Adaptive Statistical Inference Procedures for Multigroup Data and Phylogenetic Tree Inferences

Chaoyu Yu

A dissertation
submitted in partial fulfillment of the
requirements for the degree of

Doctor of Philosophy

University of Washington

2018

Reading Committee:

Peter Hoff, Chair

Mathias Drton, Chair

Lurdes Inoue

Program Authorized to Offer Degree:
Biostatistics

University of Washington

Abstract

Adaptive Statistical Inference Procedures for Multigroup Data and Phylogenetic Tree Inferences

Chaoyu Yu

Co-Chairs of the Supervisory Committee:
Professor Peter Hoff
Department of Statistics

Professor Mathias Drton
Department of Statistics

Multigroup data is a common data type in fields such as biology, the environmental sciences and the social sciences. This dissertation focuses on developing new statistical methodologies for multigroup data analysis. When data across groups are independent of each other, simultaneous statistical inferences for each group are often performed to analyze the data. We first present an adaptive multigroup confidence interval procedure. We construct confidence intervals that make use of information about across-group heterogeneity, resulting in constant coverage intervals that are narrower than standard t-intervals across groups. Then we present adaptive procedures for sign error control. We present a procedure that guarantees to control the sign error rate under a desired threshold, and another more powerful procedure that approximately controls the sign error rate under certain assumptions. When data across groups are dependent on each other, it is often of interest to capture the dependence relationships among groups. For the second part of the dissertation, we develop methodologies for such data type with a focus on phylogenetic tree inferences. We first present simple and consistent algorithms for the tree topology recovery and parameter estimation, and then present an iterative structural EM algorithm which improves the results from the simple algorithms.

TABLE OF CONTENTS

LIST OF FIGURES

ACKNOWLEDGMENTS

I would like to thank my dissertation advisors, Peter Hoff and Mathias Drton, who have patiently guided me through my doctoral study. Their encouragement, insights and support have been instrumental to my development as a statistician. I would also like to thank Lurdes Inoue for her advice and help. I am also extremely grateful for my research assistant advisors, Rozenn Lemaitre, Barbara McKnight and Ken Rice, for their guidance and generous support. I also owe my thanks to other faculty and staff in the department of Biostatistics and Statistics for making my time in graduate school wonderful.

DEDICATION

to my family

Chapter 1
INTRODUCTION

In this introductory chapter, we begin with the motivation for new statistical inference procedures for multigroup data and introduce relevant notation. Then we will introduce some graphical modeling terminologies and relevant background for phylogenetics. We finish this chapter with an outline of the dissertation.

1.1 Estimation and Inference with Multivariate Gaussian Distribution

Consider a p-variate random vector $\mathbf{Y}$ with a multivariate Gaussian distribution

$$\mathbf{Y} \sim N_p(\boldsymbol{\theta}, \sigma^2 I) \tag{1.1}$$

with an unknown mean vector $\boldsymbol{\theta}$ and known variance σ^2. The statistical tasks of interest are estimation and inference on the parameter $\boldsymbol{\theta}$ based on an observation $\mathbf{y}$ of $\mathbf{Y}$. Without loss of generality, we assume $\sigma^2 = 1$ in this chapter.

One straightforward estimator for $\boldsymbol{\theta}$, is the maximum likelihood estimator $\hat{\boldsymbol{\theta}}_{ML} = \mathbf{y}$, which is also referred to as the least square estimator. This is an unbiased estimator of the parameter $\boldsymbol{\theta}$, and has the minimum variance among all unbiased estimators of $\boldsymbol{\theta}$. However, Stein (1956) and James and Stein (1961a) shocked the statistical world by showing that $\boldsymbol{\theta}_{ML}$ is inadmissible when $p > 2$. The estimator proposed by them is usually referred to as the James-Stein estimator and takes the form

$$\hat{\boldsymbol{\theta}}_{JS} = \left(1 - \frac{p-2}{||\mathbf{y}||_2^2}\right)\mathbf{y}.$$

This estimator has a mean squared error lower than $\hat{\boldsymbol{\theta}}_{ML}$. James-Stein estimator played an important role in the development of high-dimensional statistics analysis and empirical

Bayes approaches. The empirical Bayes interpretation of James-Stein estimator can be seen by first assuming

$$\boldsymbol{\theta} \sim N(0, \tau^2 I). \tag{1.2}$$

The Bayes estimator of $\boldsymbol{\theta}$ is then

$$\hat{\boldsymbol{\theta}}_{Bayes} = \left(1 - \frac{1}{\tau^2 + 1}\right) \mathbf{y}, \tag{1.3}$$

which is the estimator that minimizes the expected squared error given $\mathbf{y}$. Since τ^2 is actually an unknown parameter, in light of empirical Bayes theories, we estimate this parameter from the data $\mathbf{y}$. Observe that the marginal distribution of $\mathbf{y}$ is

$$\mathbf{y} \sim N(0, (\tau^2 + 1)I). \tag{1.4}$$

This implies that

$$||\mathbf{y}||^2 \sim (\tau^2 + 1)\chi_p^2,$$

and we have

$$\mathrm{E}\left[\frac{p-2}{||\mathbf{y}||^2}\right] = \frac{1}{\tau^2 + 1}.$$

By replacing the unknown term $\frac{1}{\tau^2+1}$ in equation (1.3) with its unbiased estimator $\frac{p-2}{||\mathbf{y}||^2}$, we obtained the empirical Bayes estimator of $\boldsymbol{\theta}$, which is of the exactly same form as $\hat{\boldsymbol{\theta}}_{JS}$. Note that $1/(\tau^2+1) < 1$, we have that $\hat{\boldsymbol{\theta}}_{Bayes}$ is closer to $\mathbf{0}$ than $\mathbf{y}$ is. Hence we also call $\hat{\boldsymbol{\theta}}_{Bayes}$ and $\hat{\boldsymbol{\theta}}_{\boldsymbol{JS}}$ the shrinkage estimators. The technique of shrinkage estimation has a wide application in high-dimensional statistics analysis.

1.2 Simultaneous Confidence Intervals

Although there is a great amount of literature studying the empirical Bayes point estimators or shrinkage type estimators, the application of this idea to the confidence set or interval development is significantly less. In spite of its importance, the technical difficulty has been one of the major reason for this. A confidence procedure for a single parameter of interest θ with $1 - \alpha$ frequentist coverage is defined as a region C such that

$$\Pr(\theta \in C|\theta) = 1 - \alpha \ \forall \theta. \tag{1.5}$$

For multivariate case, i.e. the parameter of interest $\boldsymbol{\theta}$ is a p-variate vector, either simultaneous intervals in $\mathbb{R}$ can be constructed individually for each entry of $\boldsymbol{\theta}$, or a set C in $\mathbb{R}^p$ can be constructed for the $\boldsymbol{\theta}$ as a whole. The former is usually referred to as the confidence interval, and the latter is usually referred to as the confidence set. We focus on confidence interval construction in this dissertation.

Under the model in (1.1), the usual confidence interval for θ_i is a z-confidence interval

$$\{\theta_i : y_i - z_{1-\alpha} < \theta_i < y_i + z_{1-\alpha}\}$$

where $z_{1-\alpha}$ is the $1-\alpha$ quantile of the standard normal distribution. This confidence interval has an exactly $1-\alpha$ frequentist coverage with a width of $2z_{1-\alpha}$, and it is the narrowest interval among all unbiased intervals. This usual confidence interval is constructed based on the estimator $\hat{\theta}_i = y_i$ of θ_i. The fact that the shrinkage estimator dominates the usual maximum likelihood estimator inspired the technique of constructing confidence procedures based on shrinkage estimators, hoping to develop intervals with narrower width. This leads to the development of the empirical Bayes confidence intervals. The classic empirical Bayes confidence interval utilizes the Bayes estimator directly in constructing the confidence interval, and then plugs in an estimate of the unknown parameter. By (1.3), we can construct a confidence interval for θ_i based on the Bayes estimator when τ^2 is known

$$\left\{\theta_i : \frac{\tau^2}{1+\tau^2} y_i - z_{1-\alpha}\sqrt{\frac{\tau^2}{1+\tau^2}} < \theta_i < \frac{\tau^2}{1+\tau^2} y_i + z_{1-\alpha}\sqrt{\frac{\tau^2}{1+\tau^2}}\right\}. \tag{1.6}$$

Then we can replace the unknown term that involves τ^2 with its estimate and the resulting interval will be a classic empirical Bayes confidence interval. The confidence interval in (1.6) achieves a $1-\alpha$ Bayes coverage, which is defined as

$$\Pr(\theta \in C) = \int \Pr(\theta \in C|\theta)\pi(\theta)d\theta = 1-\alpha$$

for a confidence interval C of θ with $\pi(\theta)$ being the prior distribution of θ (Morris, 1983b). Note that compared to the frequentist coverage in (1.5), θ is integrated out in this Bayes coverage probability. Hence exact frequentist coverage guarantees exact Bayes coverage, but

not the other way around. This means that although a confidence interval procedure C has a Bayes coverage $1-\alpha$, the coverage is possibly not $1-\alpha$ for a specific parameter θ. To see this, we perform a simple simulation study. We simulate a $\boldsymbol{\theta}$ vector with 100 entries from $N(0, I)$. Then we simulate a data vector $\mathbf{y}$ from $N(\boldsymbol{\theta}, I)$, and construct the classic empirical Bayes confidence intervals for entries of $\boldsymbol{\theta}$ simultaneously at the level 0.95. We repeat this procedure 1,000 times to get the empirical coverage of this confidence procedure, and the results are summarized in Figure 1.1.

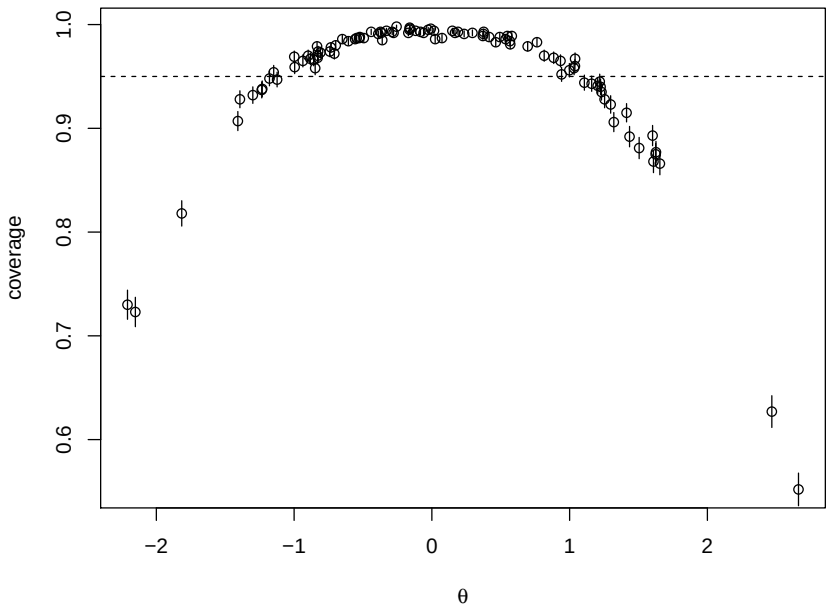

Figure 1.1: Coverage of CI for each θ_i value. The points shows the empirical estimate of the coverage probability and the solid bars are the related monte carlo error bars. The horizontal line is at 0.95.

As we can see, the classical empirical Bayes confidence interval procedure over-covers the parameter θ_i when θ_i is relatively close to 0, and under-covers the parameter θ_i when θ_i is relatively far away from 0. As a result, the coverage averaged across all the θ_is are around 0.95, but the coverage for each of the θ_is varies. One advantage of constructing

confidence intervals instead of a confidence set in such analyses is that confidence intervals do inferences on each of the θ_is individually simultaneously. Hence the fact that the classic empirical Bayes confidence interval can not guarantee the coverage for each of the θ_is is undesirable. A lot of work has been done in modifying the classic procedure trying to get better coverage property. Notable work in empirical Bayes confidence interval includes Laird and Louis (1987), He (1992) and Hwang et al. (2009), to name a few. Among them, Laird and Louis (1987) proposed a bootstrap interval procedure based on the construction of Morris (1983b). He (1992) proposed a Bayes confidence interval procedure obtained by choosing the Bayes rule against the loss

$$L(\theta, C) = k|C| - I(\theta \in C)$$

under a prior $\pi(\theta)$, where k is a constant that needs to be properly chosen to achieve the Bayes coverage. Then the empirical Bayes counterpart is obtained by approximating the unknown parameters with their estimators. Hwang et al. (2009) proposed a confidence interval procedure that not only shrinks the means but also shrink the variances. However, these procedures can not guarantee the exact frequentist coverage. In Chapter 2, we will keep pursuing this goal of deriving a confidence interval procedure with exact frequentist coverage while having narrower width than the usual confidence procedure.

1.3 Simultaneous Tests

When analyzing multigroup data, other than constructing confidence intervals simultaneously, another commonly used statistical inference procedure is simultaneous hypothesis testing. Especially in modern statistical applications, testing procedure for high throughput experiments is desired, which may involve thousands or even millions of hypotheses. For example, in genome science, scientists need to identify which genes' expression levels associate significantly with the phenotype of interest. This is usually done by simultaneously testing the hypothesis that there is no association between the phenotype level and gene's expression level for each gene.

Suppose there are p hypotheses, and for each hypothesis, there is a parameter of interest θ_i, where $i = 1, ..., p$. We are interested in making inferences on the θ_is simultaneously. For one single test, usually researchers specify a type I error rate threshold α first, which will be the probability of rejecting the null hypothesis when it is actually true. However, when p hypotheses are tested simultaneously, if we still use the same threshold α again for each of the hypothesis, the probability of making a type I error will no longer be α. In fact, in multiple testing situation, the probability of rejecting at least one true null hypothesis is called the familywise error rate (FWER). The classic procedure for controlling the FWER is the Bonferroni correction, which uses α/p instead of α as the level of the test for each hypothesis in order to control the FWER below α.

However, FWER is sometimes considered as a conservative quantity to control in multiple testing situation. In some applications, controlling the FWER leads to very few or even no discoveries. As an alternative, False Discovery Rate (FDR) has been popular since the innovative paper by Benjamini and Hochberg (1995). Suppose among the p hypotheses we make R rejections, and V of those are true nulls. We also call a rejection that is actually a true null a false discovery. The the FDR is defined as

$$FDR = E[\tfrac{V}{R \vee 1} | \theta_1, ..., \theta_p].$$

Hence unlike the FWER that only allows for at most one false discovery, the FDR may allow for certain number of false discoveries. As a result, controlling the FDR can be a more powerful way to identify more discoveries while still controlling the proportion of the false discoveries. The Benjamini-Hochberg (BH) procedure is one classic approach to control the FDR. Let $q_{(1)}, ..., q_{(p)}$ be the ordered p-values from small to large for the p hypotheses. For a desired FDR level α, the BH procedure finds the i_m such that it is the largest i for which

$$q_i \leq \frac{i}{p}\alpha, \tag{1.7}$$

and reject the hypotheses with p-values smaller than q_{i_m}. Other than the BH procedure, a large amount of work has been done on controlling the FDR, see Efron (2012), Benjamini (2010), Genovese and Wasserman (2004), Storey (2002), Storey (2007).

