_{B_i^{n,s}})} \\ &\overset{w.h.p.}{\geq} C_0(\alpha+\xi s)\sqrt{\tau} \sum_{i\in U(B^{n,s})}\sum_{j=1}^{b_i-1} D_{j,b_i} \times \left(\sum_{A^{n,s+1}\in\eta_{ext}^{i,j}(B^{n,s})} \sqrt{\frac{b_i+\tau}{(j+\tau)(b_i-j+\tau)}}\right) \end{aligned}$$

$$\geq C_0(\alpha+\xi s)\sqrt{\tau}\sum_{i\in U(B^{n,s})}\sum_{j=1}^{b_i-1}D_{j,b_i}\times\left(\sum_{A\in\eta_{ext}^{i,s}(B^{n,s})}\frac{1}{\sqrt{2(1+\tau)}}\sqrt{\frac{b_i}{j(b_i-j)}}\right)$$
$$\geq \frac{C\sqrt{\tau}}{1+\sqrt{\tau}}(\alpha+\xi s),$$

where most of the steps follow similarly as in the proof of the DPMM in Theorem 9. Thus we conclude that

$$\lim_{n\to\infty}R(s\,|\,\{x_i\}_{i=1}^n)\gtrsim\frac{(\alpha+\xi s)}{s^2}\cdot\frac{\sqrt{\tau}}{1+\sqrt{\tau}}=\frac{\sqrt{\tau}}{1+\sqrt{\tau}}\left(\frac{\alpha}{s^2}+\frac{\xi}{s}\right).$$

As a consequence, we reach the conclusion of the theorem. □

Remark. Compared to the DPMM, the PYPMM has a faster rate of table generation, and accordingly a longer tail for the number of clusters. As we can see in Theorems 11 and 12, this leads to a larger lower bound in the ratio between consecutive posterior probabilities. Essentially, the lower bound for the uniform case is raised to a constant, and that for the Gaussian case is raised to the order of $1/s$, which corresponds to our intuition that PYPMM is more likely to fit a bigger number of clusters in the posterior.

2.5 Experiments

In the current section, we discuss the results of numerical experiments that aim to empirically strengthen our understanding of the clustering behavior of the DPMM and PYPMM. Throughout this section, for simplicity of implementation, we consider the Dirichlet mixture of standard normals; i.e., we assume $f_\theta(x)=\frac{1}{\sqrt{2\pi}}e^{-(x-\theta)^2/2\sigma^2}$, where $f_\theta(x)$ is the likelihood function in Eq. (2.3). As for the prior density on the parameters, we choose $\pi(\theta)$ to be Gaussian.

To compute the posterior density given the data points, we note that if data points $\{x_i\}_{i=1}^k$ fall within the cluster, then it follows that

$$\begin{aligned}p(\theta|\{x_i\}_{i=1}^n)&\propto p(\theta)p(\{x_i\}_{i=1}^n|\theta)\\&=\frac{1}{\sqrt{2\pi}}e^{-\theta^2/2\sigma^2}\prod_{i=1}^n\frac{1}{\sqrt{2\pi}}e^{-\frac{1}{2}(x_i-\theta)^2}\\&\sim\mathcal{N}\left(\frac{\sigma^2\sum_{i=1}^n x_i}{n\sigma^2+1},\frac{\sigma^2}{n\sigma^2+1}\right).\end{aligned}$$

Moreover, to compute the marginal density of a specific x, we compute as follows:

$$\begin{aligned} f(x) &= \int f_\theta(x)\pi(\theta)\,d\theta \\ &= \int \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}(x-\theta)^2} \cdot \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{\theta^2}{2\sigma^2}}\,d\theta \\ &\sim \mathcal{N}(0, \sigma^2 + 1). \end{aligned}$$

We pick $\alpha = 1$ and $\sigma = 1$ for our simulations. Specifically, we use the "dirichletprocess" R package [23] for these simulations, where the software package runs a Gibbs sampler on the DPMM model. For all the experiments, we run the Gibbs sampler with a burn-in period with 2×10^5 burn-in steps, and then take 10000 succeeding samples with step size 100.

Finite-cluster models

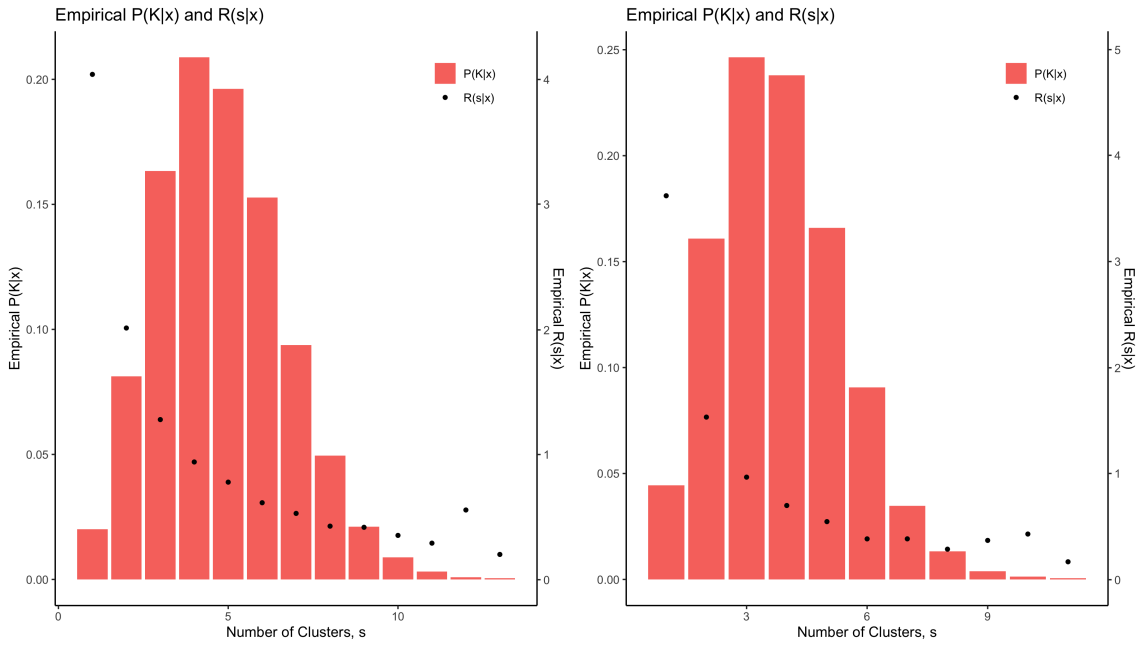
In this subsection we consider two finite-cluster distributions. In the first experiment, we generate 300 i.i.d. data points following a standard normal distribution. In the second experiment, we generate another 300 data points with the following two-cluster distribution:

$$f(x) = \frac{1}{2}, \quad \forall x \in [0,1] \cup [2,3].$$

After running the Gibbs sampler, we compute the posterior frequency of the number of clusters F_s for each s with some nonzero count of posterior frequency. This gives an empirical distribution of $\mathbb{P}(K_n = s|\{x_i\}_{i=1}^n)$, from which we may approximate $R(s|\{x_i\}_{i=1}^n)$. Note, however, that this approximation may be of low quality for large s given the corresponding low posterior frequency.

Figure 2.1 presents a plot of posterior frequencies and their ratios. The red bars correspond to the posterior frequency of the number of components, and the black dots are the ratios between the posterior frequencies, F_{s+1}/F_s.

The theoretical bound asserts that the ratio between consecutive posterior frequencies is at least of the order of $1/s^2$ under a Gaussian prior. To evaluate this, we plot $R(s) \cdot s^2$ versus s, which, roughly speaking, should look flat or have an upwards trend. In Figure 2.2a, we plot $\widehat{R}(s|x) \cdot s^2$ for both the one-cluster Gaussian and two-cluster uniform cases, where we observe the expected upward trend. In Figure 2.2b, we plot $\widehat{R}(s|x) \cdot s$ to assess whether the relationship is roughly linear. If linearity holds true then we would expect to see the dots for both settings to be roughly flat. Finally we note that it is reasonable to have erratic behavior of $\widehat{R}(s|x)$ when s is large since the posterior frequencies for large s are all very close to zero and the estimator $\widehat{R}$ is unstable in that case.