FWER or FDR control procedures control the number of discoveries that are actually nulls, i.e. $\theta = 0$. However, usually in real world, the parameter of interest θ can never be exactly 0. For example, Tukey (1991) argued that asking whether the difference between effects of A and B is zero is "foolish", since the effects of A and B are always different in some decimal place. In situation like this, the control of FDR might not be appropriate since there is actually no "true nulls". As argued by Tukey (1962), the more meaningful question would be to judge whether there is enough evidence to support a correct sign instead of a "discovery".

However, in the usual context of testing a null hypothesis $H_0 : \theta = 0$ for a parameter of interest θ, when we reject H_0, the sign of θ is usually inferred at the same time. For example, suppose $\hat{\theta} > 0$ is an estimate of θ, and the test based on $\hat{\theta}$ indicates significance, we would conclude that the true θ is positive. The usually used 0.05 level is only a threshold for type I error. The probability of claiming a wrong sign of θ is not quantified in the standard hypothesis testing procedure, and researchers are usually not aware of the probability of it.

We now use a small simulated example to illustrate this issue. Suppose the observation y is sampled from $N(\theta, 4)$. A standard z-test will reject H_0 if $|y| > 2z_{1-\alpha/2}$. Suppose we claim θ to be the same sign as y if H_0 is rejected. We calculate the ratio of the number of times of claiming the true signs to the number of rejections. We repeat this procedure 1,000 times, and do it for different θ values. Figure 1 shows the results. When the true effect size is relatively large, the probability of claiming a wrong sign is small. However, when the true effect size is small, the chance of claiming a wrong sign is very large. In this example, when the true effect is 0.01, if we estimate the sign when the observation indicates significance at level 0.05, we are getting the wrong sign around 50% of the time. This is similar to randomly guessing the sign of the true effect. In applications, so-called "significant positive associations" could actually be negative associations with a high probability, and vice-versa. The potential high probability of drawing conclusion that is opposite of the truth, even though the type I error is controlled, could lead to serious consequence in applications (Gelman and Carlin, 2014).

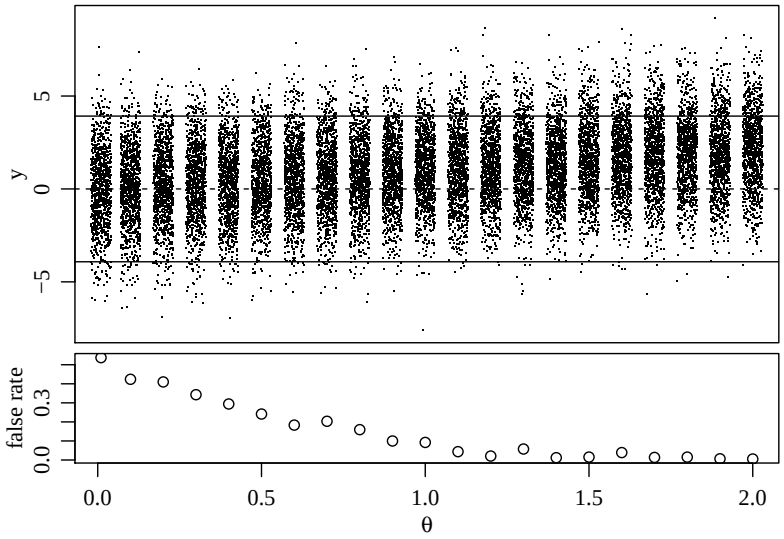

Figure 1.2: The top plot gives all the sampled y values against their corresponding θ values. The y values outside the two solid lines are rejected. The bottom plot gives empirical estimate of the probability of claiming a wrong sign for each θ value, based on 1,000 replications. The θ values we choose here are 0.01, 0.1, 0.2, ..., 2.

We call the error of claiming a wrong sign of θ a sign error. However, the probability of a sign error usually depends on the unknown value of the parameter θ, making it difficult to estimate. Gelman and Tuerlinckx (2000) defined type S error rate as the probability of making a sign error conditional on rejection. They suggested that using hierarchical Bayesian credible interval would lead to less sign errors. Gelman and Carlin (2014) suggested that researchers should use external information to come up with reasonable estimates of plausible effect sizes, and use these to assess the potential type S error rate. Owen (2016) suggested that the type S error rate can be controlled using a two-stage inference procedure, where the claim for significance and the claim for sign should be separated. For multi-group data setting, Benjamini and Yekutieli (2005) defined the mixed directional FDR as the expected proportion of discoveries in which a nonpositive parameter is declared positive or a nonnegative parameter is declared negative. They proposed to implement the Benjamini-

Hochberg algorithm directly to control the mixed directional FDR, and showed it will be controlled under a desired threshold. Following this framework, Zhao et al. (2015) used weighted p-value methods to control the mixed directional FDR, and Guo et al. (2010) extended the idea to the problem of multidimensional directional decisions. Stephens (2016) also studied the control of sign error, where the focus there was the control of local sign error instead of the sign error rate across groups. In Chapter 3, we will keep studying the sign error issue in hypothesis testing and propose new procedures that control the sign error rate.

1.4 Graphical Models and Phylogenetic Trees

A phylogenetic tree is a tree-structure graph that describes the evolutionary relationships among various organisms which are considered to have one common ancestor. Over time, the ancestral lineage splits due to stochastic genetic changes, creating new organisms. Figure 1.3 shows an example of a well-known phylogenetic tree of commonly seen organisms (University of California Museum of Paleontology, 2018). From top to bottom, it describes the evolutionary process over time. The root node represents the common ancestor of all considered animals. The internal nodes represent the divergence events, and the leaf nodes are the contemporary organisms. Species that share more common ancestors together in the tree are considered more genetically related. The application of phylogenetic trees is not limited to describing the evolution of organisms. There are broader applications in areas like public health (Ou et al., 1992), forensic science (Scaduto et al., 2010) and medicine (Amenta et al., 2015).

A graphical model is a statistical model associated with a graph that captures the dependencies among random variables (Lauritzen, 1996). Specifically, each random variable is associated with a vertex in the graph, and the edges in the graph encode conditional independence constraints. For a phylogenetic tree, the random variables that are represented by the leaf nodes can either be discrete or continuous, depending on the application. For example, DNA or protein sequence data are often recorded as discrete data while gene expression data or gene frequency data are usually continuous. Here we focus on the setting of continuous

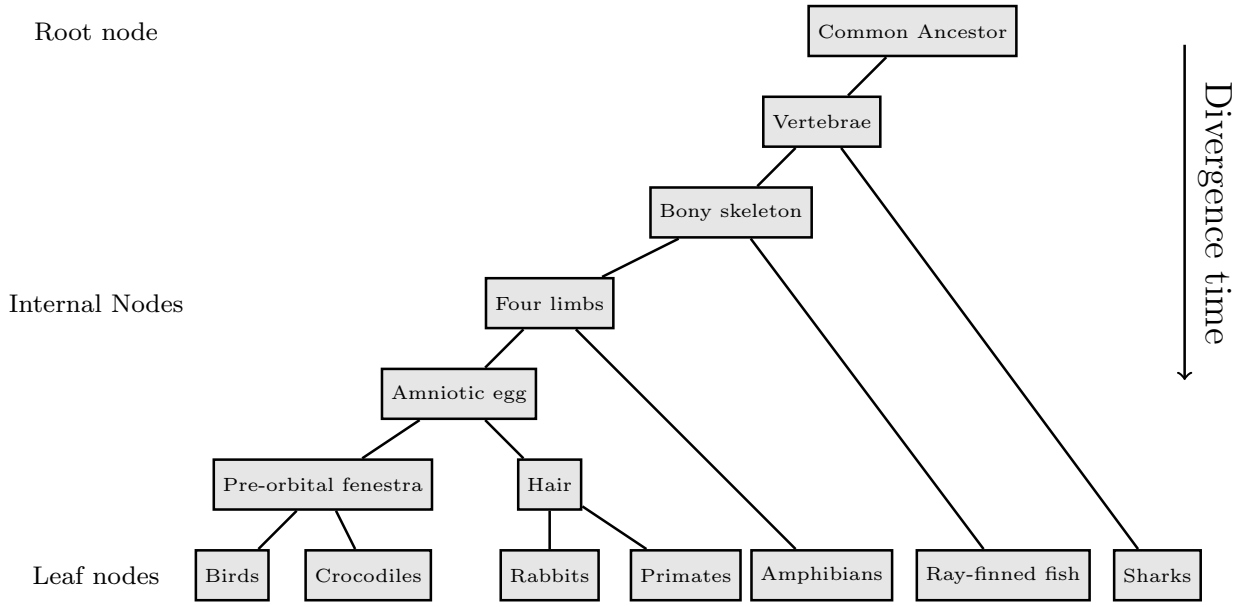

Figure 1.3: An example of a phylogenetic tree.

observations for which we assume a continuous stochastic process over a tree. Felsenstein (1973) proposed a Brownian motion tree model as phylogenetic model for the analysis of the evolution of continuous characters. This Brownian motion model has since been widely used in inferring phylogenetic trees, and our tree inferences procedure will be based on this model.

In the Brownian motion model, the character evolution is assumed to follow a Brownian motion process with mean 0 and variance σ^2 per unit time along the phylogenetic tree. Under this model, the net change along a branch after t unit of time is drawn from a normal distribution with mean 0 and variance $\sigma^2 t$. In the event of splitting, two direct descendants start to evolve by independent continuation of the Brownian motion process. The most commonly used tree type in phylogenetics is the bifurcating tree, which is a tree structure such that every internal node has exactly two descendants. The tree in Figure 1.3 is an example of a bifurcating tree. For a bifurcating tree with p leaf nodes, there are $p-1$ internal nodes and $2p-1$ branches. In applications, we usually only have observations from

the p leaf nodes. With the observed data from the leaf nodes, the two main inference tasks are inferring the topology of the tree and estimating the branch lengths.

Simple clustering method like the Neighbor-Joining approach (Saitou and Nei, 1987) are popular methods in reconstruction of the phylogenetic tree. We will propose two alternative clustering methods for the phylogenetic tree reconstruction. We then seek to improve these simple methods through use of likelihood. However Roch (2006) has shown that the maximum likelihood estimation of the tree topology under Brownian motion models is "NP hard". When p is very small, we can enumerate all possible tree topologies and evaluate them with the observed data. However, the number of tree topology grows in a factorial speed as p grows. For example, when $p = 20$, there are around 5×10^{29} phylogenetic tree topologies. Hence any exhaustive method would not work for applications with large number of leaf nodes. In Chapter 4, we propose a Structural EM algorithm for the reconstruction of the maximum likelihood phylogenetic tree with continuous data.

1.5 Summary of Our Contributions

In Chapter 2, we construct confidence intervals that have a constant frequentist coverage rate and that make use of information about across-group heterogeneity, resulting in constant-coverage intervals that are narrower than standard t-intervals across groups. Such intervals are constructed by inverting biased tests for the mean of a normal population. Given a prior distribution on the mean, Bayes-optimal biased tests can be inverted to form Bayes-optimal confidence intervals with frequentist coverage that is constant as a function of the mean. In the context of multiple groups, the prior distribution is replaced by a model of across-group heterogeneity. The parameters for this model can be estimated using data from all of the groups, and used to obtain confidence intervals with constant group-specific coverage that adapt to information about the distribution of group means.

In Chapter 3, we propose two procedures for adaptively selecting an experimentwise significance threshold in order to control the sign error rate. The first controls the sign error rate conservatively, without any distributional assumption on the parameters of interest. The

second is an empirical Bayes procedure, and achieves optimal performance asymptotically when a model for the distribution of the parameters is correctly specified. We also discuss an adaptive procedure to minimize the sign error rate when the experimentwise type I error rate is held fixed.

In Chapter 4, we consider the problem of learning the tree topology in phylogenetic analysis of continuous data. The models we treat assume that the evaluation of the continuous characters along the phylogenetic tree follows a Brownian motion process. We propose and compare three methods for recovery of the tree structure. The first is based on independence tests, and the second is based on the size of sample covariances. As a third option, we derive a Structural EM algorithm for searching the tree structure and corresponding parameter estimates that maximize the likelihood of the observed data. We examine the performance of our methods using simulated data sets and gene expression data.

Chapter 2

ADAPTIVE MULTIGROUP CONFIDENCE INTERVALS WITH CONSTANT COVERAGE

This work is published in *Biometrika* (Yu and Hoff, 2018).

2.1 Introduction

A commonly used experimental design is the one-way layout, in which a random sample $Y_{1,j}, \ldots, Y_{n_j,j}$ is obtained from each of several related groups $j \in \{1, \ldots, p\}$. The standard normal-theory model for data from such a design is that $Y_{1,j}, \ldots, Y_{n_j,j} \sim$ i.i.d. $N(\theta_j, \sigma^2)$, independently across groups. Inference for the θ_j's typically proceeds in one of two ways. The "classical" approach is to use the unbiased sample mean $\bar{y}_j$ as an estimator of θ_j, and to construct a confidence interval for θ_j by inverting the appropriate uniformly most powerful unbiased (UMPU) test, that is, constructing the standard t-interval. Such an approach essentially makes inference for each θ_j using only data from group j (although a pooled-sample estimate of σ^2 is often used). The estimator of each θ_j is unbiased, and the confidence interval for each θ_j has the desired coverage rate.

An alternative approach is to utilize data from all of the groups to infer each individual θ_j. This is typically done by invoking a hierarchical model, that is, a statistical model that describes the heterogeneity across groups. The standard one-way random effects model posits that $\theta_1, \ldots, \theta_p$ are a random sample from a normal population, so that $\theta_1, \ldots, \theta_p \sim$ i.i.d. $N(\mu, \tau^2)$. In this case, shrinkage estimators of the form

$$\hat{\theta}_j = \frac{\hat{\mu}/\hat{\tau}^2 + \bar{y}_j n_j/\hat{\sigma}^2}{1/\hat{\tau}^2 + n_j/\hat{\sigma}^2}$$

are often used, where $(\hat{\mu}, \hat{\tau}^2, \hat{\sigma}^2)$ are estimated using data from all of the groups. This estimator has a lower variance than the sample mean, but is generally biased. Confidence

intervals based on these shrinkage estimators are often derived from the hierarchical model: Letting $\tilde{\theta}_j$ be defined as

$$\tilde{\theta}_j = \frac{\mu/\tau^2 + \bar{y}_j n_j/\sigma^2}{1/\tau^2 + n_j/\sigma^2},$$

then $\mathrm{E}[(\tilde{\theta}_j - \theta_j)^2] = (1/\tau^2 + n_j/\sigma^2)^{-1}$, where the expectation integrates over both the normal model for the observed data and the normal model representing heterogeneity across the groups. This quantity is also the conditional variance of θ_j given data from group j, which suggests an empirical Bayes posterior interval for θ_j of the form $\hat{\theta}_j \pm t_{1-\alpha/2}/\sqrt{1/\hat{\tau}^2 + n_j/\hat{\sigma}^2}$, where t_γ denotes the γ-quantile of the appropriate t-distribution. Compared to the classical t-interval $\bar{y}_j \pm t_{1-\alpha/2}\sqrt{\hat{\sigma}^2/n_j}$, this interval is narrower by a factor of $\sqrt{\hat{\tau}^2/(\hat{\tau}^2 + \hat{\sigma}^2/n)}$. However, its coverage rate is not $1-\alpha$ for all groups. While the rate tends to be near the nominal level on average across all groups, the rate for a specific group j will depend on the value of θ_j. Specifically, the coverage rate will be too low for θ_j's far from the overall average θ-value, and too high for θ_j's that are close to this average (see, for example, Snijders and Bosker (2012, Section 4.8)). Other types of empirical Bayes posterior intervals have been developed by Morris (1983a), Laird and Louis (1987), He (1992) and Hwang et al. (2009). Like the interval obtained from the hierarchical normal model, these intervals are narrower than the standard t-interval but fail to have the target coverage rate for each group.

In the related problem of confidence region construction for a vector of normal means, several authors have pursued procedures that dominate those based on UMPU test inversion (Berger, 1980; Casella and Hwang, 1986). In particular, Tseng and Brown (1997) obtain a modified empirical Bayes confidence region that has exact frequentist coverage but is also uniformly smaller than the usual procedure. In this chapter we pursue similar results for the problem of multigroup confidence interval construction. Specifically, we develop a confidence interval procedure that has the desired coverage rate for every group, but also adapts to the heterogeneity across groups, thereby achieving shorter confidence intervals than the classical approach on average across groups. More precisely, our goal is to obtain a multigroup confidence interval procedure $\{C^1(\boldsymbol{Y}), \ldots, C^p(\boldsymbol{Y})\}$, based on data $\boldsymbol{Y}$ from all of

the groups, that attains the target frequentist coverage rate for each group and all values of $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_p)$, so that

$$\Pr(\theta_j \in C^j(\boldsymbol{Y})|\boldsymbol{\theta}) = 1 - \alpha \quad \forall\, \boldsymbol{\theta} \in \mathbb{R}^p,\ \forall\, j \in \{1, \ldots, p\}, \tag{2.1}$$

and is also more efficient than the standard t-interval on average across groups, so that

$$\mathrm{E}[|C^j(\boldsymbol{Y})|] < 2t_{1-\alpha/2}, \tag{2.2}$$

where $|C|$ denotes the width of an interval C, and the expectation is with respect to an unknown distribution describing the across-group heterogeneity of the θ_j's. The interval procedures we propose satisfy the constant coverage property (2.1) exactly. Property (2.2) will hold approximately, depending on what the across-group distribution is and how well it is estimated.

The intuition behind our procedure is as follows: While the standard t-interval for a single group is uniformly most accurate among unbiased interval procedures (UMAU), it is not uniformly most accurate among all procedures. We define classes of biased hypothesis tests for a normal mean, inversion of which generates $1 - \alpha$ frequentist t-intervals that are more accurate than the standard UMAU t-interval for some values of the parameter space, but less accurate elsewhere. The class of tests can be chosen to minimize an expected width with respect to a prior distribution for the population mean, yielding the confidence interval procedure (CIP) that is Bayes-optimal among all CIPs that have $1 - \alpha$ frequentist coverage. We call the Bayes-optimal frequentist procedure a "frequentist assisted by Bayes" (FAB) interval procedure. In a multigroup setting, the "prior" for the population mean is replaced by a model for across-group heterogeneity. The parameters in this model can be estimated using data from all of the groups, yielding an empirical FAB confidence interval procedure that maintains a coverage rate that is constant as a function of the group means.

Several authors have studied constant coverage CIPs in the single-group case that differ from the UMAU procedure. Such procedures generally make use of some sort of prior knowledge about the population mean. In particular, our work builds upon that of Pratt

(1963), who studied the Bayes-optimal z-interval for the case that σ^2 is known. Other related work includes Farchione and Kabaila (2008) and Kabaila and Tissera (2014), who developed procedures that make use of non-probabilistic prior knowledge that the mean is near a pre-specified parameter value (e.g. zero). Their procedures have shorter expected widths near this special value, but revert to the UMAU procedures when the data are far from this point. Evans et al. (2005) obtained minimax CIPs for cases where prior knowledge takes the form of bounds on the parameter values.

The FAB t-interval we construct is a straightforward extension of the Bayes-optimal z-interval developed by Pratt (1963). In the next section, we review the FAB z-interval of Pratt and extend the idea to construct a FAB t-interval for the case that σ^2 is unknown. In Section 2.3 we use the FAB t-interval procedure to obtain group-specific confidence intervals that have constant coverage rates for all groups and all values of $\boldsymbol{\theta}$, and are also asymptotically optimal as the number of groups increases. In Section 2.4 we illustrate the use of the FAB interval procedure with an example dataset, and compare its performance to that of the UMAU and empirical Bayes procedures often used for multigroup data. Then we examine the robustness of the proposed confidence interval procedure. A discussion follows in Section 2.6 . Proofs are given in Appendix.

2.2 FAB confidence intervals

Consider a model for a random variable Y that is indexed by a single unknown scalar parameter $\theta \in \mathbb{R}$. A $1-\alpha$ confidence region procedure (CRP) for θ based on Y is a set-valued function $C(y)$ such that $\Pr(\theta \in C(Y)|\theta) = 1-\alpha$ for all $\theta \in \mathbb{R}$. As is well-known, a CRP can be constructed by inversion of a collection of hypothesis tests. For each $\theta \in \mathbb{R}$, let $A(\theta)$ be the acceptance region of an α-level test of $H_\theta : Y \sim P_\theta$ versus $K_\theta : Y \sim P_{\theta'}, \theta' \neq \theta$. Then $C(y) = \{\theta : y \in A(\theta)\}$ is a $1-\alpha$ CRP. We take the risk $R(\theta, C)$ of a $1-\alpha$ CRP to be its expected Lebesgue measure

$$R(\theta, C) = \int \int 1(y \in A(\theta'))\ d\theta'\ P_\theta(dy).$$

For our model of current interest, $Y \sim N(\theta, \sigma^2)$ with σ^2 known, there does not exist a CRP that uniformly minimizes this risk over all values of θ. However, there exist optimal CRPs within certain subclasses of procedures. For example, the standard z-interval, given by $C_z(y) = (y + \sigma z_{\alpha/2}, y + \sigma z_{1-\alpha/2})$, minimizes the risk among all unbiased CRPs derived by inversion of unbiased tests of H_θ versus K_θ, and so is the uniformly most accurate unbiased (UMAU) CRP.

That the interval is unbiased means $\Pr(\theta' \in C_z(Y)|\theta) \leq 1 - \alpha$ for all θ and θ', and that it is UMAU means $R(\theta, C_z) = 2\sigma z_{1-\alpha/2} \leq R(\theta, \tilde{C})$ for any other unbiased CRP $\tilde{C}$ and every θ. But while C_z is best among unbiased CRPs, the lack of a UMP test of H_θ versus K_θ means there will be CRPs corresponding to collections of biased level-α tests that have lower risks than C_z for *some* values of θ. This suggests that if we have prior information that θ is likely to be near some value μ, we may be willing to incur larger risks for θ-values far from μ in exchange for small risks near μ. With this in mind, we consider the Bayes risk $R(\pi, C) = \int R(\theta, C)\, \pi(d\theta)$, where π is a prior distribution that describes how close θ is likely to be to μ. This Bayes risk may be related to the marginal (Bayes) probability of accepting H_θ as follows:

$$\begin{aligned} R(\pi, C) = \int R(\theta, C)\, \pi(\theta) d\theta &= \int\int\int 1(y \in A(\theta'))\, d\theta'\, P_\theta(dy)\, \pi(d\theta) \\ &= \int\int\int 1(y \in A(\theta')) P_\theta(dy) \pi(d\theta)\, d\theta' \\ &= \int \Pr(Y \in A(\theta'))\, d\theta'. \end{aligned}$$

The Bayes-optimal $1 - \alpha$ CRP is obtained by choosing $A(\theta)$ to minimize $\Pr(y \in A(\theta))$ for each $\theta \in \mathbb{R}$, or equivalently, to maximize the probability that H_θ is rejected under the prior predictive (marginal) distribution P_π for Y that is induced by π. This means that the optimal $A(\theta)$ is the acceptance region of the most powerful test of the simple hypothesis $H_\theta : Y \sim P_\theta$ versus the simple hypothesis $K_\pi : Y \sim P_\pi$. The confidence region obtained by inversion of this collection of acceptance regions is Bayes optimal among all CRPs having $1-\alpha$ frequentist coverage. We describe such a procedure as "frequentist, assisted by Bayes",

or FAB.

Using this logic, Pratt (1963) obtained and studied the Bayes-optimal optimal CRP for the model $Y \sim N(\theta, \sigma^2)$ with σ^2 known and prior distribution $\theta \sim N(\mu, \tau^2)$. Under this distribution for θ, the marginal distribution for Y is $N(\mu, \tau^2 + \sigma^2)$. The Bayes-optimal CRP is therefore given by inverting acceptance regions $A(\theta)$ of the most powerful tests of $H_\theta : Y \sim N(\theta, \sigma^2)$ versus $K_\pi : Y \sim N(\mu, \tau^2 + \sigma^2)$ for each θ. This optimal CRP is an interval, the endpoints of which may be obtained by solving two nonlinear equations. We refer to this CRP as Pratt's FAB z-interval.

The procedure used to obtain the FAB z-interval, and the form used by Pratt, are not immediately extendable to the more realistic situation in which $Y_1 \ldots, Y_n \sim$ i.i.d $N(\theta, \sigma^2)$ where both θ and σ^2 are unknown. The primary reason is that in this case the Bayes-optimal acceptance region depends on the unknown value of σ^2, or to put it another way, the null hypothesis H_θ is composite. However, the situation is not too difficult to remedy: Below we re-express Pratt's z-interval in terms of a function that controls where the type I error is "spent". We then define a class of t-intervals based on such functions, from which we obtain the Bayes-optimal t-interval for the case that σ^2 is unknown.

2.2.1 The Bayes-optimal w-function

For the model $\{Y \sim N(\theta, \sigma^2),\ \theta \in \mathbb{R}\}$ we may limit consideration of CRPs to those obtained by inverting collections of two-sided tests:

Lemma 2.2.1. *Suppose the distribution of Y belongs to a one-parameter exponential family with parameter $\theta \in \mathbb{R}$. For any confidence region procedure $\tilde{C}$ there exists a procedure C, obtained by inverting a collection of two-sided tests, that has the same coverage as $\tilde{C}$ and a risk less than or equal to that of $\tilde{C}$.*

For the normal model of interest, an interval $A(\theta) = (\theta - \sigma u, \theta - \sigma l)$ will be the acceptance region of a two-sided level-α test if and only if u and l satisfy $\Phi(u) - \Phi(l) = 1 - \alpha$, or equivalently, if $u = z_{1-\alpha w}$ and $l = z_{\alpha(1-w)}$ for some value of $w \in (0, 1)$, where Φ is the

standard normal CDF and $z_\gamma = \Phi^{-1}(\gamma)$. It is important to note that the value of w, and thus l and u, can vary with θ and still yield a $1-\alpha$ confidence region: Let $w : \mathbb{R} \to (0,1)$ and define

$$A_w(\theta) = (\theta - \sigma z_{1-\alpha w(\theta)}, \theta - \sigma z_{\alpha(1-w(\theta))}). \tag{2.3}$$

Then for each θ, $A_w(\theta)$ is the acceptance region of a level-α test of H_θ versus K_θ. Inversion of $A_w(\theta)$ yields a $1-\alpha$ CRP given by

$$C_w(y) = \{\theta : y + \sigma z_{\alpha(1-w(\theta))} < \theta < y + \sigma z_{1-\alpha w(\theta)}\}. \tag{2.4}$$

This confidence region can be seen as a generalization of the usual UMAU z-interval, given by $C_{1/2}(y) = \{\theta : y + \sigma z_{\alpha/2} < \theta < y + \sigma z_{1-\alpha/2}\}$, corresponding to a constant w-function of $w(\theta) = 1/2$. Given a prior distribution for θ, the Bayes-optimal w-function corresponds to the Bayes-optimal CRP. For the prior distribution $\theta \sim N(\mu, \tau^2)$ considered by Pratt, the optimal w-function depends on $\psi = (\mu, \tau^2, \sigma^2)$ and is given as follows:

Proposition 2.2.1. *Let $Y \sim N(\theta, \sigma^2)$, $\theta \sim N(\mu, \tau^2)$ and let $w : \mathbb{R} \to (0,1)$. Then $R(\psi, C_{w_\psi}) \leq R(\psi, C_w)$ where $w_\psi(\theta)$ is given by $w_\psi(\theta) = g^{-1}(2\sigma(\theta-\mu)/\tau^2)$ with $g(w) = \Phi^{-1}(\alpha w) - \Phi^{-1}(\alpha(1-w))$. The function $w_\psi(\theta)$ is a continuous strictly increasing function of θ.*

As stated in Pratt (1963) but not proven, $C_{w_\psi}(y)$ is actually an interval for each $y \in \mathbb{R}$, and so C_{w_ψ} is a confidence interval procedure (CIP). In fact, a CRP C_w will be a CIP as long as the w-function is continuous and nondecreasing:

Lemma 2.2.2. *Let $w : \mathbb{R} \to (0,1)$ be a continuous nondecreasing function. Then the set $C_w(y) = \{\theta : y + \sigma z_{\alpha(1-w(\theta))} < \theta < y + \sigma z_{1-\alpha w(\theta)}\}$ is an interval and can be written as (θ^L, θ^U), where θ^L and θ^U are solutions to $\theta^L = y + \sigma z_{\alpha(1-w(\theta^L))}$ and $\theta^U = y + \sigma z_{1-\alpha w(\theta^U)}$.*

A bit of algebra shows that Pratt's FAB z-interval can be expressed as $C_{w_\psi} = (\theta^L, \theta^U)$,

where θ^L and θ^U solve

$$\theta^U = \frac{y + \sigma\Phi^{-1}(1 - \alpha + \Phi(\frac{y-\theta^U}{\sigma}))}{1 + 2\sigma^2/\tau^2} + \mu\frac{2\sigma^2/\tau^2}{1 + 2\sigma^2/\tau^2}$$
$$\theta^L = \frac{y + \sigma\Phi^{-1}(\alpha - \Phi(\frac{\theta^L-y}{\sigma}))}{1 + 2\sigma^2/\tau^2} + \mu\frac{2\sigma^2/\tau^2}{1 + 2\sigma^2/\tau^2}.$$

Solutions to these equations can be found with a zero-finding algorithm, and noting the fact that $\theta^L < y + \sigma z_\alpha$ and $y + \sigma z_{1-\alpha} < \theta^U$.

Some aspects of the FAB z-interval procedure are displayed graphically in Figure 2.1. The left panel gives the w-functions corresponding to the Bayes-optimal 95% CIPs for $\sigma^2 = 1$, $\mu = 0$ and $\tau^2 \in \{1/4, 1, 4\}$. At varying rates depending on τ^2, the w-functions approach zero or one as θ moves towards $-\infty$ and ∞, respectively. The level-α tests corresponding to these w-functions are "spending" more of their type I error on y-values that are likely under the $N(\mu, \sigma^2 + \tau^2)$ prior predictive distribution of Y. This makes the intervals narrower than the usual interval when y is near μ, and wider when y is far from μ, as shown in the middle panel of the figure. In particular, at $y = \mu$, the 95% FAB z-interval with $\tau^2 = 1/4$ has a width of 3.29, which is about 84% of that of the UMAU interval. Average performance across y-values is given by risk, or expected confidence interval width, displayed in the top right plot. Expected widths of the FAB z-intervals are lower than those of the UMAU intervals for values of θ near μ (15% lower for $\theta = \mu$ and $\tau^2 = 1/4$), but can be much higher for θ-values far away from μ, particularly for small values of τ^2. Relative to small values of τ^2, the larger value of $\tau^2 = 4$ enjoys better performance than the UMAU interval over a wider range of θ-values, but the improvement is not as large near μ. Additional calculations (available from the replication code for this chapter) show that the performance of the FAB interval near μ improves as α increases, as compared to the UMAU interval. For example, with $\tau^2 = 1/4$ and $\alpha = 0.50$, the width of the FAB interval at $y = \mu$ is about 25% of that of the UMAU interval, and its risk at $\theta = \mu$ is 60% that of the UMAU interval.