(a) Gaussian, one cluster

(b) Uniform, two clusters

Figure 2.1: Plots for empirical $\mathbb{P}(K_n|x)$ and $R(s|x)$

Dirichlet process mixture models

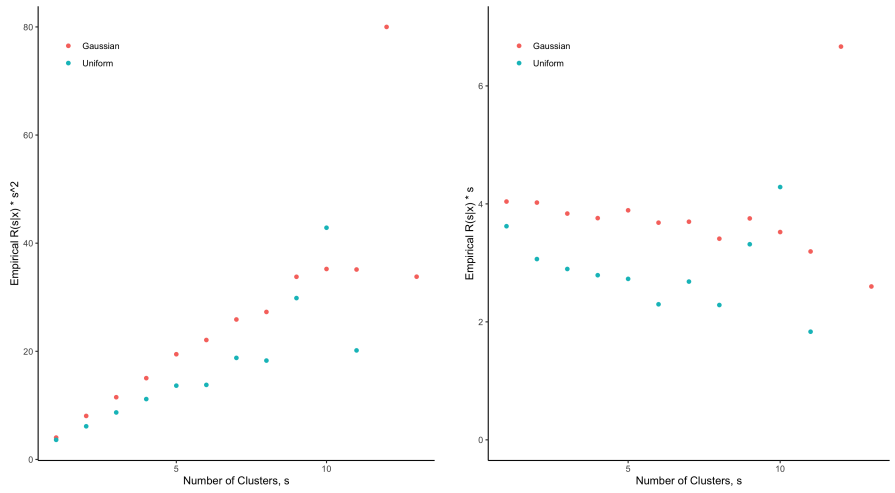
It is also of interest to investigate the behavior when the data follow an infinite-component distribution.

To study this case, we investigate a Dirichlet process data-generating model, with the $\alpha = 3$, $\theta_i \overset{i.i.d.}{\sim} N(0, 25)$ for each i, and

$$x_i|\theta_j, i \in A_j^{n,s} \sim N(\theta_j, 1).$$

Again, we generate 300 samples for three independent experiments, which result in 11, 14, and 17 clusters respectively.

We again plot the empirical values of $\mathbb{P}(K_n|x)$ and $R(s|x)$ in Figure 2.3, and investigate the nature of the growth rate of the posterior number of clusters in Figure 2.4. As we see in the latter figure, the ratios between consecutive posterior frequencies of the number of clusters are at least of order $1/s^2$.



(a) Empirical $\widehat{R}(s|x) * s^2$ (b) Empirical $\widehat{R}(s|x) * s$

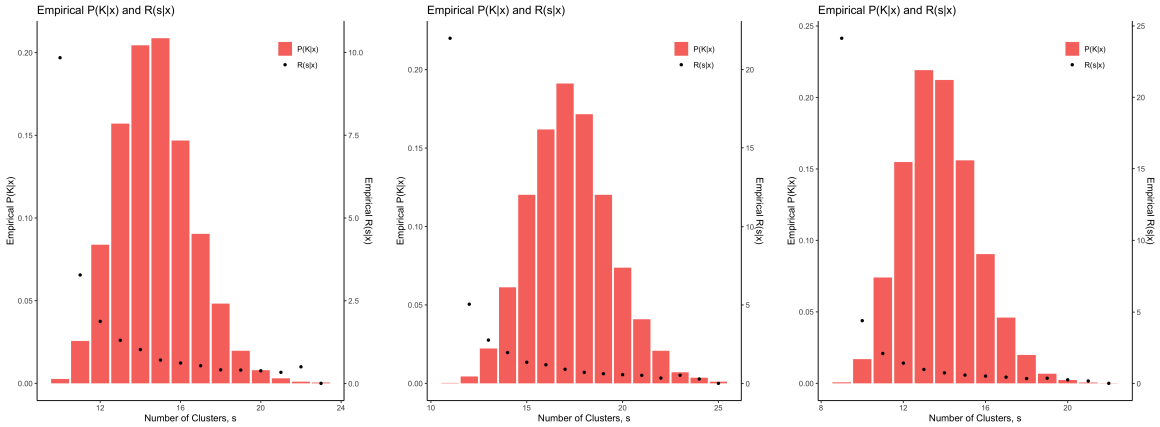
Figure 2.2: Plots of $\widehat{R}(s|x)$ with respect to different order of powers

2.6 Discussion

We have established lower bounds on the ratio of posterior probabilities $R(s|\{x_i\}_{i=1}^n)$ for the number of clusters for both the Dirichlet process mixture model and Pitman-Yor process mixture model, under several choices of prior distributions on the parameter space. While the complicated combinatorial structure of the DPMM and PYPMM preclude general characterizations of the posterior distribution, we have shown that it is possible to obtain analytical results for the posterior distribution of the number of clusters. We obtained both asymptotic and nonasymptotic results for the distribution of this important quantity.

An interesting open problem is to consider whether our rates are optimal or can be further improved. In particular, in the case of a Gaussian prior, our simulation suggests that instead of the $1/s^2$ rate predicted by our theory, a tighter $1/s$ rate may be possible. This seems plausible given that the $1/s$ rate arises in the case of uniform prior.

Our results provide a strong negative response to the question of whether basic nonparametric Bayesian models such as the DPMM or the PYPMM are able to infer the true number of clusters when the true number of clusters is finite. Indeed, our results provide a quantitative refutation of this naive hope. Additional mechanisms, such as truncation or some form of regularization, will be necessary to obtain consis-

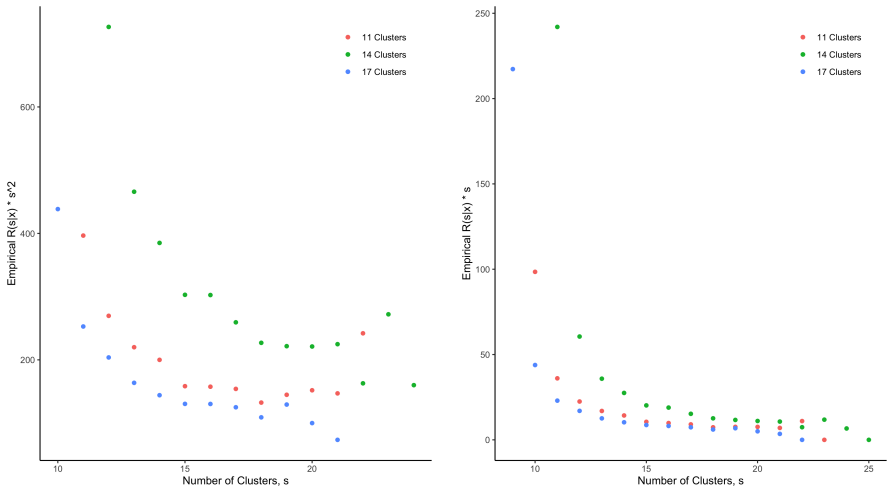


(a) 11 clusters (b) 14 clusters (c) 17 clusters

Figure 2.3: Plots for empirical $\mathbb{P}(K_n|x)$ and $R(s|x)$ in three simulations from Dirichlet process.

tent inference of the number of clusters in the finite setting. Indeed, recent work has shown that truncation can yield consistency with the number of clusters when the true data-generation mechanism is a finite mixtures [24]. However, the truncation method brings additional tuning parameters into the picture, including notions of separation of clusters, which may not be easily determined in practice. The problem of consistent estimation of the number of clusters remains open.

There is, however, another interesting open problem, which arises when the true distribution contains an infinite number of components. Do the DPMM or the PYPMM guarantee an infinite number of clusters in this case? Does the rate of growth of clusters in the posterior match that of the data-generating distribution, at least asymptotically? What about the case in which the truth is Dirichlet process mixture of normals or simply a general Dirichlet process? Does DPMM generate a posterior number of clusters at the same rate as implied by the Dirichlet process? In answering these questions, it will be necessary to derive some sort of upper bound on the $R(s|X)$, and focus on the asymptotic regime.



(a) Empirical $R(s|x) * s^2$ (b) Empirical $R(s|x) * s$

Figure 2.4: Plots of $\widehat{R}(s|x)$ with respect to different order of powers in three simulations from the Dirichlet process.

Bibliography

[1] Kasra Alishahi, Ahmad Reza Ehyaei, and Ali Shojaie. "A generalized benjamini-hochberg procedure for multivariate hypothesis testing". In: *arXiv preprint arXiv:1606.02386* (2016).

[2] C. Antoniak. "Mixtures of Dirichlet processes with applications to Bayesian nonparametric problems". In: *Annals of Statistics* 2 (1974), pp. 1152–1174.

[3] Sivaraman Balakrishnan and Larry Wasserman. "Hypothesis testing for densities and high-dimensional multinomials: Sharp local minimax rates". In: *arXiv preprint arXiv:1706.10003* (2017).

[4] Sivaraman Balakrishnan and Larry Wasserman. "Hypothesis Testing for High-Dimensional Multinomials: A Selective Review". In: *arXiv preprint arXiv:1712.06120* (2017).

[5] Rina Foygel Barber, Emmanuel J Candès, et al. "Controlling the false discovery rate via knockoffs". In: *The Annals of Statistics* 43.5 (2015), pp. 2055–2085.

[6] Yoav Benjamini and Yosef Hochberg. "Controlling the false discovery rate: a practical and powerful approach to multiple testing". In: *Journal of the royal statistical society. Series B (Methodological)* (1995), pp. 289–300.

[7] D. Blackwell and J. MacQueen. "Ferguson distributions via Polya urn schemes". In: *Annals of Statistics* 1 (1973), pp. 353–355.

[8] Bradley Efron. *Local false discovery rates*. 2005.

[9] Bradley Efron. "Microarrays, empirical Bayes and the two-groups model". In: *Statistical science* (2008), pp. 1–22.

[10] Michael D Escobar and Mike West. "Bayesian density estimation and inference using mixtures". In: *Journal of the American Statistical Association* 90 (1995), pp. 577–588.

[11] Thomas S Ferguson. "A Bayesian analysis of some nonparametric problems". In: *Annals of Statistics* 1 (1973), pp. 209–230.

[12] Manuel AR Ferreira and Shaun M Purcell. "A multivariate test of association". In: *Bioinformatics* 25.1 (2008), pp. 132–133.

[13] William Fithian and Daniel Ting. "Family learning: nonparametric statistical inference with parametric efficiency". In: *arXiv preprint arXiv:1711.10028* (2017).

[14] E. Fox et al. "A sticky HDP-HMM with application to speaker diarization". In: *Annals of Applied Statistics* 5 (2011), pp. 1020–1056.

[15] Max Grazier G'Sell et al. "Sequential selection procedures and false discovery rate control". In: *Journal of the royal statistical society: series B (statistical methodology)* 78.2 (2016), pp. 423–444.

[16] S. van de Geer. *Empirical processes in M-estimation*. Cambridge University Press, 2000.

[17] S. Ghosal, J. K. Ghosh, and A. van der Vaart. "Convergence rates of posterior distributions". In: *Annals of Statistics* 28 (2000), pp. 500–531.

[18] S. Ghosal and A. van der Vaart. "Convergence rates of posterior distributions for noniid observations". In: *Annals of Statistics* 35 (2007), pp. 192–223.

[19] S. Ghosal and A. van der Vaart. "Entropies and rates of convergence for maximum likelihood and Bayes estimation for mixtures of normal densities". In: *Annals of Statististics* 29 (2001), pp. 1233–1263.

[20] S. Ghosal and A. van der Vaart. "Posterior convergence rates of Dirichlet mixtures at smooth densities". In: *Annals of Statistics* 35 (2007), pp. 697–723.

[21] Subhashis Ghosal. "The Dirichlet process, related priors and posterior asymptotics". In: *Bayesian Nonparametrics*. Cambridge University Press Walker, Cambridge, 2010.

[22] J.K. Ghosh and R.V. Ramamoorthi. *Bayesian Nonparametrics*. Springer Series in Statistics. Springer, 2003. ISBN: 9780387955377. URL: `https://books.google.com/books?id=4jhE7N23hRcC`.

[23] J. R. Gordon and D. Markwick. *dirichletprocess: Build Dirichlet process objects for Bayesian modelling*. R package version 0.3.0. 2019. URL: `https://CRAN.R-project.org/package=dirichletprocess`.

[24] A. Guha, N. Ho, and L. Nguyen. "On posterior contraction of parameters and interpretability in Bayesian mixture modeling". In: *Arxiv preprint Arxiv: 1901.05078* (2019).

[25] Wenge Guo, Gavin Lynch, and Joseph P Romano. "A new approach for large scale multiple testing with application to FDR control for graphically structured hypotheses". In: *arXiv preprint arXiv:1812.00258* (2018).

[26] John P Huelsenbeck and Peter Andolfatto. "Inference of population structure under a Dirichlet process model". In: *Genetics* 175 (2007), pp. 1787–1802.

[27] Nikolaos Ignatiadis et al. "Data-driven hypothesis weighting increases detection power in genome-scale multiple testing". In: *Nature methods* 13.7 (2016), pp. 577–580.

[28] Arnold Janssen. "Global power functions of goodness of fit tests". In: *Annals of Statistics* (2000), pp. 239–253.

[29] Nicolas Lartillot and Hervé Philippe. "A Bayesian mixture model for across-site heterogeneities in the amino-acid replacement process". In: *Molecular Biology and Evolution* 21.6 (2004), pp. 1095–1109.

[30] Lihua Lei and William Fithian. "AdaPT: an interactive procedure for multiple testing with side information". In: *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 80.4 (2018), pp. 649–679.

[31] Lihua Lei and William Fithian. "Power of ordered hypothesis testing". In: *International Conference on Machine Learning*. 2016, pp. 2924–2932.

[32] Lihua Lei, Aaditya Ramdas, and William Fithian. "A general interactive framework for false discovery rate control under structural constraints". In: *Biometrika* (2020).

[33] Ang Li and Rina Foygel Barber. "Accumulation tests for FDR control in ordered hypothesis testing". In: *Journal of the American Statistical Association* 112.518 (2017), pp. 837–849.

[34] Ang Li and Rina Foygel Barber. "Multiple testing with the structure adaptive Benjamini-Hochberg algorithm". In: *arXiv preprint arXiv:1606.07926* (2016).

[35] Zhonghua Liu and Xihong Lin. "A Geometric Perspective on the Power of Principal Component Association Tests in Multiple Phenotype Studies". In: *Journal of the American Statistical Association* (2019), pp. 1–32.

[36] A.Y. Lo. "On a class of Bayesian nonparametric estimates I: Density estimates". In: *Annals of Statistics* 12 (1984), pp. 351–357.

[37] Adam E Locke et al. "Genetic studies of body mass index yield new insights for obesity biology". In: *Nature* 518.7538 (2015), p. 197.

[38] S. MacEachern and P. Mueller. "Estimating mixture of Dirichlet process models". In: *Journal of Computational and Graphical Statistics* 7 (1998), pp. 223–238.

[39] Alisa K Manning et al. "A genome-wide approach accounting for body mass index identifies genetic variants influencing fasting glycemic traits and insulin resistance". In: *Nature genetics* 44.6 (2012), p. 659.

[40] Mario Medvedovic and Siva Sivaganesan. "Bayesian infinite mixture model based clustering of gene expression profiles". In: *Bioinformatics* 18 (2002), pp. 1194–1206.

[41] Jeffrey W Miller and Matthew T Harrison. "A simple example of Dirichlet process mixture inconsistency for the number of components". In: *Advances in Neural Information Processing Systems*. 2013, pp. 199–206.

[42] Jeffrey W Miller and Matthew T Harrison. "Inconsistency of Pitman-Yor process mixtures for the number of components". In: *Journal of Machine Learning Research* 15.1 (2014), pp. 3333–3370.

[43] Paul F O'Reilly et al. "MultiPhen: joint model of multiple phenotypes can increase discovery in GWAS". In: *PloS one* 7.5 (2012), e34861.

[44] Edoardo Otranto and Giampiero M Gallo. "A nonparametric Bayesian approach to detect the number of regimes in Markov switching models". In: *Econometric Reviews* 21 (2002), pp. 477–496.

[45] J. Paisley et al. "Nested hierarchical Dirichlet processes". In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 37 (2015), pp. 256–270.

[46] Jim Pitman and Marc Yor. "The two-parameter Poisson-Dirichlet distribution derived from a stable subordinator". In: *Annals of Probability* (1997), pp. 855–900.

[47] Shaun Purcell et al. "PLINK: a tool set for whole-genome association and population-based linkage analyses". In: *The American journal of human genetics* 81.3 (2007), pp. 559–575.

[48] Aaditya Ramdas et al. "DAGGER: A sequential algorithm for FDR control on DAGs". In: *arXiv preprint arXiv:1709.10250* (2017).