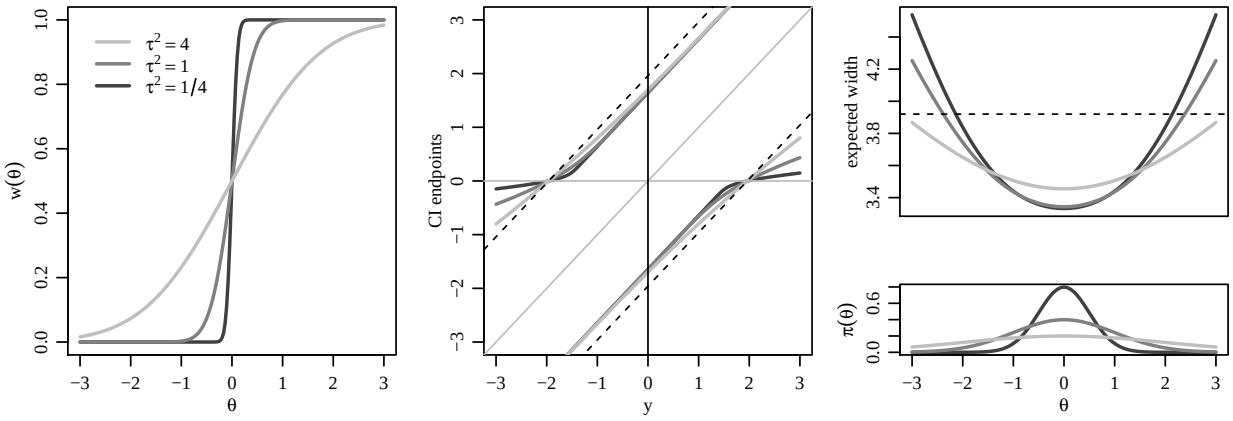

Figure 2.1: Descriptions of the FAB z-procedure. The left plot gives Bayes-optimal w-functions for three values of τ^2, at level $\alpha = 0.05$. The middle plot gives the corresponding confidence interval procedures, with the UMAU procedure given by dashed lines. The top plot on the right gives the risk functions (expected widths) of the 95% FAB z-intervals for the three values of τ^2, with the corresponding prior densities plotted below.

2.2.2 FAB t-intervals

Adoption of Pratt's z-interval has been limited, possibly due to two factors: First, in most applications the population variance is unknown, and second, the prior distribution for θ must be specified. We now address this first issue by developing a FAB t-interval. Suppose we have a sample $Y_1, \ldots, Y_n \sim$ i.i.d. $N(\theta, \sigma^2)$, with sufficient statistics $(\bar{Y}, S^2)$, the sample mean and (unbiased) sample variance. The standard UMAU t-interval is given by

$$\{\theta : \bar{y} + \tfrac{s}{\sqrt{n}} t_{\alpha/2} < \theta < \bar{y} + \tfrac{s}{\sqrt{n}} t_{1-\alpha/2}\}. \tag{2.5}$$

This interval is symmetric around $\bar{y}$, with the same tail-area probability $(\alpha/2)$ defining the lower and upper endpoints. The development of the w-function described in the previous subsection suggests viewing the UMAU t-interval as belonging to the larger class of CRPs, given by

$$C_w(\bar{y}, s^2) = \{\theta : \bar{y} + \tfrac{s}{\sqrt{n}} t_{\alpha(1-w(\theta))} < \theta < \bar{y} + \tfrac{s}{\sqrt{n}} t_{1-\alpha w(\theta)}\}, \tag{2.6}$$

for some $w : \mathbb{R} \to (0,1)$. Any procedure thus defined satisfies $\Pr(\theta \in C_w(\bar{Y}, S^2)|\theta) = 1 - \alpha$ for any value of θ, and the standard t-interval in (2.5) is a special case when $w = 1/2$. Additionally, C_w is a CIP as long as w is a continuous nondecreasing function:

Lemma 2.2.3. *Let $w : \mathbb{R} \to (0,1)$ be a continuous nondecreasing function. Then the set $C_w(\bar{y}, s^2) = \{\theta : \bar{y} + \frac{s}{\sqrt{n}} t_{\alpha(1-w(\theta))} < \theta < \bar{y} + \frac{s}{\sqrt{n}} t_{1-\alpha w(\theta)}\}$ is an interval and can be written as (θ^L, θ^U), where θ^L and θ^U are solutions to $\theta^L = \bar{y} + \frac{s}{\sqrt{n}} t_{\alpha(1-w(\theta^L))}$ and $\theta^U = \bar{y} + \frac{s}{\sqrt{n}} t_{1-\alpha w(\theta^U)}$.*

For a given w-function, the endpoints of the interval can be reëxpressed as

$$F\left(\frac{\bar{y} - \theta^U}{s/\sqrt{n}}\right) = \alpha w(\theta^U) \tag{2.7}$$

$$F\left(\frac{\bar{y} - \theta^L}{s/\sqrt{n}}\right) = 1 - \alpha(1 - w(\theta^L)), \tag{2.8}$$

where F is the CDF of the t_{n-1} distribution. Using the same logic as at the beginning of Section 2.2, the Bayes risk of a CRP for a prior distribution π on θ and σ^2 is

$$R(\pi, C) = \int \Pr((\bar{Y}, S^2) \in A(\theta'))\, d\theta', \tag{2.9}$$

where $\Pr((\bar{Y}, S^2) \in A(\theta'))$ is the prior predictive (marginal) probability of $(\bar{Y}, S^2)$ being in the acceptance region $A(\theta')$ under the prior distribution π. Given a prior π that corresponds to a continuous, nondecreasing w-function, the Bayes-optimal FAB interval can be obtained numerically by using an iterative algorithm to solve (2.7) and (2.8). However, this requires computation of the w-function, which for each θ is the minimizer in w of $\Pr((\bar{Y}, S^2) \in A_w(\theta))$, where

$$A_w(\theta) = \left\{(\bar{y}, s^2) : t_{\alpha w} < \frac{\bar{y} - \theta}{s/\sqrt{n}} < t_{1-\alpha(1-w)}\right\}. \tag{2.10}$$

Obtaining the optimal w-function will generally involve numerical integration. Consider a $N(\mu, \tau^2)$ prior on θ and so conditionally on σ^2 we have $\bar{Y} \sim N(\mu, \sigma^2/n + \tau^2)$ and $(n-1)S^2/\sigma^2 \sim \chi^2_{n-1}$. From this we can show that $c(\bar{Y} - \theta)/(S/\sqrt{n})$ has a noncentral t_{n-1} distribution with noncentrality parameter $\lambda = c\frac{\mu - \theta}{\sigma/\sqrt{n}}$, where $c = \sqrt{\sigma^2/n}/\sqrt{\sigma^2/n + \tau^2}$. Therefore, the probability of the event $\{(\bar{Y}, S^2) \in A(\theta)\}$, conditional on σ^2, can be written

as

$$\Pr(\{\bar{Y}, S^2\} \in A(\theta)|\sigma^2) = F_\lambda(ct_{1-\alpha(1-w)}) - F_\lambda(ct_{\alpha w}),$$

where F_λ is the CDF of the noncentral t_{n-1} distribution with parameter $\lambda = c\frac{\mu-\theta}{\sigma/\sqrt{n}}$. The Bayes-optimal w-function is therefore given by

$$w_\pi(\theta) = \arg\min_w \int \left(F_\lambda(ct_{1-\alpha(1-w)}) - F_\lambda(ct_{\alpha w})\right) p_\pi(\sigma^2)\, d\sigma^2, \tag{2.11}$$

where $p_\pi(\sigma^2)$ is the prior density over σ^2.

In the replication material for this chapter we provide R-code for obtaining $w_\pi(\theta)$ and the corresponding Bayes-optimal t-interval $C_\pi(\bar{y}, s^2)$ for the class of priors where θ and σ^2 are *a priori* independently distributed as normal and inverse-gamma random variables. Here, we provide some descriptions of this FAB t-interval procedure for some parameter values that make the interval comparable to the z-interval from Section 2.2.1. Specifically, we consider the case that $n = 10$, $1/\sigma^2 \sim \text{gamma}(1, 10)$ and $\theta \sim N(0, \tau^2)$ for $\tau^2 \in \{1/4, 1, 4\}$. This makes the prior median of σ^2 near 10, and the variance of $\bar{Y}$ near 1 (and so the variance of $\bar{Y}$ here is comparable to the variance of Y in Section 2.2.1). The left panel of Figure 2.2 gives the w-functions, which are very similar to those of the FAB z-procedure displayed in Figure 2.1, but with somewhat larger derivatives near μ. The second panel gives the FAB t-intervals as functions of $\bar{y}$, with s^2 fixed at 10. Again, the intervals resemble the corresponding z-intervals, but are slightly wider due to the use of t-quantiles instead of z-quantiles. The effect of not knowing σ^2 is more noticeable in the plot of the risk functions, given in the right-upper plot. While the shapes of the risk functions are similar to those of the analogous z-intervals, the risks (expected widths) are larger due to the fact that the width of a t-interval is dependent on S^2, which is proportional to a χ^2_9 random variable having non-trivial skew.

2.3 Empirical FAB intervals for multigroup data

A potential obstacle to the adoption of FAB confidence intervals is the aversion that many researchers have to specifying a distribution over θ. However, in multigroup data settings, probabilistic information about the mean θ_j of one group is may be obtained from data of

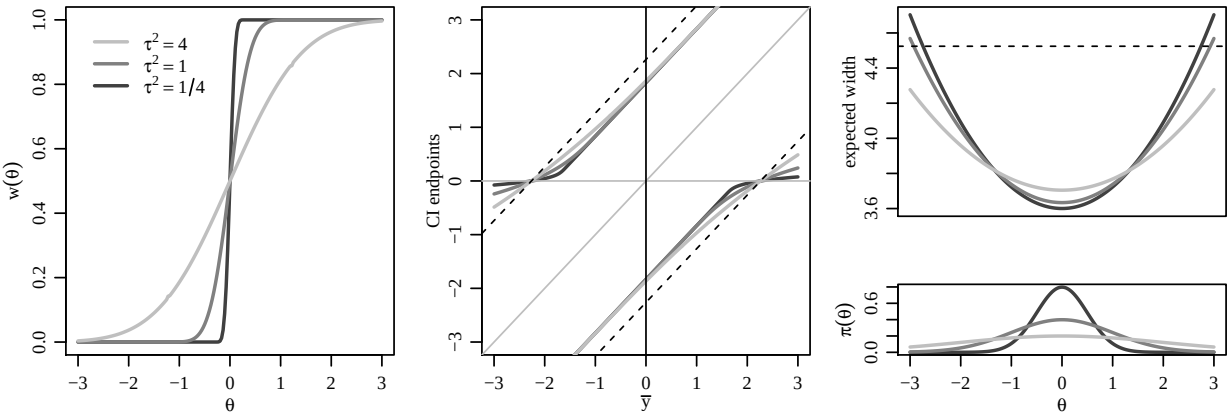

Figure 2.2: Descriptions of the FAB t-procedure. The left plot gives Bayes-optimal w-functions for three values of τ^2, at level $\alpha = 0.05$. The middle plot gives the corresponding confidence interval procedures with s^2 fixed at 10. The top plot on the right gives the expected widths of the 95% FAB t-intervals for the three values of τ^2, with the corresponding prior densities plotted below.

the other groups. This information can be used to specify a probability distribution π for the likely values of θ_j, from which an empirical FAB interval may be constructed. Such an interval will have exact $1-\alpha$ coverage for every value of θ_j, but a shorter expected width for values that are deemed likely by π. For the usual homoscedastic hierarchical normal model having a common within-group variance, we develop such a procedure that may be used in practice, and show that it is risk-optimal asymptotically in the number of groups. We also develop a similar procedure for the case of heteroscedastic groups.

2.3.1 Asymptotically optimal procedure for homoscedastic groups

Consider the case of p normal populations with means $\theta_1, \ldots, \theta_p$ and common variance and sample size, so that $Y_{1,j}, \ldots Y_{n,j} \sim$ i.i.d. $N(\theta_j, \sigma^2)$ independently across groups (common sample sizes are used here solely to simplify notation). The standard hierarchical normal model posits that the heterogeneity across groups can be described by a normal distribution,

so that $\theta_1, \ldots, \theta_p \sim$ i.i.d. $N(\mu, \tau^2)$. In the multigroup setting, this normal distribution is not considered to be a prior distribution for a single θ_j, but instead is a statistical model for the across-group heterogeneity of $\theta_1, \ldots, \theta_p$. The parameters describing the across- and within-group heterogeneity are $\psi = (\mu, \tau^2, \sigma^2)$.

For each group j let C^j be a $1-\alpha$ CRP for θ_j that possibly depends on data from all of the other groups. Letting $\boldsymbol{C} = \{C^1, \ldots, C^p\}$ we define the risk of such a multigroup confidence procedure as

$$R(\boldsymbol{C}, \psi) = \frac{1}{p} \sum_{j=1}^{p} \mathrm{E}[|C^j(\boldsymbol{Y})|],$$

where $\boldsymbol{Y}$ is the data from all of the groups and the expectation is over both $\boldsymbol{Y}$ and $\theta_1, \ldots, \theta_p$. Under the hierarchical normal model, the risk at a value of ψ is minimized by letting each C^j be equal to $C_{w_\psi}(\bar{y}_j)$, the FAB z-interval defined in Section 2.2 but with $\psi = (\mu, \tau^2, \sigma^2/n)$, since $\mathrm{Var}[\bar{Y}_j|\theta_j] = \sigma^2/n$. The oracle multigroup confidence procedure is then $\boldsymbol{C}_{w_\psi} = \{C_{w_\psi}(\bar{y}_1), \ldots, C_{w_\psi}(\bar{y}_p)\}$, which has risk

$$R(\boldsymbol{C}_{w_\psi}, \psi) = \frac{1}{p} \sum_{j=1}^{p} \mathrm{E}[|C_{w_\psi}(\bar{y}_j)|] = \mathrm{E}[|C_{w_\psi}(\bar{y})|],$$

where $\bar{y} \sim N(\theta, \sigma^2/n)$ and $\theta \sim N(\mu, \tau^2)$. While this oracle procedure is generally unavailable in practice, estimates of ψ may be obtained from the data and used to construct a multigroup CIP that achieves the oracle risk asymptotically as $p \to \infty$. To show how to do this, we first construct a $1-\alpha$ CIP for a single θ based on $\bar{Y} \sim N(\theta, \sigma^2/n)$ and independent estimates S^2 and $\hat{\psi}$ of σ^2 and ψ. We show how the risk of this CIP converges to the oracle risk as $S^2 \xrightarrow{a.s.} \sigma^2$ and $\hat{\psi} \xrightarrow{a.s.} \psi$, and then show how to use this fact to construct an asymptotically optimal multigroup CIP.