[49] A. Rodriguez, D. Dunson, and A.E. Gelfand. "The nested Dirichlet process". In: *Journal of the American Statistical Association* 103 (2008), pp. 1131–1154.

[50] J. Rousseau. "Rates of convergence for the posterior distributions of mixtures of Betas and adaptive nonparametric estimation of the density". In: *Annals of Statistics* 38 (2010), pp. 146–180.

[51] Sam Roweis. "EM algorithms for PCA and SPCA". In: *Advances in neural information processing systems* (1998), pp. 626–632.

[52] X. Shen and L. Wasserman. "Rates of convergence of posterior distributions". In: *Annals of Statistics* 29 (2001), pp. 687–714.

[53] Dmitry Shungin et al. "New genetic loci link adipose and insulin biology to body fat distribution". In: *Nature* 518.7538 (2015), p. 187.

[54] John D Storey, Jonathan E Taylor, and David Siegmund. "Strong control, conservative point estimation and simultaneous conservative consistency of false discovery rates: a unified approach". In: *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 66.1 (2004), pp. 187–205.

[55] Y. W. Teh et al. "Hierarchical Dirichlet processes". In: *Journal of the American Statistical Association* 101 (2006), pp. 1566–1581.

[56] Tanya M Teslovich et al. "Biological, clinical and population relevance of 95 loci for blood lipids". In: *Nature* 466.7307 (2010), p. 707.

[57] S. Walker, A. Lijoi, and I. Prunster. "On rates of convergence for posterior distributions in infinite-dimensional models". In: *Annals of Statistics* 35 (2007), pp. 738–746.

[58] Asaf Weinstein, Rina Barber, and Emmanuel Candes. "A power and prediction analysis for knockoffs with lasso statistics". In: *arXiv preprint arXiv:1712.06465* (2017).

Appendix A

Appendix

A.1 FDR Control

Proof to Lemma 1

Proof. The probability mass function for (V, U) is:

$$\mathbb{P}(V = v, U = k - v) = \frac{\binom{a}{v}\binom{b}{k-v}}{\binom{a+b}{k}},$$

where we define the binomial term $\binom{a}{k}$ and $\binom{b}{k-v}$ to be 0 when $k > a$ and $k - v > b$.

Let V^-, V^+ be the minimum and maximum numbers such that $\mathbb{P}(V = v, U = k - v) > 0$. Then we have:

$$\begin{aligned}
\mathbb{E}\Big[\frac{V}{1+U}\Big] &= \sum_{v=V^-}^{V^+} \frac{v}{1+k-v} \cdot \frac{\binom{a}{v}\binom{b}{k-v}}{\binom{a+b}{k}} \\
&= \frac{a}{1+b} \cdot \Big(1\{V^- > 0\}\frac{\binom{a-1}{V^--1}\binom{b+1}{k-v+1}}{\binom{a+b}{k}} + \sum_{v=V^-+1}^{V^+} \frac{\binom{a-1}{v-1}\binom{b+1}{k-v+1}}{\binom{a+b}{k}}\Big) \\
&\leq \frac{a}{1+b},
\end{aligned}$$

and thus we complete the first part, where we note that the result is not contingent on the choice of k, which can be any of $\{0, \cdots, a+b\}$. With a similar approach,

$$\mathbb{E}\Big[\frac{V}{1+U}\cdot\frac{b-U}{1+a-V}\Big] = \sum_{v=V^-}^{V^+}\frac{v\cdot(b-(k-v))}{(1+k-v))\cdot(1+a-v)}\cdot\mathbb{P}(V=v) \tag{A.1}$$

$$= \sum_{v=V^-}^{V^+}\frac{v\cdot(b-(k-v))}{(1+k-v))\cdot(1+a-v)}\cdot\frac{\binom{a}{v}\binom{b}{k-v}}{\binom{a+b}{k}} \tag{A.2}$$

$$= \sum_{v=V^-+1}^{V^+}\frac{\binom{a}{v-1}\binom{b}{k-v+1}}{\binom{a+b}{k}} \tag{A.3}$$

$$\leq 1, \tag{A.4}$$

where in (A.3) the term corresponding to V^- is 0 since either V^- is 0 or $b-(k-V^-)$ is 0. Thus the proof is complete. □

A.2 Linear Multivariate Gaussian Testing

As we have proved above, the FDR control of BONuS does not rely on specific model assumptions. However, the gain in power depends on the model we choose. In the following, we demonstrate a particular theoretical example in multiple testing for multivariate Gaussian distributions, first illustrating how problems can arise in a high dimensional testing and then demonstrating that the parameters are learnable via an maximum likelihood estimator (MLE) approach, where we also give a nonasymptotic upper bound on the error of the MLE.

Problem Statement and Result

Consider the Bayes two-group model with:

$$\mu^{(i)} \,|\, H^{(i)} \sim H^{(i)}\cdot\mathcal{N}_d(0,\mathbf{\Psi}) \tag{A.5}$$

$$X^{(i)} \,|\, \mu^{(i)}, H^{(i)} \sim \mathcal{N}_d(\mu^{(i)}, I_d) \tag{A.6}$$

An important case is that $\mathbf{\Psi}$ is a diagonal matrix that has only k nonzero terms, which corresponds to a feature selection problem, and when $k \ll d$, it reduces to a sparse problem. Here we consider a more general version where $\mathbf{\Psi}$ is a rank k matrix.

In this case, we have $\boldsymbol{X}^{(i)}|H^{(i)} = 1 \sim \mathcal{N}_d(0, \boldsymbol{\Psi} + I_d)$, so we may derive the likelihood ratio as:

$$\frac{f_\Lambda(\boldsymbol{x})}{f_{\mathbf{0}}(x)} \propto \boldsymbol{x}'\boldsymbol{x} - \boldsymbol{x}'(\boldsymbol{\Psi} + I_d)^{-1}\boldsymbol{x} \tag{A.7}$$

$$= \boldsymbol{x}'\Big(I_d - (\boldsymbol{\Psi} + I_d)^{-1}\Big)\boldsymbol{x} \tag{A.8}$$

If we know the matrix $\boldsymbol{\Psi}$, then we can use $\boldsymbol{X}'\Big(I_d-(\boldsymbol{\Psi}+I_d)^{-1}\Big)\boldsymbol{X}$ as an oracle test statistic. Hence, the problem reduces to finding the matrix $\boldsymbol{\Psi}$ from the data mixed with the synthetic samples. To formulate our problem more precisely, note that the data $\boldsymbol{X}^{(1)}, \ldots, \boldsymbol{X}^{(n)}$ can be thought to be i.i.d. samples from two-components mixture models with true density function $p_{G_*} := (1-\lambda^*)N(0, I_d) + \lambda^* N(0, I_d + \boldsymbol{\Psi}^*)$ where $G_* := (\lambda^*, \boldsymbol{\Psi}^*)$ such that $\lambda^* \in (0, 1)$ and $\boldsymbol{\Psi}^* \in \Omega$ are unknown parameters. Here, Ω is a set of positive definite matrices whose eigenvalues are upper bounded by some fixed constant. If we mix a fixed proportion of synthetic samples with the true samples, the resulting mixture is still a Gaussian mixture, so without loss of generality we may just ignore the synthetic samples here.

In this case of Gaussian 2-mixture model, we can show that the maximum likelihood estimator for $\boldsymbol{\Psi}$ is rate-optimal up to a $\Big(\log(n)\Big)^{1/2}$ factor. We use MLE to obtain an estimation of G_*, i.e., we have the following estimator

$$\widehat{G}_n := \underset{G\in(0,1)\times\Omega}{\arg\min} \sum_{i=1}^{n} \log(p_G(\boldsymbol{X}^{(i)})) \tag{A.9}$$

To talk about density estimation from MLE method, we will utilize the classical result from [16]. In particular, we have the following result regarding the density estimation $p_{\widehat{G}_n}$.

Proposition 13. *There exist some positive constants C and c depending only on d, Ω such that*

$$P\bigg(h(p_{\widehat{G}_n}, p_{G_*}) > C\Big(\frac{\log n}{n}\Big)^{1/2}\bigg) \leq \exp(-c\log n).$$

The proof of the above proposition is a direct application of Theorem 7.14 in [16]. Note that the standard result holds that the optimal convergence rate for parameter estimation in finite mixture model with a known number of component is $n^{-1/2}$, so MLE is rate optimal up to a $\Big(\log(n)\Big)^{1/2}$ factor.

Convergence rates of MLE Given the setup of MLE in equation (A.9), we have the following result regarding the convergence rates of $\widehat{G}_n$.

Theorem 14. *Assume that $\widehat{G}_n$ is given as in equation* (A.9)*. Then, the following holds:*

$$(a)\ \mathbb{P}\left(|\widehat{\lambda}_n - \lambda^*|\|\widehat{\boldsymbol{\Psi}}_n\|\|\boldsymbol{\Psi}^*\| > C_1\left(\frac{\log n}{n}\right)^{1/2}\right) \leq \exp(-c_1 \log n),$$

$$(b)\ \mathbb{P}\left(\lambda^*\|\boldsymbol{\Psi}^*\|\|\widehat{\boldsymbol{\Psi}}_n - \boldsymbol{\Psi}^*\| > C_1\left(\frac{\log n}{n}\right)^{1/2}\right) \leq \exp(-c_1 \log n).$$

Here, the probability $\mathbb{P}$ is taken with respect to density function p_{G_*}. Furthermore, C_1, c_1 are positive constants depending only on d and Ω.

The interesting feature in Theorem 14 is that both constants C_1 and c_1 are independent of λ^* and $\boldsymbol{\Psi}^*$. Therefore, the above results give a rigorous way to evaluate the convergence rates of $\widehat{\lambda}_n$ and $\widehat{\boldsymbol{\Psi}}_n$ when either λ^* goes to 0 or $\boldsymbol{\Psi}^*$ goes go $\vec{0}$ with the sample size. To further obtain the sole dependence of the convergence rate of $\widehat{\lambda}_n$ on $\|\boldsymbol{\Psi}^*\|$ in part (a) of Theorem 14, we will need to enforce more conditions on λ^* and $\|\boldsymbol{\Psi}^*\|$. In particular, we denote

$$\Theta_n(l_n) = \left\{G = (\lambda, \boldsymbol{\Psi}) : \ \frac{l_n}{\|\boldsymbol{\Psi}\|^2\sqrt{n}} \leq \lambda\right\}.$$

We have the following convergence result of $\widehat{\lambda}_n$ when $G_* \in \Theta_n(l_n)$:

Proposition 15. *Assume that the sequence l_n is chosen such that $l_n/\sqrt{\log(n)} \to \infty$ as $n \to \infty$. Then, as n is sufficiently large such that $C_1\sqrt{\log n}/l_n < 1/2$, we obtain*

$$P\left(|\widehat{\lambda}_n - \lambda^*|\|\boldsymbol{\Psi}^*\|^2 > 2C_1\left(\frac{\log n}{n}\right)^{1/2}\right) \leq 2\exp(-c_1 \log n)$$

as long as $G_* \in \Theta_n(l_n)$ where C_1, c_1 are two positive constants defined in Theorem 14.

Note that, the condition of $\Theta(l_n)$ is to guarantee that $\widehat{\boldsymbol{\Psi}}_n$ is the consistent estimator of $\boldsymbol{\Psi}^*$. The detail proof of Proposition 15 is deferred to Section A.2. A minimax result is also available but not related to our discussion here.

Proof of Theorem 14

Our approach to obtain the convergence rate of $\widehat{G}_n$ to G_* is based on the comparison between density estimation and parameter estimation, i.e., we would like to see how close $\widehat{G}_n$ to G_* as long as $p_{\widehat{G}_n}$ is close to p_{G_*}. In particular, we have the following result regarding such approach.

Theorem 16. *For any $G = (\lambda, \mathbf{\Psi})$ and $G_* = (\lambda^*, \mathbf{\Psi}^*)$, we denote*

$$\begin{aligned}\mathcal{D}(G, G_*) &:= l\|\mathbf{\Psi}\|^2 + \lambda^*\|\mathbf{\Psi}^*\|^2 - \min\{\lambda, \lambda^*\}\bigg(\|\mathbf{\Psi}\|^2 + \|\mathbf{\Psi}^*\|^2\bigg) \\ &+\bigg(\lambda\|\mathbf{\Psi}\| + \lambda^*\|\mathbf{\Psi}^*\|\bigg)\|\mathbf{\Psi} - \mathbf{\Psi}^*\|.\end{aligned} \tag{A.10}$$

Then, there exists a positive constant C depending only on d and Ω such that

$$\|p_G - p_{G_*}\|_1 \geq C \cdot \mathcal{D}(G, G_*)$$

for all G and G_*.

Remark: We can verify that

$$\mathcal{D}(G, G_*) \asymp \mathcal{D}_1(G, G_*) = |\lambda - \lambda^*|\|\mathbf{\Psi}\|\|\mathbf{\Psi}^*\| + \bigg(\lambda\|\mathbf{\Psi}\| + \lambda^*\|\mathbf{\Psi}^*\|\bigg)\|\mathbf{\Psi} - \mathbf{\Psi}^*\|$$

for any $\mathbf{\Psi}$ and $\mathbf{\Psi}^*$. Therefore. we also can obtain the lower bound of L_1 norm between p_G and p_{G_*} in terms of $D_1(G, G_*)$. This particular lower bound is useful for deriving the convergence rates of MLE estimation later.

To achieve the conclusion of the theorem, we firstly demonstrate the following result

Proposition 17. *Denote $\overline{G} = (\overline{\lambda}, \overline{\mathbf{\Psi}})$ such that $\overline{\lambda} \in [0, 1]$ and $\overline{\mathbf{\Psi}}$ can be identical to $\vec{0}$ where $\vec{0}$ denotes matrix with all elements to be 0. Then, the following holds*

$$\varliminf_{\epsilon \to 0} \inf_{G, G_*} \bigg\{ \frac{\|p_G - p_{G_*}\|_\infty}{\mathcal{D}(G, G_*)} : \mathcal{D}(G, \overline{G}) \vee \mathcal{D}(G_*, \overline{G}) \leq \epsilon \bigg\} > 0.$$

Proof. Throughout this proof, we denote $f_1(\boldsymbol{x}|\mathbf{\Psi})$ to be the density of $N(0, I_d + \mathbf{\Psi})$. Here, we only consider the most challenging setting that $\overline{\mathbf{\Psi}} = \vec{0}$ as the proof for other possibilities of $\overline{\mathbf{\Psi}}$ and $\overline{\lambda}$ can be argued in the similar fashion. Assume that the conclusion of Proposition 17 does not hold. It implies that we can find two sequences $G_n = (\lambda_n, \mathbf{\Psi}_n)$ and $G_{*,n} = (\lambda_n^*, \mathbf{\Psi}_n^*)$ such that $\mathcal{D}(G_n, \overline{G}) = \lambda_n\|\mathbf{\Psi}_n\|^2 \to 0$,

$\mathcal{D}(G_{*,n}, \overline{G}) = \lambda_n^* \|\boldsymbol{\Psi}_n^*\|^2 \to 0$, and $\|p_{G_n} - p_{G_{*,n}}\|_\infty / \mathcal{D}(G_n, G_{*,n}) \to 0$ as $n \to \infty$. For the transparency of presentation, we denote $A_n = \|\boldsymbol{\Psi}_n\|$, $B_n = \|\boldsymbol{\Psi}_n^*\|$, and $C_n = \|\boldsymbol{\Psi}_n - \boldsymbol{\Psi}_n^*\|$. Now, we have three main cases regarding the convergence behaviors of $\boldsymbol{\Psi}_n$ and $\boldsymbol{\Psi}_n^*$

Case 1: Both $A_n \to 0$ and $B_n \to 0$, i.e., $\boldsymbol{\Psi}_n$ and $\boldsymbol{\Psi}_n^*$ vanish to $\vec{0}$ as $n \to \infty$. Due to the symmetry between λ_n and λ_n^*, we assume without loss of generality that $\lambda_n^* \geq \lambda_n$ for infinite values of n. Without loss of generality, we replace these subsequences of $G_n, G_{*,n}$ by the whole sequences of G_n and $G_{*,n}$. Now, the formulation of $\mathcal{D}(G_n, G_{*,n})$ is

$$\mathcal{D}(G_n, G_{*,n}) = (\lambda_n^* - \lambda_n) B_n^2 + \Big(\lambda_n A_n + \lambda_n^* B_n\Big) C_n.$$