The ingredients of our FAB CIP for a single population mean θ are as follows: Let $\bar{Y} \sim N(\theta, \sigma^2/n)$ and $qS^2/\sigma^2 \sim \chi^2_q$ be independent. Consider the $1-\alpha$ CRP for θ given by

$$C_w(\bar{y}, s^2) = \{\theta : \bar{y} + \tfrac{s}{\sqrt{n}} t_{\alpha(1-w(\theta))} < \theta < \bar{y} + \tfrac{s}{\sqrt{n}} t_{1-\alpha w(\theta)}\}, \tag{2.12}$$

where the t-quantiles are those of the t_q-distribution. As described in Section 2.2.2, this procedure has $1-\alpha$ coverage for every value of θ and is an interval if $w : \mathbb{R} \to (0, 1)$

is a continuous nondecreasing function. This holds for non-random w-functions as well as for random w-functions that are independent of $\bar{Y}$ and S^2. In particular, suppose we have estimates $\hat{\psi} = (\hat{\mu}, \hat{\tau}^2, \hat{\sigma}^2/n)$ that are independent of $\bar{Y}$ and S^2. We can then let $w = w_{\hat{\psi}}$, the w-function of the Bayes optimal z-interval assuming a prior distribution $\theta \sim N(\hat{\mu}, \hat{\tau}^2)$ and that $\text{Var}[\bar{Y}|\theta] = \hat{\sigma}^2/n$. Note that we are not assuming $(\mu, \tau^2, \sigma^2/n)$ actually equals $(\hat{\mu}, \hat{\tau}^2, \hat{\sigma}^2/n)$, we are just using these values to approximate the optimal w-function by $w_{\hat{\psi}}$ and the optimal CIP by $C_{w_{\hat{\psi}}}$.

The random interval $C_{w_{\hat{\psi}}}(\bar{Y}, S^2)$ differs from the optimal interval $C_{w_\psi}(\bar{Y})$ in three ways: First, the former uses S^2 instead of σ^2 to scale the endpoints of the interval. Second, the former uses t-quantiles instead of standard normal quantiles. Third, the former uses $\hat{\psi}$ to define the w-function, instead of ψ. Now as q increases, $S^2 \xrightarrow{a.s.} \sigma^2$ and the t-quantiles in (2.12) converge to the corresponding z-quantiles. If we are also in a scenario where $\hat{\psi}$ can be indexed by q and $\hat{\psi} \xrightarrow{a.s.} \psi$, then we expect that $w_{\hat{\psi}}$ converges to w_ψ and that the risk of $C_{w_{\hat{\psi}}}$ converges to the oracle risk:

Proposition 2.3.1. *Let $\bar{Y} \sim N(\theta, \sigma^2/n)$, $qS^2/\sigma^2 \sim \chi^2_q$, and $\hat{\psi}$ be independent for each value of q, with $\hat{\psi} \xrightarrow{a.s.} \psi$ as $q \to \infty$. Then*

1. *$C_{w_{\hat{\psi}}}$ defined in (2.12) is a $1-\alpha$ CIP for each value of θ and q;*

2. *$E[|C_{w_{\hat{\psi}}}|] \to E[|C_{w_\psi}|]$ as $q \to \infty$.*

We now return to the problem of constructing an asymptotically optimal multigroup procedure. Let $\bar{Y}_j$ and S_j^2 be the sample mean and variance for a given group j. Divide the remaining groups into two sets, with $p_1 - 1$ in the first set and $p_2 = p - p_1$ in the second. Pool the group-specific sample variances of the first set of groups with S_j^2 to obtain an estimate $\tilde{S}_j^2$ of σ^2, so that

$$\tfrac{p_1(n-1)}{\sigma^2}\tilde{S}_j^2 \sim \chi^2_{p_1(n-1)}.$$

From the remaining groups, obtain a strongly consistent estimate $\hat{\psi}_j$ of ψ (such as the MLE or a moment-based estimate). Then $\bar{Y}_j$, $\tilde{S}_j^2$ and $\hat{\psi}_j$ are independent for each value of p.

Therefore, a $1-\alpha$ CIP for θ_j is given by

$$C_{w_{\hat{\psi}_j}}(\bar{y}_j, \tilde{s}_j^2) = \{\theta_j : \bar{y}_j + \tfrac{\tilde{s}_j}{\sqrt{n}} t_{\alpha(1-w_{\hat{\psi}_j}(\theta_j))} < \theta_j < \bar{y}_j + \tfrac{\tilde{s}_j}{\sqrt{n}} t_{1-\alpha w_{\hat{\psi}_j}(\theta_j)}\}, \tag{2.13}$$

where the quantiles are those of the $t_{p_1(n-1)}$ distribution. If p_1 is chosen so that it remains a fixed fraction of p as p increases, then $\tilde{S}_j^2$ and $\hat{\psi}_j$ converge to σ^2 and ψ respectively, and the t-quantiles converge to the corresponding standard normal quantiles. By Proposition 2.3.1, the risk of this interval converges to that of the oracle risk. Repeating this construction for each group j results in a multigroup confidence procedure that has $1-\alpha$ coverage for each group *conditional* on $(\theta_1, \ldots, \theta_p)$, but is also asymptotically optimal on average across the $N(\mu, \tau^2)$ population of θ-values.

In practice for finite p, different choices of p_1 and p_2 will affect the resulting confidence intervals. Since the minimal width of each interval is directly tied to the degrees of freedom $p_1(n-1)$ of the variance estimate $\tilde{S}_j^2$, we suggest choosing p_1 to ensure that the quantiles of the $t_{p_1(n-1)}$ distribution are reasonably close to those of the standard normal distribution. If either p or n are large, this can be done while still allowing p_2 to be large enough for $(\hat{\mu}, \hat{\tau}^2, \hat{\sigma}^2/n)$ to be useful.

2.3.2 A procedure for heteroscedastic groups

If a researcher is unwilling to assume a common within-group variance, constant $1-\alpha$ group-specific coverage can still be ensured by using intervals of the form

$$C_{w_j}(\bar{y}_j, s_j^2) = \{\theta_j : \bar{y}_j + \tfrac{s_j}{\sqrt{n_j}} t_{\alpha(1-w_j(\theta_j))} < \theta_j < \bar{y}_j + \tfrac{s_j}{\sqrt{n_j}} t_{1-\alpha w_j(\theta_j)}\}, \tag{2.14}$$

where w_j is an estimate of the Bayes-optimal w-function discussed at the end of Section 2.2.2, estimated with data from groups other than j. We recommend obtaining w_j from a hierarchical model for both the group-specific means and variances, as this allows across-group sharing of information about both of these quantities. For example, the replication material for this chapter provides code to obtain estimates of the w-function that is optimal

for the following model of across-group heterogeneity:

$$\theta_1, \ldots, \theta_p \sim \text{i.i.d. } N(\mu, \tau^2) \tag{2.15}$$
$$1/\sigma_1^2, \ldots, 1/\sigma_p^2 \sim \text{i.i.d. gamma}(a, b).$$

We estimate the across-group heterogeneity parameters (μ, τ^2, a, b) as follows: For each group j let

$$X_j^2 = \sum_i (Y_{i,j} - \bar{Y}_j)^2 \sim \sigma_j^2 \chi^2_{n_j - 1}.$$

If $1/\sigma_j^2 \sim \text{gamma}(a, b)$ independently for each j then the marginal density of $X_1^2, \ldots, X_p^2$ can be shown to be

$$p(x_1^2, \ldots, x_p^2 | a, b) \prod_{j=1}^{p} c(x_j^2) \frac{\Gamma(a + (n_j - 1)/2) b^a}{\Gamma(a)(b + x_j^2/2)^{a + (n_j - 1)/2}},$$

where c is a function that does not depend on a or b. This quantity can be maximized to obtain marginal maximum likelihood estimates of $\hat{a}$ and $\hat{b}$. Now if $\sigma_1^2, \ldots, \sigma_p^2$ were known, then a maximum likelihood estimate of (μ, τ^2) could be obtained based on the fact that under the hierarchical model, $\bar{Y}_j \sim N(\mu, \sigma_j^2/n_j + \tau^2)$ independently across groups. Since the σ_j^2's are not known we use empirical Bayes estimates, given by $\hat{\sigma}_j^2 = (\hat{b} + x_j^2/2)/(\hat{a} + (n_j - 1)/2)$, to obtain the "plug-in" marginal likelihood estimates $(\hat{\mu}, \hat{\tau}^2)$:

$$(\hat{\mu}, \hat{\tau}^2) = \arg\max_{\mu, \tau^2} \prod_j \frac{1}{\sqrt{\hat{\sigma}_j^2/n_j + \tau^2}} \phi\left(\frac{\bar{y}_j - \mu}{\sqrt{\hat{\sigma}_j^2/n_j + \tau^2}} \right),$$

where ϕ is the standard normal probability density function.

To create a FAB t-interval for a given group j, we obtain estimates $(\hat{\mu}_j, \hat{\tau}_j^2, \hat{a}_j, \hat{b}_j)$ using the procedure described above with data from all groups except group j. The w-function w_j for group j is taken to be the Bayes-optimal w-function defined by Equation 2.11, under the estimated prior $\theta_j \sim N(\hat{\mu}_j, \hat{\tau}_j^2)$ and $1/\sigma_j^2 \sim \text{gamma}(\hat{a}_j, \hat{b}_j)$. The independence of $(\bar{Y}_j, S_j^2)$ and $(\hat{\mu}_j, \hat{\tau}_j^2, \hat{a}_j, \hat{b}_j)$ ensures that the resulting FAB t-interval has exact $1 - \alpha$ coverage, conditional on $\theta_1, \ldots, \theta_p$ and $\sigma_1^2, \ldots, \sigma_p^2$.

We speculate that this procedure enjoys similar optimality properties to those of the approach for homoscedastic groups described in Section 2.3.1: If the hierarchical model given

by (2.15) is correct, then as the number p of groups increases, the estimates $(\hat{\mu}_j, \hat{\tau}_j^2, \hat{a}_j, \hat{b}_j)$ will converge to (μ, τ^2, a, b) and the interval for a given group will converge to the corresponding Bayes-optimal interval. So far we have been unable to prove this, the primary difficulty being that the Bayes-optimal w function given by Equation 2.11 is a non-standard integral involving the non-central t-distribution, and is not easily studied analytically.

2.4 Radon data example

A study by the U.S. Environmental Protection Agency measured radon levels in a random sample of homes. Price et al. (1996) use a subsample of these data to estimate county-specific mean radon levels (on a log scale) in the state of Minnesota. This dataset consists of log radon values measured in 919 homes, each being located in one of $p = 85$ counties. County-specific sample sizes ranged from 1 to 116 homes. In this section we obtain a 95% FAB confidence interval for each county-specific mean radon level, based on data from all of the counties, and compare these intervals to the corresponding UMAU intervals. Also, in a simulation study based on these data, we compare the expected widths of these two types of intervals to empirical Bayes posterior intervals, and show how the latter do not provide constant coverage across values of the county-specific means.

2.4.1 County-specific confidence intervals

Letting $Y_{i,j}$ be the radon measurement for home i in county j, we assume throughout this section that $Y_{1,j} \ldots, Y_{n_j,j} \sim$ i.i.d. $N(\theta_j, \sigma_j^2)$ and that the data are independently sampled across counties. Under the assumptions of a constant across-county variance and the normal hierarchical model $\theta_1, \ldots, \theta_p \sim$ i.i.d. $N(\mu, \tau^2)$, the maximum likelihood estimates of σ^2, μ and τ^2 are $\hat{\sigma}^2 = 0.637$, $\hat{\mu} = 1.313$ and $\hat{\tau}^2 = 0.096$. The estimate of the across-county variability is substantially smaller than the estimate of within-county variability, suggesting that there is useful information to be shared across the groups. However, Levene's test of heteroscedasticity (an F-test using the absolute difference between the data and group-specific medians) rejects the null of homoscedasticity with a p-value of 0.011. For this reason,

we use the FAB t-interval procedure described in Section 2.3.2 for each group, having the form $\{\theta_j : \bar{y}_j + \sqrt{s_j^2/n_j} \times t_{\alpha(1-w_j(\theta_j))} < \theta_j < \bar{y}_j + \sqrt{s_j^2/n_j} \times t_{1-\alpha w_j(\theta_j)}\}$, where $\alpha = .05$, $\bar{y}_j$ and s_j^2 are the sample mean and variance from county j, and w_j is the optimal w-function assuming $\theta_j \sim N(\hat{\mu}_j, \hat{\tau}_j^2)$ and $1/\sigma_j^2 \sim \text{gamma}(\hat{a}_j, \hat{b}_j)$, where $\hat{\mu}_j, \hat{\tau}_j^2, \hat{a}_j$ and $\hat{b}_j$ are estimated from the counties other than county j. Such intervals have 95% coverage for each county, assuming only within-group normality.

We constructed FAB and UMAU intervals for each county that had a sample size greater than one, i.e. counties for which we could obtain an unbiased within-sample variance estimate. Intervals for counties with sample sizes greater than two are displayed in Figure 2.3 (intervals based on a sample size of two were excluded from the figure because their widths make smaller intervals difficult to visualize). The UMAU intervals are wider than the FAB intervals for 77 of the 82 counties having a sample size greater than 1, and are 30% wider on average across counties. Generally speaking, the counties for which the FAB intervals provide the biggest improvement are those with smaller sample sizes and sample means near the across-group average. Conversely, the five counties for which the UMAU intervals are narrower than the FAB interval are those with moderate to large sample sizes, and sample means somewhat distant from the across-group average.

2.4.2 Risk performance and comparison to posterior intervals

Assuming within-group normality, the FAB interval procedure described above has 95% coverage for each group j and for all values of $\theta_1, \ldots, \theta_p$. Furthermore, the procedure is designed to approximately minimize the expected risk under the hierarchical model $\theta_1, \ldots, \theta_p \sim$ i.i.d. $N(\mu, \tau^2)$, among all 95% CRPs. However, one may wonder how well the FAB procedure works for fixed values of $\theta_1, \ldots, \theta_p$. This question is particularly relevant in cases where the hierarchical model is misspecified, or if a hierarchical model is not appropriate (e.g., if the groups are not sampled). We investigate this for the radon data with a simulation study in which we take the county-specific sample means and variances as the true county-specific values, that is, we set $\theta_j = \bar{y}_j$ and $\sigma_j^2 = s_j^2$ for each county j. We then simulate n_j observations

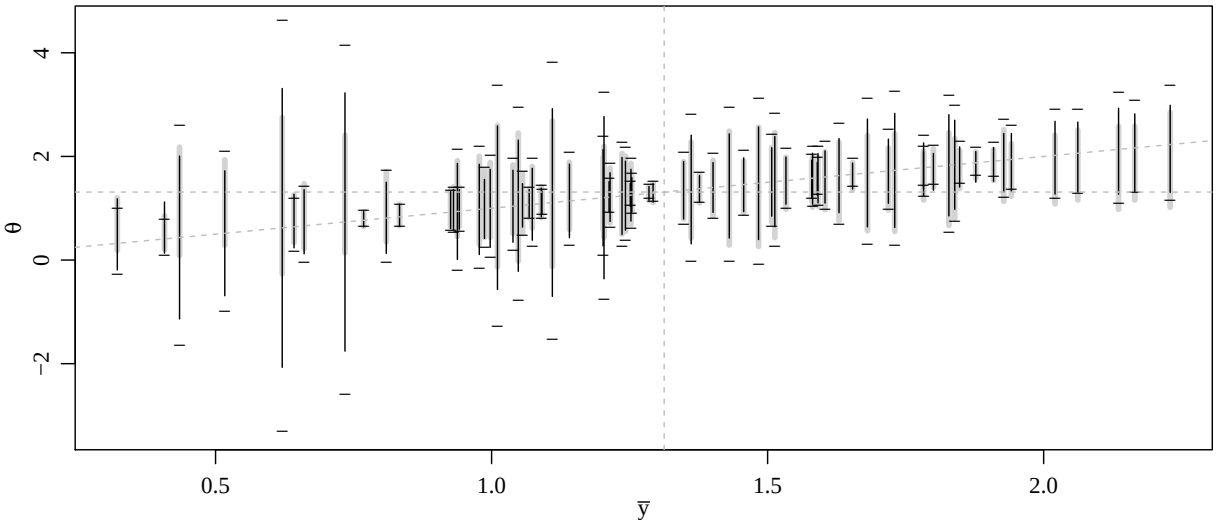

Figure 2.3: FAB and UMAU 95% confidence intervals for the radon dataset. The UMAU intervals are plotted as wide gray lines, the FAB intervals as narrow black lines. Vertical and horizontal dashed lines are drawn at $\sum \bar{y}_j/p$, and the other dashed gray line is the line of $\bar{y} = \theta$.

for each county j from the model $Y_{1,j}, \ldots, Y_{n_j,j} \sim$ i.i.d. $N(\theta_j, \sigma_j^2)$.