Now, by means of Taylor expansion up to the second order, we get

$$\begin{aligned}
\frac{p_{G_n}(\boldsymbol{x}) - p_{G_{*,n}}(\boldsymbol{x})}{\mathcal{D}(G_n, G_{*,n})} &= \frac{(\lambda_n^* - \lambda_n)[f_1(\boldsymbol{x}|\vec{0}) - f_1(\boldsymbol{x}|\boldsymbol{\Psi}_n^*)] + \lambda_n[f_1(\boldsymbol{x}|\boldsymbol{\Psi}_n) - f_1(\boldsymbol{x}|\boldsymbol{\Psi}_n^*)]}{\mathcal{D}(G_n, G_{*,n})} \\
&= \frac{(\lambda_n^* - \lambda_n)\Big(\sum\limits_{|\alpha|=1}^{2} \frac{(-\boldsymbol{\Psi}_n^*)^\alpha}{\alpha!} \frac{\partial^{|\alpha|} f_1}{\partial \boldsymbol{\Psi}^\alpha}(\boldsymbol{x}|\boldsymbol{\Psi}_n^*) + R_1(\boldsymbol{x})\Big)}{\mathcal{D}(G_n, G_{*,n})} \\
&+ \frac{\lambda_n\Big(\sum\limits_{|\alpha|=1}^{2} \frac{(\boldsymbol{\Psi}_n - \boldsymbol{\Psi}_n^*)^\alpha}{\alpha!} \frac{\partial^{|\alpha|} f_1}{\partial \boldsymbol{\Psi}^\alpha}(\boldsymbol{x}|\boldsymbol{\Psi}_n^*) + R_2(\boldsymbol{x})\Big)}{\mathcal{D}(G_n, G_{*,n})},
\end{aligned}$$

where $R_1(\boldsymbol{x})$ and $R_2(\boldsymbol{x})$ are Taylor remainders that satisfy $R_1(\boldsymbol{x}) = O(B_n^{2+\gamma})$ and $R_2(\boldsymbol{x}) = O(C_n^{2+\gamma})$ for some positive number γ due to the smoothness of Gaussian kernel. From the formation of $\mathcal{D}(G_n, G_{*,n})$, since $A_n + B_n \geq C_n$ (triangle inequality), as $A_n \to 0$ and $B_n \to 0$ it is clear that

$$\begin{aligned}
(\lambda_n - \lambda_n^*)|R_1(\boldsymbol{x})|/\mathcal{D}(G_n, G_{*,n}) &\leq |R_1(\boldsymbol{x})|/B_n^2 = O(B_n^\gamma) \to 0, \\
\lambda_n |R_2(\boldsymbol{x})|/\mathcal{D}(G_n, G_{*,n}) \leq |R_2(\boldsymbol{x})|/\left\{(A_n + B_n)C_n\right\} &= O\Big(C_n^{2+\gamma}/C_n^2\Big) = O(C_n^\gamma) \to 0,
\end{aligned}$$

as $n \to \infty$ for all $\boldsymbol{x} \in \mathcal{X}$. Therefore, we achieve for all $\boldsymbol{x} \in \mathcal{X}$ that

$$\Big((\lambda_n - \lambda_n^*)|R_1(\boldsymbol{x})| + \lambda_n |R_2(\boldsymbol{x})|\Big)/\mathcal{D}(G_n, G_{*,n}) \to 0.$$

Hence, we can treat $[p_{G_n}(\boldsymbol{x}) - p_{G_{*,n}}(\boldsymbol{x})]/\mathcal{D}(G_n, G_{*,n})$ as a linear combination of $\frac{\partial^{|\alpha|} f_1}{\partial \boldsymbol{\Psi}^\alpha}(\boldsymbol{x}|\boldsymbol{\Psi}_n^*)$ for all $\boldsymbol{x}$ and $\alpha \in \mathbb{N}^{d\times d}$ such that $1 \leq |\alpha| \leq 2$. Assume that all the coefficients of these terms go to 0 as $n \to \infty$. By studying the vanishing behaviors of the coefficients of $\frac{\partial^{|\alpha|} f_1}{\partial \boldsymbol{\Psi}^\alpha}(\boldsymbol{x}|\boldsymbol{\Psi}_n^*)$ as $|\alpha| = 1$, we achieve the following limits

$$\Big(\lambda_n(\boldsymbol{\Psi}_n)_{uv} - \lambda_n^*(\boldsymbol{\Psi}_n^*)_{uv}\Big)/\mathcal{D}(G_n, G_{*,n}) \to 0$$

for all $1 \leq u, v \leq d$ where A_{uv} denotes the (u, v)-th element of matrix A. For any two pairs $(u_1, v_1), (u_2, v_2)$ (not neccessarily distinct) such that $1 \leq u_1, u_2, v_1, v_2 \leq d$, the coefficients of $\frac{\partial^{|\alpha|} f_1}{\partial \boldsymbol{\Psi}^\alpha}(\boldsymbol{x}|\boldsymbol{\Psi}_n^*)$ when $(\alpha)_{u_1v_1} = (\alpha)_{u_2v_2} = 1$ leads to

$$\frac{(\lambda_n^* - \lambda_n)(-\boldsymbol{\Psi}_n^*)_{u_1v_1}(-\boldsymbol{\Psi}_n^*)_{u_2v_2} + \lambda_n(\boldsymbol{\Psi}_n - \boldsymbol{\Psi}_n^*)_{u_1v_1}(\boldsymbol{\Psi}_n - \boldsymbol{\Psi}_n^*)_{u_2v_2}}{\mathcal{D}(G_n, G_{*,n})} \to 0. \quad \text{(A.11)}$$

When $(u_1, v_1) \equiv (u_2, v_2)$, the above limits lead to

$$\Big[(\lambda_n^* - \lambda_n)(\boldsymbol{\Psi}_n^*)_{u_1v_1}^2 + \lambda_n(\boldsymbol{\Psi}_n - \boldsymbol{\Psi}_n^*)_{u_1v_1}^2\Big]/\mathcal{D}(G_n, G_{*,n}) \to 0.$$

Therefore, we would have that

$$\Big[(\lambda_n^* - \lambda_n)\|\boldsymbol{\Psi}_n^*\|^2 + \lambda_n\|\boldsymbol{\Psi}_n - \boldsymbol{\Psi}_n^*\|^2\Big]/\mathcal{D}(G_n, G_{*,n}) \to 0. \quad \text{(A.12)}$$

Now, as $\Big(\lambda_n(\boldsymbol{\Psi}_n)_{uv} - \lambda_n^*(\boldsymbol{\Psi}_n^*)_{uv}\Big)/\mathcal{D}(G_n, G_{*,n}) \to 0$ for all $1 \leq u, v \leq d$, we obtain that

$$\begin{aligned}
\Big(\lambda_n(\boldsymbol{\Psi}_n)_{u_1v_1}(\boldsymbol{\Psi}_n)_{u_2v_2} - \lambda_n^*(\boldsymbol{\Psi}_n^*)_{u_1v_1}(\boldsymbol{\Psi}_n)_{u_2v_2}\Big)/\mathcal{D}(G_n, G_{*,n}) &\to 0, \\
\Big(\lambda_n(\boldsymbol{\Psi}_n)_{u_1v_1}(\boldsymbol{\Psi}_n^*)_{u_2v_2} - \lambda_n^*(\boldsymbol{\Psi}_n^*)_{u_1v_1}(\boldsymbol{\Psi}_n^*)_{u_2v_2}\Big)/\mathcal{D}(G_n, G_{*,n}) &\to 0
\end{aligned}$$

for any two pairs $(u_1, v_1), (u_2, v_2)$. The above results imply that

$$\begin{aligned}
\Big[(\lambda_n^* - \lambda_n)(-\boldsymbol{\Psi}_n^*)_{u_1v_1}(-\boldsymbol{\Psi}_n^*)_{u_2v_2} + \lambda_n(\boldsymbol{\Psi}_n - \boldsymbol{\Psi}_n^*)_{u_1v_1}(\boldsymbol{\Psi}_n - \boldsymbol{\Psi}_n^*)_{u_2v_2} \\
+(\lambda_n - \lambda_n^*)(\boldsymbol{\Psi}_n^*)_{u_1v_1}(\boldsymbol{\Psi}_n)_{u_2v_2}\Big]/\mathcal{D}(G_n, G_{*,n}) \to 0. \quad \text{(A.13)}
\end{aligned}$$