We generated 10,000 such simulated datasets. For each dataset, we computed the widths of the 95% FAB and UMAU confidence intervals for each county having a sample size greater than one. Additionally, for comparison we also computed empirical Bayes posterior intervals, which are often used in hierarchical modeling. The posterior interval for group j is given by $\hat{\theta}_j \pm t_{1-\alpha/2} \times (1/\hat{\tau}^2 + n_j/\hat{s}_j^2)^{-1/2}$, where $\hat{\theta}_j$ is the empirical Bayes estimator given by

$$\hat{\theta}_j = \frac{\hat{\mu}/\hat{\tau}^2 + \bar{y}_j n_j/s_j^2}{1/\hat{\tau}^2 + n_j/s_j^2},$$

and $t_{1-\alpha/2}$ is the $1-\alpha/2$ quantile of the t_{n-1}-distribution. As discussed in the Introduction, such intervals are always narrower than the corresponding UMAU intervals but will not have $1-\alpha$ frequentist coverage for each group. Instead, such intervals generally have $1-\alpha$ coverage on average, or in expectation with respect to the hierarchical model over the θ_j's.

The results of this simulation study are displayed in Figure 2.4. The left panel of the figure gives the expected widths of the FAB and Bayes procedures relative to those of the UMAU procedure. Based on the 10,000 simulated datasets, the estimated expected widths across counties were about 2.28, 1.60 and 1.61, respectively for the UMAU, FAB and Bayes procedures respectively. As with the actual interval widths for the non-simulated data, expected widths of the FAB intervals are smaller than those of the UMAU intervals for most counties (79 out of 82). The Bayes intervals are always narrower than the UMAU intervals for all groups by construction. However, while they tend to be narrower than the FAB intervals for θ_j's far from $\bar{\theta} = \sum \theta_j/p$, near this average they are often wider than the FAB intervals. This is not too surprising - the FAB intervals are at their narrowest near this overall average, while the Bayes intervals tend to over-cover here. This latter issue is illustrated in the right panel of the figure, which shows how the Bayes credible intervals do not have constant coverage across groups. This is because the Bayes intervals are centered around biased estimates that are shrunk towards the estimated overall mean $\bar{\theta}$. If θ_j is far from $\bar{\theta}$ then the bias is high and the coverage is too low, whereas if θ_j is near $\bar{\theta}$ the coverage is too high since the variability of the shrinkage estimate $\hat{\theta}_j$ is lower than that of

$\bar{y}_j$. The group-specific coverage rates of the Bayes intervals vary from about 91% to 98%, although the average coverage rate across groups is approximately 95%. In summary, the UMAU procedure provides constant $1-\alpha$ coverage across groups, but wider intervals than those obtained from the FAB and Bayes procedures. The Bayes procedure provides narrower intervals but non-constant coverage. The FAB procedure provides both narrower intervals and constant coverage.

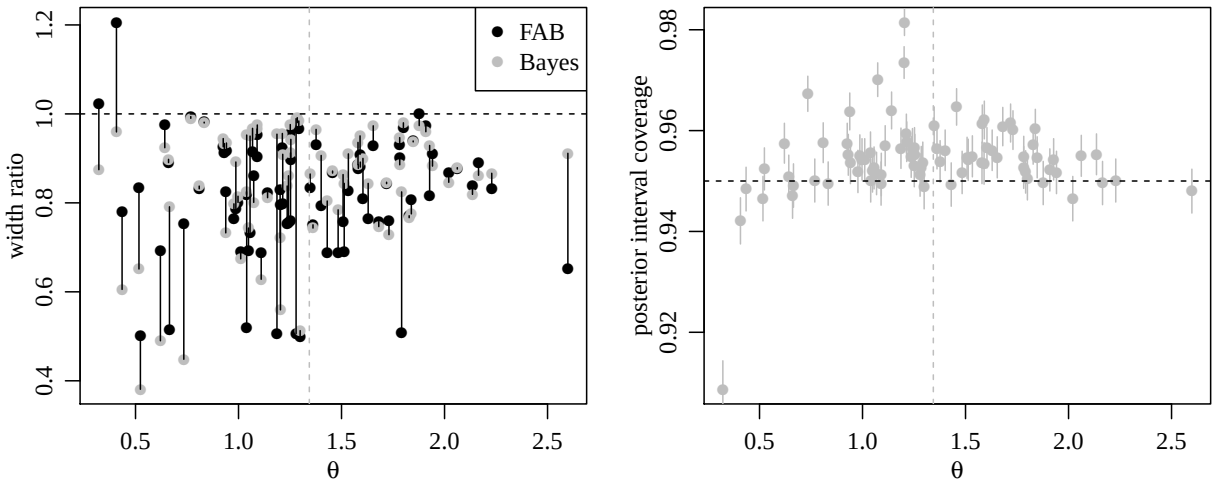

Figure 2.4: Simulation results. The left panel gives relative expected interval widths of the FAB and Bayes procedures relative to the UMAU procedure. The right panel indicates how coverage rates of Bayes posterior intervals are not constant across groups. Points are coverage rates based on 10,000 simulated datasets, and vertical lines are nominal 95% intervals representing Monte Carlo standard error. Vertical lines are drawn at $\sum \theta_j/p$ in each panel.

2.5 Robustness of the FAB Confidence Interval

In this section, we examine the robustness of the FAB confidence interval procedure C_{w_ϕ}. We first discuss what happens if the parameter τ^2 that describes the variance of the prior is misspecified. Then we examine the performance of the Bayes-optimal confidence interval

when the distribution of the parameter of interest is not Gaussian. Note that the coverage of the Bayes-optimal confidence is always guaranteed. Hence we only examine the width of the confidence interval here. In this discussion, we focus on the FAB z-interval.

2.5.1 Bayes risk of the Bayes-optimal Confidence Interval Under Misspecified Priors

Previously, under the prior of $\theta \sim N(0, \tau^2)$, we have shown that C_{w_ϕ} is Bayes-optimal with exact frequentist coverage $1-\alpha$. Now suppose two Bayesians, Bayesian A and Bayesian B, have different prior information. Bayesian A's prior π_A is $N(0, \tau_A^2)$ and used this prior to construct the confidence interval procedure (CIP) C_{w_ϕ}, which will be denoted as $C^{\tau_A^2}$ in this section. Bayesian B's prior π_B is $N(0, \tau_B^2)$ where τ_B^2 does not equal to τ_A^2. Now using Bayes risk as a criterion, Bayesian B examines the performance of $C^{\tau_A^2}$ constructed by Bayesian A, specifically as compared to $C_{1/2}$, which will be denoted as C^∞ in this section. From the perspective of Bayesian B, $C^{\tau_A^2}$ will not be Bayes-optimal, but it still may be better than C^∞ in terms of its Bayes risk.

First, the frequentist coverage of $C^{\tau_A^2}$ still maintains the exact frequentist coverage rate regardless of the distribution of θ. For the width of $C^{\tau_A^2}$ averaged over the distribution $N(0, \tau_B^2)$, we will show that as long as τ_A^2 is greater than some value $s(\tau_B^2)$, the Bayes width of $C^{\tau_A^2}$ will be shorter than C^∞. Here, $s(\tau_B^2)$ is a function of only τ_B^2 when the level α is fixed. We will also show that this condition is easy to achieve as $s(\tau_B^2)$ is much smaller than τ_B^2 itself. Thus Bayesian B will consider $C^{\tau_A^2}$ be a better procedure than C^∞, unless τ_A^2 is much smaller than τ_B^2.

First, it can be shown, with the help of the following lemma, that $s(\tau_B^2) = 0$ when τ_B^2 is within some range.

Lemma 2.5.1. *For any $\tau_A^2 > 0$,*

$$R(\pi_B, C^{\tau_A^2}) < 2\Phi^{-1}(1-\alpha)\Phi\left(\frac{\Phi^{-1}(1-\alpha)}{\sqrt{1+\tau_B^2}}\right) + 2\sqrt{1+\tau_B^2}\,\phi\left(\frac{\Phi^{-1}(1-\alpha)}{\sqrt{1+\tau_B^2}}\right). \tag{2.16}$$

If we let the right side of (2.16) $\leq 2\Phi^{-1}(1-\alpha/2)$, we can identify a critical point τ^{*2}, such that when $\tau_B^2 \leq \tau^{*2}$, $r(\pi_B, C^{\tau_A^2}) \leq 2\Phi^{-1}(1-\alpha/2)$. In this case, for any $\tau_A^2 > 0$, $r(\pi_B, C^{\tau_A^2})$ is less than $r(\pi_B, C^{\infty})$, i.e. $s(\tau_B^2) = 0$. For example, when $\alpha = 0.05$, $\tau^{*2} = 4.2$. Thus, if the prior variance of Bayesian B is less than 4.2, she/he will consider $C^{\tau_A^2}$ as a better CIP than the UMAU one, for any τ_A^2 used by Bayesian A.

When τ_B^2 is greater than τ^{*2}, $s(\tau_B^2)$ is no longer 0, but we will numerically show that $s(\tau_B^2)$ is much smaller than τ_B^2 itself. For a fixed α, the Bayes risk $r(\pi_B, C^{\tau_A^2})$ is determined by the values of τ_A^2 and τ_B^2. Thus for a given τ_B^2 value, we can identify the value of $s(\tau_B^2)$ numerically.

Figure 2.5 gives the curve of $s(\tau_B^2)$ as a function of τ_B^2. As we can see from the figure, $s(\tau_B^2)$ is much smaller than τ_B^2. For example, when $\tau_B^2 = 6$, $s(\tau_B^2) = 0.87$. This means although Bayesian B has prior variance 6, she/he will favor $C^{\tau_A^2}$ over C^{∞} as long as τ_A^2 is greater than 0.87. In summary, Bayesian B will prefer Bayesian A's optimal CIP to the UMAU CIP as long as the prior variance of Bayesian A is not dramatically smaller than that of Bayesian B.

2.5.2 Model Misspecification

Previously, we evaluate the FAB CI under the assumption that the distribution of θ is normal. However, in applications the true distribution of the parameter θ is often unknown, and it is likely to be non-normal. In this section, we examine the FAB CI when $\pi(\theta)$ is not normal. Suppose we still have observations of Y, with $Y \sim N(\theta, 1)$ and $\theta \sim \pi$ without assuming π is normal. We still construct the FAB CI as in previous section, but will evaluate it under a non-normal distribution $\pi(\theta)$.

Without loss of generality, we assume π to be a mean 0 distribution. In Section 2.5.1, we have discussed the misspecification of the prior, where we assume the true distribution of θ is normal but the variance τ^2 may be misspecified. For empirical Bayes case, we do not pre-specify τ^2, instead we estimate it from the observations. Asymptotically, we can get the accurate τ^2 value from the observations. So in that case, evaluating the Bayes risk of $C_{w_{\hat{\phi}}}$ is

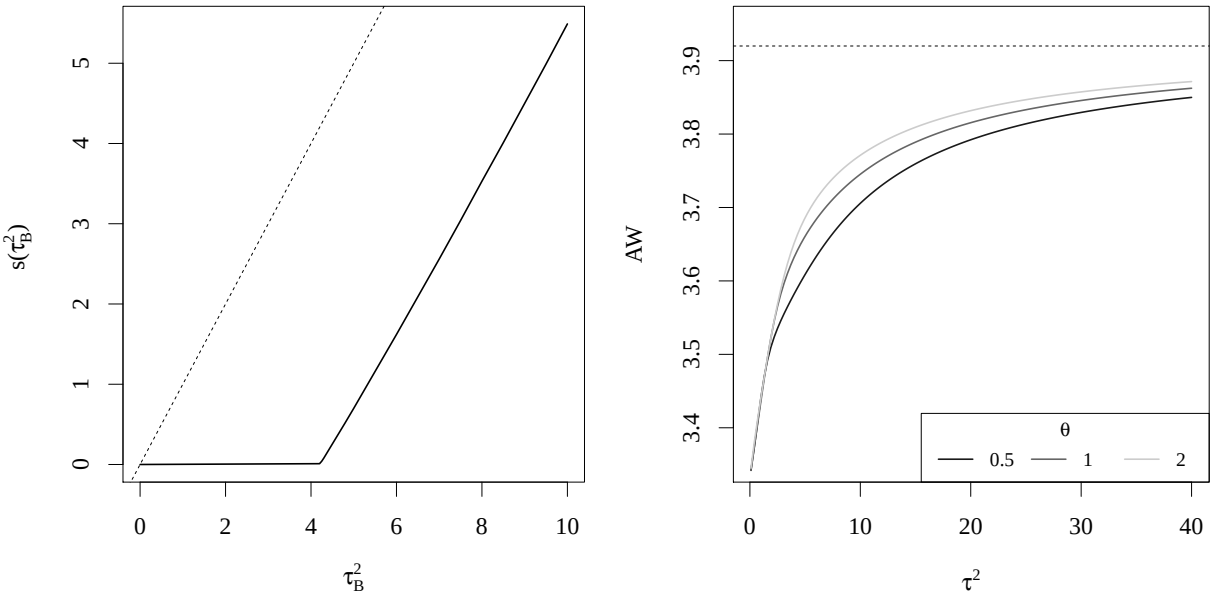

Figure 2.5: The solid line in the graph on the left shows the curve of $s(\tau^B)$; the dashed line is a 45° line through the origin. The graph on the right shows $AW(\theta, \tau^2)$ as a function of τ^2 for different fixed θ values; the dashed line represents the width of the UMAU procedure.

equivalent to evaluating the Bayes risk of C_{w_ϕ}, with the variance of the prior π is correctly specified but the distribution of π is not necessarily normal. Our speculation is that the Bayes risk of C_{w_ϕ} is lower than that of $C_{1/2}$ regardless of the distribution of θ, as long as we have an accurate estimate of the variance. To show this speculation holds for all distribution families, a lemma is introduced below to simplify the target from all families to one specific family.