By combining the results from (A.11) and (A.13), we ultimately achieve for any two pairs (u_1, v_1) and (u_2, v_2) that

$$(\lambda_n - \lambda_n^*)(\boldsymbol{\Psi}_n^*)_{u_1v_1}(\boldsymbol{\Psi}_n)_{u_2v_2}/\mathcal{D}(G_n, G_{*,n}) \to 0. \tag{A.14}$$

Using the results from equation (A.11) and (A.14), we have

$$\begin{aligned}
\frac{\lambda_n(\boldsymbol{\Psi}_n)_{u_1v_1}(\boldsymbol{\Psi}_n - \boldsymbol{\Psi}_n^*)_{u_2v_2}}{\mathcal{D}(G_n, G_{*,n})} \to \frac{(\lambda_n^* - \lambda_n)(\boldsymbol{\Psi}_n)_{u_1v_1}(\boldsymbol{\Psi}_n^*)_{u_2v_2}}{\mathcal{D}(G_n, G_{*,n})} &\to 0, \\
\frac{\lambda_n^*(\boldsymbol{\Psi}_n^*)_{u_1v_1}(\boldsymbol{\Psi}_n - \boldsymbol{\Psi}_n^*)_{u_2v_2}}{\mathcal{D}(G_n, G_{*,n})} \to \frac{(\lambda_n^* - \lambda_n)(\boldsymbol{\Psi}_n^*)_{u_1v_1}(\boldsymbol{\Psi}_n)_{u_2v_2}}{\mathcal{D}(G_n, G_{*,n})} &\to 0
\end{aligned}$$

for any two pairs (u_1, v_1) and (u_2, v_2). Therefore, it leads to

$$\begin{aligned}
\frac{\sum\limits_{(u_1,v_1),(u_2,v_2)} \lambda_n |(\boldsymbol{\Psi}_n)_{u_1v_1}||(\boldsymbol{\Psi}_n - \boldsymbol{\Psi}_n^*)_{u_2v_2}|}{\mathcal{D}(G_n, G_{*,n})} &= \frac{\lambda_n \sum\limits_{(u_1,v_1)} |(\boldsymbol{\Psi}_n)_{u_1v_1}| \sum\limits_{(u_2,v_2)} |(\boldsymbol{\Psi}_n - \boldsymbol{\Psi}_n^*)_{u_2v_2}|}{\mathcal{D}(G_n, G_{*,n})} \to 0, \\
\frac{\sum\limits_{(u_1,v_1),(u_2,v_2)} \lambda_n^* |(\boldsymbol{\Psi}_n^*)_{u_1v_1}||(\boldsymbol{\Psi}_n - \boldsymbol{\Psi}_n^*)_{u_2v_2}|}{\mathcal{D}(G_n, G_{*,n})} &= \frac{\lambda_n^* \sum\limits_{(u_1,v_1)} |(\boldsymbol{\Psi}_n^*)_{u_1v_1}| \sum\limits_{(u_2,v_2)} |(\boldsymbol{\Psi}_n - \boldsymbol{\Psi}_n^*)_{u_2v_2}|}{\mathcal{D}(G_n, G_{*,n})} \to 0.
\end{aligned}$$

The above results indicate that

$$\lambda_n \|\boldsymbol{\Psi}_n\| \|\boldsymbol{\Psi}_n - \boldsymbol{\Psi}_n^*\|/\mathcal{D}(G_n, G_{*,n}) \to 0, \qquad \lambda_n^* \|\boldsymbol{\Psi}_n^*\| \|\boldsymbol{\Psi}_n - \boldsymbol{\Psi}_n^*\|/\mathcal{D}(G_n, G_{*,n}) \to 0 \tag{A.15}$$

Combining the results from (A.12) and (A.15), we have

$$1 = \mathcal{D}(G_n, G_{*,n})/\mathcal{D}(G_n, G_{*,n}) \to 0,$$

which is a contradiction. As a consequence, not all the coefficients of $\dfrac{\partial^{|\alpha|} f_1}{\partial \boldsymbol{\Psi}^\alpha}(\boldsymbol{x}|\boldsymbol{\Psi}_n^*)$ go to 0 as $1 \leq |\alpha| \leq 2$. By denoting m_n to be the maximum of the absolute values of the coefficients of $\dfrac{\partial^{|\alpha|} f_1}{\partial \boldsymbol{\Psi}^\alpha}(\boldsymbol{x}|\boldsymbol{\Psi}_n^*)$ we achieve for all $\boldsymbol{x}$ that

$$\frac{1}{m_n} \frac{p_{G_n}(\boldsymbol{x}) - p_{G_{*,n}}(\boldsymbol{x})}{\mathcal{D}(G_n, G_{*,n})} \to \sum_{|\alpha|=1}^{2} \tau_\alpha \frac{\partial^{|\alpha|} f_1}{\partial \boldsymbol{\Psi}^\alpha}(\boldsymbol{x}|0) = 0$$

where $\tau_\alpha \in \mathbb{R}$ are some coefficients such that not all of them are 0. We can check that the previous equation only holds when $\tau_\alpha = 0$ for all $1 \leq |\alpha| \leq 2$, which is a contradiction. As a consequence, Case 1 cannot happen.

Case 2: Exactly one of A_n and B_n goes to 0, i.e., there exists at least one component among $\boldsymbol{\Psi}_n$ and $\boldsymbol{\Psi}_n^*$ that does not converge to $\vec{0}$ as $n \to \infty$. Due to the symmetry of A_n and B_n, we assume without loss of generality that $A_n \not\to 0$ and $B_n \to 0$, which is equivalent to $\boldsymbol{\Psi}_n \to \boldsymbol{\Psi}' \neq \vec{0}$ while $\boldsymbol{\Psi}_n^* \to \vec{0}$ as $n \to \infty$. We denote

$$\mathcal{D}'(G_n, G_{*,n}) = |\lambda_n^* - \lambda_n| B_n + \lambda_n A_n + \lambda_n^* B_n.$$

Since $[p_{G_n}(\boldsymbol{x}) - p_{G_{*,n}}(\boldsymbol{x})]/\mathcal{D}(G_n, G_{*,n}) \to 0$ and $\mathcal{D}(G_n, G_{*,n}) \lesssim \mathcal{D}'(G_n, G_{*,n})$, it follows that the ratio $[p_{G_n}(\boldsymbol{x}) - p_{G_{*,n}}(\boldsymbol{x})]/\mathcal{D}'(G_n, G_{*,n}) \to 0$ for all $\boldsymbol{x}$. By means of Taylor expansion up to the first order, we have

$$\begin{aligned}
\frac{p_{G_n}(\boldsymbol{x}) - p_{G_{*,n}}(\boldsymbol{x})}{\mathcal{D}'(G_n, G_{*,n})} &= \frac{(\lambda_n^* - \lambda_n)[f_1(\boldsymbol{x}|\vec{0}) - f_1(\boldsymbol{x}|\boldsymbol{\Psi}_n^*)] + \lambda_n f_1(\boldsymbol{x}|\boldsymbol{\Psi}_n) - \lambda_n f_1(\boldsymbol{x}|\boldsymbol{\Psi}_n^*)}{\mathcal{D}'(G_n, G_{*,n})} \\
&= \frac{(\lambda_n^* - \lambda_n)\left(\sum\limits_{|\alpha|=1} \frac{(-\boldsymbol{\Psi}_n^*)^\alpha}{\alpha!} \frac{\partial^{|\alpha|} f_1}{\partial \boldsymbol{\Psi}^\alpha}(\boldsymbol{x}|\boldsymbol{\Psi}_n^*) + R_1'(\boldsymbol{x})\right)}{\mathcal{D}'(G_n, G_{*,n})} \\
&+ \frac{\lambda_n f_1(\boldsymbol{x}|\boldsymbol{\Psi}_n) - \lambda_n f_1(\boldsymbol{x}|\boldsymbol{\Psi}_n^*)}{\mathcal{D}'(G_n, G_{*,n})}
\end{aligned}$$

where $R_1'(\boldsymbol{x})$ is Taylor remainder that satisfies $(\lambda_n^* - \lambda_n)|R_1'(\boldsymbol{x})|/\mathcal{D}'(G_n, G_{*,n}) = O(B_n^{\gamma'}) \to 0$ for some positive number $\gamma' > 0$. Since $\boldsymbol{\Psi}_n$ and $\boldsymbol{\Psi}_n^*$ do not have the same limit, they will be different when n is large enough, i.e., $n \geq M'$ for some value of M'. Now, as $n \geq M'$, $[p_{G_n}(\boldsymbol{x}) - p_{G_{*,n}}(\boldsymbol{x})]/\mathcal{D}'(G_n, G_{*,n})$ becomes a linear combination of $\frac{\partial^{|\alpha|} f_1}{\partial \boldsymbol{\Psi}^\alpha}(\boldsymbol{x}|\boldsymbol{\Psi}_n^*)$ for all $|\alpha| \leq 1$ and $f_1(\boldsymbol{x}|\boldsymbol{\Psi}_n)$. If all of the coefficients of these terms go to 0, we would have $\lambda_n/\mathcal{D}'(G_n, G_{*,n}) \to 0$ and $(\lambda_n^* - \lambda_n)(-\boldsymbol{\Psi}_n^*)_{uv}/\mathcal{D}'(G_n, G_{*,n}) \to 0$ for all $1 \leq u, v \leq d$. It implies that $(\lambda_n^* - \lambda_n)B_n/\mathcal{D}'(G_n, G_{*,n}) \to 0$, $\lambda_n A_n/\mathcal{D}'(G_n, G_{*,n}) \to 0$, and $\lambda_n B_n/\mathcal{D}'(G_n, G_{*,n}) \to 0$. These results lead to

$$1 = \left(|\lambda_n^* - \lambda_n| B_n + \lambda_n A_n + \lambda_n^* B_n\right)/\mathcal{D}'(G_n, G_{*,n}) \to 0,$$

a contradiction. Therefore, not all the coefficients of $\frac{\partial^{|\alpha|} f_1}{\partial \boldsymbol{\Psi}^\alpha}(\boldsymbol{x}|\boldsymbol{\Psi}_n^*)$ and $f_1(\boldsymbol{x}|\boldsymbol{\Psi}_n)$ go to 0. By defining m_n' to be the maximum of these coefficients, we achieve for all $\boldsymbol{x}$ that

$$\frac{1}{m_n'} \frac{p_{G_n}(\boldsymbol{x}) - p_{G_{*,n}}(\boldsymbol{x})}{\mathcal{D}'(G_n, G_{*,n})} \to \eta' f_1(\boldsymbol{x}|\vec{0}) + \sum_{|\alpha|=0}^{1} \tau_\alpha' \frac{\partial^{|\alpha|} f_1}{\partial \boldsymbol{\Psi}^\alpha}(\boldsymbol{x}|\boldsymbol{\Psi}') = 0,$$

where η' and τ_α' are coefficients such that not all of them are 0, which is a contradiction to the first order identifiability of Gaussian distribution with only covariance parameter. As a consequence, Case 2 cannot hold.

Case 3: Both A_n and B_n do not go to 0, i.e., $\boldsymbol{\Psi}_n$ and $\boldsymbol{\Psi}_n^*$ do not converge to $\vec{0}$ as $n \to \infty$. Since $\mathcal{D}_n(G_n, G_{*,n}) \lesssim \mathcal{K}(G_n, G_{*,n}) = |\lambda_n - \lambda_n^*| + (\lambda_n + \lambda_n^*)C_n$ and $[p_{G_n}(\boldsymbol{x}) - p_{G_{*,n}}(\boldsymbol{x})]/\mathcal{D}(G_n, G_{*,n}) \to 0$, we achieve that $[p_{G_n}(\boldsymbol{x}) - p_{G_{*,n}}(\boldsymbol{x})]/\mathcal{K}(G_n, G_{*,n}) \to 0$ for all $\boldsymbol{x}$. From here, by using the same argument as that of Case 1 and Case 2, we also reach the contradiction. Therefore, Case 3 cannot happen.

In sum, we achieve the conclusion of the proposition. □

Now, assume that the conclusion of Theorem 16 does not hold. It implies that we can find two sequences G'_n and $G'_{*,n}$ such that $F_n = \|p_{G'_n} - p_{G'_{*,n}}\|_1/\mathcal{D}(G'_n, G'_{*,n}) \to 0$ as $n \to \infty$. Since Ω is bounded set of positive definite matrices, we can find subsequences of G'_n and $G'_{*,n}$ such that $\mathcal{D}(G'_n, \overline{G}_1)$ and $\mathcal{D}(G'_{*,n}, \overline{G}_2)$ vanish to 0 as $n \to \infty$ where $\overline{G}_1, \overline{G}_2$ are some parameters in $[0, 1] \times \Omega$. Because $F_n \to 0$, we obtain $\|p_{G'_n} - p_{G'_{*,n}}\|_1 \to 0$ as $n \to \infty$. By means of Fatou's lemma, we have

$$0 = \lim_{n\to\infty} \int |p_{G'_n}(x) - p_{G'_{*,n}}(x)|dx \geq \int \liminf_{n\to\infty} |p_{G'_n}(\boldsymbol{x}) - p_{G'_{*,n}}(\boldsymbol{x})|d\boldsymbol{x} = ||p_{\overline{G}_1} - p_{\overline{G}_2}||_1.$$

Due to the fact that Gaussian is identifiable, the above equation implies that $\overline{G}_1 \equiv \overline{G}_2$. However, from the result of Proposition 17, regardless of the value of $\overline{G}_1$ we would have $F_n \not\to 0$ as $n \to \infty$, which is a contradiction. Therefore, we obtain the conclusion of the theorem.

Proof of Proposition 15

Let n be such that $C_1\sqrt{\log n}/l_n < 1/2$ where C_1 is positive constant in Theorem 14. Then, we obtain that

$$\begin{aligned}
P\left(\|\widehat{\boldsymbol{\Psi}}_n\| < \frac{\|\boldsymbol{\Psi}^*\|}{2}\right) &= P\left(\|\boldsymbol{\Psi}^*\| - \|\widehat{\boldsymbol{\Psi}}_n\| > \frac{\|\boldsymbol{\Psi}^*\|}{2}\right) \leq P\left(\|\boldsymbol{\Psi}^* - \widehat{\boldsymbol{\Psi}}_n\| > \frac{\|\boldsymbol{\Psi}^*\|}{2}\right). \\
&\leq P\left(\|\boldsymbol{\Psi}^* - \widehat{\boldsymbol{\Psi}}_n\| > \frac{C_1\sqrt{\log n}\|\boldsymbol{\Psi}^*\|}{l_n}\right) \\
&\leq P\left(\|\boldsymbol{\Psi}^* - \widehat{\boldsymbol{\Psi}}_n\| > \frac{C_1\sqrt{\log n}\|\boldsymbol{\Psi}^*\|}{\lambda^*\|\boldsymbol{\Psi}^*\|^2\sqrt{n}}\right) \\
&= P\left(\lambda^*\|\boldsymbol{\Psi}^*\|\|\widehat{\boldsymbol{\Psi}}_n - \boldsymbol{\Psi}^*\| > C_1\left(\frac{\log n}{n}\right)^{1/2}\right) \leq \exp(-c_1 \log n)
\end{aligned}$$

where c_1 is positive constant defined in Theorem 14. The above inequality leads to

$$
\begin{aligned}
& P\left(|\widehat{\lambda}_n - \lambda^*| \|\mathbf{\Psi}^*\|^2 > 2C_1 \left(\frac{\log n}{n}\right)^{1/2}\right) \qquad \text{(A.16)} \\
= \; & P\left(|\widehat{\lambda}_n - \lambda^*| \|\mathbf{\Psi}^*\|^2 > 2C_1 \left(\frac{\log n}{n}\right)^{1/2}, \|\widehat{\mathbf{\Psi}}_n\| \geq \frac{\|\mathbf{\Psi}^*\|}{2}\right) \\
& + P\left(|\widehat{\lambda}_n - \lambda^*| \|\mathbf{\Psi}^*\|^2 > 2C_1 \left(\frac{\log n}{n}\right)^{1/2}, \|\widehat{\mathbf{\Psi}}_n\| < \frac{\|\mathbf{\Psi}^*\|}{2}\right) \\
\leq \; & P\left(|\widehat{\lambda}_n - \lambda^*| \|\widehat{\mathbf{\Psi}}_n\| \|\mathbf{\Psi}^*\|^2 > C_1 \left(\frac{\log n}{n}\right)^{1/2}\right) + P\left(\|\widehat{\mathbf{\Psi}}_n\| < \frac{\|\mathbf{\Psi}^*\|}{2}\right) \\
\leq \; & 2\exp(-c_1 \log n).
\end{aligned}
$$

We obtain the conclusion of the proposition.