Lemma 2.5.2. *Suppose*

$$\theta|\lambda \sim f(\theta|\lambda)$$

and the distribution function of θ is a mixture of f

$$\pi(\theta) = \int_\Lambda f(\theta|\lambda) g(d\lambda).$$

If we have

$$R(f(\theta|\lambda), C_1) \leq R(f(\theta|\lambda), C_2)$$

for all λ*, then*

$$R(\pi, C_1) \leq R(\pi, C_2).$$

Thus if we can find a class of distributions with variance τ^2, under which the Bayes risk of C_{w_ϕ} is lower than that of $C_{1/2}$, then under a mixture of distributions from this class, C_{w_ϕ} will still has lower Bayes risk than $C_{1/2}$. Moreover, if any distribution with mean zero and variance τ^2 can be expressed as the mixture of distributions from this particular class, then our speculation is proved. One such class is the class of two-point distributions with mean zero and variance τ^2. This distribution family can be parameterized as

$$\Pr(\theta = x) = \tfrac{\theta_l}{\theta_l+\theta_h}\delta(x = \theta_h) + \tfrac{\theta_h}{\theta_l+\theta_h}\delta(x = \theta_l).$$

Since we require $\mathrm{E}[\theta] = 0$ and $\mathrm{Var}[\theta] = \tau^2$, together we can solve that:

$$\theta_l\theta_h = \tau^2.$$

Thus the distribution can be written as:

$$\Pr(\theta = x) = \tfrac{\tau^2}{\tau^2+\theta_h^2}\delta(x = \theta_h) + \tfrac{\theta_h^2}{\tau^2+\theta_h^2}\delta(x = \theta_l).$$

Recall the expected width of our CI at θ is:

$$\begin{aligned} EW(\theta, \tau^2) = \int_\Omega & \Phi(\theta - t - \Phi^{-1}(g^{-1}(\frac{2t}{\tau^2})\alpha)) \\ & - \Phi(\theta - t + \Phi^{-1}((1 - g^{-1}(\frac{2t}{\tau^2}))\alpha))dt. \end{aligned}$$

Thus the expected width averaged with respect to the two-point distribution prior, i.e. the Bayes risk, is

$$AW(\theta_h, \tau^2) = \frac{\tau^2}{\tau^2 + \theta_h^2} EW(\theta_h, \tau^2) + \frac{\theta_h^2}{\tau^2 + \theta_h^2} EW(\tau^2/\theta_h, \tau^2).$$

The shape of $AW(\theta_h, \tau^2)$ is described in Figure 2.5. As we can see from the graph, when fixing θ_h values, $AW(\theta_h, \tau^2)$ is monotonically increasing in τ^2. As we have explained before, $\lim_{\tau^2 \to \infty} EW(\theta_h, \tau^2) = 2\Phi^{-1}(1-\alpha/2)$, thus $\lim_{\tau^2 \to \infty} AW(\theta_h, \tau^2) \to 2\Phi^{-1}(1-\alpha/2)$. Therefore $AW(\theta_h, \tau^2) \leq 2\Phi^{-1}(1-\alpha/2)$. Thus for the two-point distributions, the average width is less than or equal to the width of the UMAU procedure. Together with Lemma 2.5.2, we see that FAB confidence interval could work for any distribution that can be written as the mixture of the class of two-points distributions. These support the speculation that C_{w_ϕ} always has a better Bayes risk than $C_{1/2}$, though more theoretical work are needed in the future to prove this rigorously.

2.6 Discussion

Standard analyses of multilevel data utilize multigroup confidence interval procedures that either have constant coverage but do not share information across groups, or share information across groups but lack constant coverage. These latter procedures typically do maintain a pre-specified coverage rate on average across groups, but the value of this property is unclear if one wants to make group-specific inferences. The FAB procedures developed in this chapter have coverage rates that are constant in the mean parameter, and so maintain constant coverage for each group selected into the dataset, while also making use of across-group information. The FAB procedures are approximately optimal among constant coverage procedures if the across-group heterogeneity is well-represented by a normal hierarchical model.

If the across-group heterogeneity is not well-represented by a hierarchical normal model, then the FAB procedure will still maintain the chosen constant coverage rate but may not be optimal. We speculate that in such cases, the FAB procedure based on a hierarchical normal model, while not optimal, will still have better risk than the UMAU procedure when the across-group heterogeneity corresponds to any probability distribution with a finite second moment. This is partly because the UMAU procedure is a limiting case of the FAB procedure as the across-group variance goes to infinity. We have developed an analytical argument of this and have gathered computational evidence, but a complete proof of the dominance of

a misspecified FAB procedure over the UMAU procedure is still a work in progress. Of course, the basic idea behind the FAB procedure could be implemented with alternative models describing across-group heterogeneity, such as models that allow for sparsity among the group-level parameters. We have implemented a few such procedures computationally, but studying them analytically is challenging.

Replication code for this chapter can be found at the following website: *pdhoff.github.io*. The multigroup FAB procedures discussed in Sections 2.3 and 2.4 are provided by the R-package `fabCI`.

Appendix: Proofs

Proof of Lemma 2.2.1. This lemma follows from Ferguson (1967, Section 5.3), which says that for any level-α test of a point null hypothesis for a one-parameter exponential family, there exists a two-sided test of equal or greater power. Let $\{\tilde{\phi}_\theta(y) : \theta \in \mathbb{R}\}, \{\tilde{A}(\theta) : \theta \in \mathbb{R}\}$ be the test functions and acceptance regions corresponding to the CRP $\tilde{C}(y)$. The coverage of $\tilde{C}$ is

$$\Pr(Y \in \tilde{A}(\theta)|\theta) = 1 - \mathrm{E}[\tilde{\phi}_\theta(Y)|\theta]. \tag{2.17}$$

By Theorem 2 from Ferguson (1967, Section 5.3), for each $\theta \in \mathbb{R}$ there exists a two-sided test ϕ_θ such that

$$\mathrm{E}[\phi_\theta(Y)|\theta] = \mathrm{E}[\tilde{\phi}_\theta(Y)|\theta]. \tag{2.18}$$

Denote the acceptance regions corresponding to these two-sided test as $\{A(\theta) : \theta \in \mathbb{R}\}$. Inverting these regions gives a CIP $C(y)$. The coverage of $C(y)$ is

$$\Pr(Y \in A(\theta)|\theta) = 1 - \mathrm{E}[\phi_\theta(Y)|\theta]. \tag{2.19}$$

Hence by (2.17), (2.19), and (2.18), the coverage of $C(y)$ is the same as the coverage of $\tilde{C}(y)$. The width of $\tilde{C}(y)$ is:

$$W(y) = \int_{\mathbb{R}} 1(t \in \tilde{C}(y))dt = \int_{\mathbb{R}} 1(y \in \tilde{A}(t))dt = \int_{\mathbb{R}} (1 - \tilde{\phi}_t(y))dt.$$

The expected width of $\tilde{C}(y)$ is:

$$\mathrm{E}[\tilde{W}|\theta] = \int_{\mathbb{R}} W(y)p(y|\theta)dy = \int_{\mathbb{R}}\int_{\mathbb{R}} (1-\tilde{\phi}_t(y))p(y|\theta)dydt \tag{2.20}$$

where $p(y|\theta)$ is the density of Y given θ. Similarly, the expected width of $C(y)$ is

$$\mathrm{E}[W|\theta] = \int_{\mathbb{R}}\int_{\mathbb{R}} (1-\phi_t(y))p(y|\theta)dydt. \tag{2.21}$$

Again, by Theorem 2 from Ferguson (1967, Section 5.3), for every $\theta \in \mathbb{R}$

$$\int_{\mathbb{R}} \phi_t(y)p(y|\theta)dy \geq \int_{\mathbb{R}} \tilde{\phi}(y)p(y|\theta)dy.$$

Thus

$$\int_{\mathbb{R}} (1-\phi_t(y))p(y|\theta)dy \leq \int_{\mathbb{R}} (1-\tilde{\phi}_t(y))p(y|\theta)dy. \tag{2.22}$$

Therefore by (2.20), (2.21), (2.22), we have $\mathrm{E}[W|\theta] \leq \mathrm{E}[\tilde{W}|\theta]$. □

Proof of Proposition 2.2.1. Without loss of generality, we prove the proposition for the simple case when $\mu = 0$ and $\sigma^2 = 1$. Other cases can be obtained by reparametrizing as $\tilde{Y} = (Y-\mu)/\sigma$, $\tilde{\theta} = (\theta-\mu)/\sigma$ and $\tilde{\tau}^2 = \tau^2/\sigma^2$ so that $\tilde{Y} \sim N(\tilde{\theta}, 1)$ and $\tilde{\theta} \sim N(0, \tilde{\tau}^2)$.

The Bayes optimal procedure minimizes the Bayes risk $R(\psi, C_w) = \int \Pr(Y \in A(\theta))\, d\theta$, where Y has the marginal density $N(0, 1+\tau^2)$. For a given w-function, the Bayes risk is

$$\begin{aligned} R(\psi, C_w) &= \int_{\mathbb{R}} \Phi(\frac{\theta - l}{\sqrt{1+\tau^2}}) - \Phi(\frac{\theta - u}{\sqrt{1+\tau^2}})\ d\theta \\ &= \int_{\mathbb{R}} \Phi(\frac{\theta - \Phi^{-1}(\alpha(1-w))}{\sqrt{1+\tau^2}}) - \Phi(\frac{\theta - \Phi^{-1}(1-\alpha w)}{\sqrt{1+\tau^2}})\ d\theta. \end{aligned} \tag{2.23}$$

We will show that, as a function of w, the integrand H is minimized at $w_\psi(\theta)$ as given in the proposition statement. First, we obtain the derivative of H with respect to w:

$$\begin{aligned} H'(w) =& \exp(-\frac{1}{2}\frac{(\theta - \Phi^{-1}(\alpha(1-w)))^2}{1+\tau^2})\frac{1}{\sqrt{1+\tau^2}}\frac{\alpha}{\exp(-\frac{1}{2}(\Phi^{-1}(\alpha(1-w)))^2)} \\ &- \exp(-\frac{1}{2}\frac{(\theta - \Phi^{-1}(1-\alpha w))^2}{1+\tau^2})\frac{1}{\sqrt{1+\tau^2}}\frac{\alpha}{\exp(-\frac{1}{2}(\Phi^{-1}(1-\alpha w))^2)}. \end{aligned}$$

Setting this to zero and simplifying shows that a critical point w_ψ satisfies

$$2\theta/\tau^2 = \Phi^{-1}(w\alpha) - \Phi^{-1}((1-w)\alpha). \tag{2.24}$$

Let the right side of (2.24) be $g(w)$. It's not difficult to verify that $g(w)$ a continuous and strictly increasing function of w, with range $(-\infty, \infty)$. Thus there is a unique solution $w_\psi(\theta)$ to the equation above, $w_\psi(\theta) = g^{-1}(2\theta/\tau^2)$, which is a continuous and strictly increasing function of θ. Since $H'(w)$ is continuous on $(0, 1)$ with only one root, and $\lim_{w\to 0} \ H'(w) = -\infty$, $\lim_{w\to 1} \ H'(w) = \infty$, then $H(w)$ is minimized by $w_\psi(\theta)$. Therefore $w_\psi(\theta)$ minimizes the Bayes risk, and C_{w_ψ} is the Bayes-optimal procedure among all CRPs. □

Proof of Lemma 2.2.2. $C_w(y)$ can be written as

$$C_w(y) = \{\theta : y < \theta - \sigma l(\theta) \text{ and } \theta - \sigma u(\theta) < y\}.$$

Letting $f_1(\theta) = \theta - \sigma u(\theta)$, $f_2(\theta) = \theta - \sigma l(\theta)$, we first prove that $C_w(y)$ can also be written as

$$C_w(y) = \{\theta : f_2^{-1}(y) < \theta \text{ and } \theta < f_1^{-1}(y)\}.$$

Note that both Φ^{-1} and $w(\theta)$ are continuous nondecreasing functions. Therefore $f_1(\theta) = \theta - \Phi^{-1}(1 - \alpha w(\theta))$ is a strictly increasing continuous function, with $\lim_{\theta\to-\infty} \theta - \Phi^{-1}(1 - \alpha w(\theta)) = -\infty$ and $\lim_{\theta\to+\infty} \theta - \Phi^{-1}(1 - \alpha w(\theta)) = +\infty$. Hence, f_1^{-1} exists, and is a strictly increasing continuous function with range $(-\infty, \infty)$. Thus $f_1(\theta) < y$ can also be expressed as $\theta < f_1^{-1}(y)$. Similarly, $y < f_2(\theta)$ can also be expressed as $f_2^{-1}(y) < \theta$. Next, in order to show that $C_w(y)$ is an interval, we need to show that $f_2^{-1}(y) < f_1^{-1}(y)$. To see this, we only need to show

$$\theta - \sigma\Phi^{-1}(1 - \alpha w(\theta)) < \theta - \sigma\Phi^{-1}(\alpha(1 - w(\theta))),$$

or that $\Phi^{-1}(\alpha w(\theta)) < \Phi^{-1}(1-\alpha(1-w(\theta)))$. This follows since $\Phi^{-1}(x)$ is a strictly increasing function. Thus $C_w(y) = \{\theta : f_2^{-1}(y) < \theta < f_1^{-1}(y)\}$, which is an interval. □

Proof of Lemma 2.2.3. The proof is basically the same as the proof of Lemma 2.2. We only need to replace y with $\bar{y}$, σ with $s/\sqrt{n}$, and the z-quantiles with t-quantiles, and then use the same logic as in the proof of Lemma 2.2. □

The proof of Proposition 2.3.1 requires the following lemma that bounds the width of the FAB t-interval:

Lemma 2.6.1. *The width of $C_{w_\psi}(\bar{y}, s^2)$ satisfies*

$$|C_{w_\psi}(\bar{y}, s^2)| < |\bar{y} - \mu| + \tfrac{s}{\sqrt{n}}(|t(\alpha/2)| + |t(1-\alpha/2)|), \tag{2.25}$$

where t-quantiles are those of the t_q-distribution.

Proof. For notational convenience, for this proof and the proof of Proposition 3.1, we write t_α as $t(\alpha)$. By previous results, the endpoints θ^L and θ^U of $C_{w_\psi}(\bar{y}, s^2)$ are solutions to

$$\begin{aligned} \theta^U - \tfrac{s}{\sqrt{n}} t(1 - \alpha w_\psi(\theta^U)) &= \bar{y} \\ \theta^L - \tfrac{s}{\sqrt{n}} t(\alpha(1 - w_\psi(\theta^L))) &= \bar{y}. \end{aligned} \tag{2.26}$$

Here $w_\psi(\theta)$ is defined as $w_\psi(\theta) = g^{-1}(\frac{2(\theta-\mu)}{\tau^2/\sigma})$, where $g(w) = \Phi^{-1}(\alpha w) - \Phi^{-1}(\alpha(1-w))$. At the upper endpoint, we have $w_\psi(\theta^U) = F((\bar{y} - \theta^U)/(s/\sqrt{n}))/\alpha$, where F is the CDF of the t_q-distribution. When $\theta^U > \mu$, we have $w_\psi(\theta^U) > g^{-1}(0) = 1/2$. Thus $\theta^U < \bar{y} - \frac{s}{\sqrt{n}} t(\alpha/2)$. Also, $g^{-1}(\frac{2(\theta^U-\mu)}{\tau'^2/\sigma}) < 1$, so $\theta^U > \bar{y} - \frac{s}{\sqrt{n}} t(\alpha)$. When $\theta^U < \mu$, $\bar{y} - \frac{s}{\sqrt{n}} t(\alpha/2) < \theta^U$. This implies that

$$\begin{aligned} &\bar{y} - \tfrac{s}{\sqrt{n}} t(\alpha) < \theta^U < \bar{y} - \tfrac{s}{\sqrt{n}} t(\alpha/2) && \text{if } \theta^U > \mu \\ &\bar{y} - \tfrac{s}{\sqrt{n}} t(\alpha/2) < \theta^U < \mu && \text{if } \theta^U < \mu. \end{aligned}$$

Similarly we have

$$\begin{aligned} &\mu < \theta^L < \bar{y} - \tfrac{s}{\sqrt{n}} t(1 - \alpha/2) && \text{if } \theta^L > \mu \\ &\bar{y} - \tfrac{s}{\sqrt{n}} t(1 - \alpha/2) < \theta^L < y - \tfrac{s}{\sqrt{n}} t(1 - \alpha) && \text{if } \theta^L < \mu. \end{aligned}$$

Therefore

$$|C_{w_\psi}(\bar{y}, s)| = \theta^U - \theta^L < |\bar{y} - \mu| + \tfrac{s}{\sqrt{n}}(|t(\alpha/2)| + |t(1-\alpha/2)|).$$

□

Proof of Proposition 2.3.1. That $C_{w_{\hat{\psi}}}$ is a $1 - \alpha$ CIP follows by construction of the interval and that $\hat{\psi}$ is independent of $\bar{Y}$ and S^2. To prove the convergence of the risk, we denote the

endpoints of the oracle CIP C_{w_ψ} as θ^U and θ^L, which are the solutions to

$$\theta^U - \tfrac{\sigma}{\sqrt{n}}\Phi^{-1}(1-\alpha w_\psi(\theta^U)) = \bar{Y}$$
$$\theta^L - \tfrac{\sigma}{\sqrt{n}}\Phi^{-1}(\alpha(1-w_\psi(\theta^L))) = \bar{Y}.$$

We denote the endpoints of $C_{w_{\hat{\psi}}}$ as θ_q^U and θ_q^L, which are the solutions to

$$\theta_q^U - \tfrac{S}{\sqrt{n}}t(1-\alpha w_{\hat{\psi}}(\theta_q^U)) = \bar{Y}$$
$$\theta_q^L - \tfrac{S}{\sqrt{n}}t(\alpha(1-w_{\hat{\psi}}(\theta_q^L))) = \bar{Y}.$$

We first prove that $|C_{w_{\hat{\psi}}}| - |C_{w_\psi}| = (\theta_q^U - \theta^U) + (\theta^L - \theta_q^L) \xrightarrow{a.s.} 0$ as $q \to \infty$ for each fixed $\bar{Y}$. We can write the upper endpoints as $\theta^U = G(\psi, \bar{Y}, \sigma^2)$, and $\theta_q^U = G_q(\hat{\psi}, \bar{Y}, S^2)$, where G and G_q are continuous functions of their parameters. The difference between G and G_q is that the former is obtained based on z-quantiles, while the later is based on t-quantiles. We have

$$|\theta_q^U - \theta^U| = |G_q(\hat{\psi}, \bar{Y}, S^2) - G(\psi, \bar{Y}, \sigma^2)| \tag{2.27}$$
$$\leq |G_q(\hat{\psi}, \bar{Y}, S^2) - G(\hat{\psi}, \bar{Y}, S^2)| + |G(\hat{\psi}, \bar{Y}, S^2) - G(\psi, \bar{Y}, \sigma^2)|. \tag{2.28}$$

The first term in (2.28) converges to zero because the convergence of $G_q \to G$ is uniform, and the second term converges to zero because $(\hat{\psi}, S^2) \xrightarrow{a.s.} (\psi, \sigma^2)$. Elaborating on the convergence of the first term, note that G_q is a monotone sequence of continuous functions: Given $q_2 > q_1$, we have $t_{q_2}(1-\alpha w) < t_{q_1}(1-\alpha w)$. Hence

$$\theta - \tfrac{S}{\sqrt{n}}t_{q_2}(1-\alpha w_{\hat{\psi}}(\theta)) > \theta - \tfrac{S}{\sqrt{n}}t_{q_1}(1-\alpha w_{\hat{\psi}}(\theta)).$$

Therefore $G_{q_2}(\hat{\psi}, \bar{Y}, S^2) < G_{q_1}(\hat{\psi}, \bar{Y}, S^2)$, and so by Dini's theorem, $G_q \to G$ uniformly on a compact set of $(\hat{\psi}, S^2)$ values. Since $(\hat{\psi}, S^2) \xrightarrow{a.s.} (\psi, \sigma^2)$, with probability one there is an integer Q such that when $q > Q$, $|\hat{\psi} - \psi| \leq c_1$ and $|S^2 - \sigma^2| \leq c_2$ for any to positive constants c_1 and c_2. Thus, G_q converges to G uniformly on this compact set and the first term in (2.28) converges to zero.

Now we show the expected width converges to the oracle width by integrating over $\bar{Y}$. This is done by finding a dominating function for $|C_{w_{\hat{\psi}}}(\bar{Y}, S^2)|$ and applying the dominated

convergence theorem. By the previous lemma we know that

$$|C_{w_{\hat{\psi}}}(\bar{Y}, S^2)| < |\bar{Y}| + |\hat{\mu}| + \tfrac{S}{\sqrt{n}}(|t(\alpha/2)| + |t(1-\alpha/2)|).$$

Note that

$$|t(\alpha/2)| + |t(1-\alpha/2)| < |t_1(\alpha/2)| + |t_1(1-\alpha/2)|,$$

where t_1 is the t-quantile with one degree of freedom. Similar to the argument earlier in this proof, given two constants $c_1, c_2 > 0$, we can find a Q such that when $q > Q$, we have $|\hat{\mu}| < |\mu| + c_1$ and $S^2 < \sigma^2 + c_2$ a.s.. Now we have an dominating function for $|C_{w_{\hat{\psi}}}(\bar{Y}, S^2)|$

$$|C_{w_{\hat{\psi}}}(\bar{Y}, S^2)| < \bar{W}(\bar{Y}, S^2, \hat{\psi}) = |\bar{Y}| + |\mu| + c_1 + \tfrac{\sqrt{\sigma^2+c_2}}{\sqrt{n}}(|t_1(\alpha/2)| + |t_1(1-\alpha/2)|).$$

Since $|\bar{Y}|$ is a folded normal random variable with finite mean, it's easy to see that this dominating function is integrable. Therefore, by dominated convergence theorem we have $\lim_{q\to\infty} \mathrm{E}[|C_{w_{\hat{\psi}}}|] = \mathrm{E}[|C_{w_{\psi}}|]$.

□

Proof of Lemma 2.5.1 . We look at the worst situation for $C^{\tau_A^2}$ in this case, i.e. we identify the τ_A^2 value that maximizes the Bayes risk of $C^{\tau_A^2}$ with respect to π_B. When $\theta > 0$, $C_{w_\pi} = g^{-1}(\frac{2\theta}{\tau_A^2}) \in (\frac{1}{2}, 1)$. By the proof of Prop. 2.2.1, $H(w)$ is maximized when $w = 1$, which corresponds to the limiting procedure when $\tau_A^2 \to 0$. Similarly, when $\theta < 0$, $H(w)$ is also maximized when $\tau_A^2 \to 0$. Thus when $\tau_A^2 \to 0$, the Bayes risk is maximized. Therefore,

$$\begin{aligned} R(\pi_B, C^{\tau_A^2}) &< \int_{-\infty}^{0} \Phi(\frac{\theta + \Phi^{-1}(1-\alpha)}{\sqrt{1+\tau_B^2}})d\theta + \int_{0}^{+\infty} 1 - \Phi(\frac{\theta - \Phi^{-1}(1-\alpha)}{\sqrt{1+\tau_B^2}})d\theta \\ &= 2\int_{0}^{+\infty} \Phi(\frac{-\theta + \Phi^{-1}(1-\alpha)}{\sqrt{1+\tau_B^2}})d\theta \end{aligned}$$

This is an integral form for Gaussian function, by the well-known equality:

$$\int \Phi(a+bx)dx = b^{-1}((a+bx)\Phi(a+bx) + \phi(a+bx)) + C.$$

we can get rid of the integral,

$$R(\pi_B, C^{\tau_A^2}) < 2\Phi^{-1}(1-\alpha)\Phi(\frac{\Phi^{-1}(1-\alpha)}{\sqrt{1+\tau_B^2}}) + 2\sqrt{1+\tau_B^2}\phi(\frac{\Phi^{-1}(1-\alpha)}{\sqrt{1+\tau_B^2}}). \tag{2.29}$$

This provides a upper bound of the Bayes risk with respect to $\theta \sim N(0, \tau_B^2)$ for $C^{\tau_A^2}$ constructed using any τ_A^2 values. □

Proof of Lemma 2.5.2 . We use $R(\pi, C)$ to represent the Bayes risk of procedure C, and use $R(\theta, C)$ to represent the frequentist risk of procedure C at θ.

$$\begin{aligned} R(\pi, C) &= \int_\Omega R(\theta, C)\pi(\theta)d\theta \\ &= \int_\Omega R(\theta, C) \int_\Lambda f(\theta|\lambda)g(d\lambda)d\theta \\ &= \int_\Lambda \Big(\int_\Omega R(\theta, C) f(\theta|\lambda)d\theta \Big) g(d\lambda) \\ &= \int_\Lambda \Big(R(f(\theta|\lambda), C) \Big) g(d\lambda). \end{aligned}$$

Therefore if we have

$$R(f(\theta|\lambda), C_1) \leq R(f(\theta|\lambda), C_2)$$

for all $f(\theta|\lambda)$, then

$$R(\pi, C_1) \leq R(\pi, C_2).$$

□

Chapter 3

ADAPTIVE SIGN ERROR CONTROL

3.1 Introduction

We consider multiparameter inference for the normal means model,

$$\boldsymbol{Y}|\boldsymbol{\theta} \sim N(\boldsymbol{\theta}, \boldsymbol{I}), \tag{3.1}$$

where $\boldsymbol{Y} = (Y_1, \ldots, Y_m)$ and $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_m)$. Simultaneous inference for $\theta_1, \ldots, \theta_m$ often begins by testing $H_i : \theta_i = 0$ for each $i = 1, \ldots, m$ at level α, that is, we reject H_i if $|Y_i|$ exceeds the $1 - \alpha/2$ standard normal quantile, $z_{1-\alpha/2}$. This controls the experimentwise type I error rate to be equal to α. A popular method for choosing α is the Benjamini Hochberg (BH) procedure (Benjamini and Hochberg, 1995). The BH procedure is an adaptive method for selecting a value of α that will bound the false discovery rate (FDR), which is defined as FDR $= E[\frac{V}{R \vee 1}|\theta_1, ..., \theta_p]$, where R is the number of rejections and V is the number of false rejections, that is, the number of null hypotheses that are rejected but true. There is a large literature on FDR control, see Efron (2012), Benjamini (2010), Genovese and Wasserman (2004), Storey (2002) and Storey (2007). However, in many applications it is likely that none of the θ_i's are truly equal to exactly zero. For example, in the case where each Y_i represents a difference in sample averages between two treatments, Tukey (1991) argued that evaluating if $\theta_i = 0$ is "foolish" since the effects of two different factors are always different, however minutely. In such cases, Tukey (1962) suggests that a more meaningful task is to judge whether or not there is enough evidence to infer the sign of θ_i, instead of whether or not it is zero. However, if significance tests are used in this way, then FDR control is inappropriate since it is always zero if there are no true nulls. Instead, the relevant error control is not the FDR, but a sign error rate (Gelman and Tuerlinckx, 2000; Gelman and Carlin, 2014; Owen,

2016).

Benjamini and Yekutieli (2005) showed that the Benjamini-Hochberg algorithm can be used to control the pure directional FDR, defined as the expected proportion of discoveries in which a positive parameter is declared negative or a negative parameter is declared positive. We refer to this procedure as the BY procedure in this chapter. Some follow-up work includes Zhao et al. (2015) who used weighted p-value methods, and Guo et al. (2010) who extended the idea to making multidimensional directional decisions. Weinstein et al. (2013) derived new selection-adjusted confidence intervals by minimizing an objective function comprised of the length of the acceptance region and a penalty term for the magnitude of the observation. They showed in examples that these procedures have correct coverage on selected parameters, and have more power to determine the sign, but they did not assess the sign error rate directly. These procedures also do not utilize information across experiments and so are not adaptive. Stephens (2016) proposed an empirical Bayes procedure for sign error control to gain more power. However, the focus there was control of the local sign error instead of the sign error rate across experiments.

In the next section, we discuss the distribution of the sign error proportion (SEP) under a hierarchical model for the Y_i's and θ_i's, and relate this to a marginal sign error rate (MSER). We then propose an adaptive nonparametric procedure that controls the MSER below a desired threshold regardless of the distribution of the θ_i's. This procedure is more powerful than BY procedure in terms of the number of rejections made, and therefore in terms of the number of signs inferred. The power can be further improved if one is willing to assume a parametric model for the distribution of the θ_i's. We show that a model-based approach to MSER control can achieve an optimal power asymptotically, if a model for the θ_i's is chosen correctly. In Section 3.3, we numerically compare the nonparametric procedure and parametric procedures to the BY procedure and an oracle MSER control procedure in a simulation study. In Section 3.4, we discuss an adaptive procedure for the somewhat different task of sign inference subject to fixed experimentwise type I error rate. We show how the acceptance region of a level-α test of each H_i may be adaptively chosen to minimize the

MSER or maximize the power, that is, the number of sign discoveries. A discussion follows in Section 3.5. Proofs are given in Appendix.

3.2 Sign Error Rate Control Procedures

3.2.1 Marginal Sign Error Rate

We are interested in inferring the sign of each θ_i in the normal means model in (3.1). We test $H_i : \theta_i = 0$ using the usual level-α z-test, and estimate $sign(\theta_i)$ by $sign(Y_i)$ if the test rejects and do not estimate the sign otherwise. We use the pair (R_i, S_i) to denote the outcome of this procedure, where $R_i = 1$ if H_i is rejected, and $R_i = 0$ otherwise. We use S_i to denote the sign estimate, with possible values 1 (positive), -1 (negative), and 0 (sign not estimated). Note that $S_i = 0$ if $R_i = 0$. A sign error is made if $S_i \cdot sign(\theta_i) = -1$. Let E_i be the binary indicator of a sign error, so that $E_i = R_i(1 - S_i \cdot sign(\theta_i))/2$. The results across experiments are summarized with (R, E), where $R = \sum_{i=1}^{m} R_i$ is the total number of rejections and $E = \sum_{i=1}^{m} E_i$ is the total number of sign errors among the m experiments. In what follows, we assume that none of the θ_i's are truly equal to zero. The properties of our procedures in cases where there are some true nulls are discussed in Section 3.4.

Define the sign error proportion as

$$\text{SEP} = E/(R \vee 1). \tag{3.2}$$

Ideally, we want to keep SEP under a desired threshold. Given a data vector $\boldsymbol{Y}$ and a experimentwise significance threshold, the number of rejections R is known but the number of sign errors E is unknown since each E_i depends on the unknown true parameter θ_i. Therefore, SEP is an unobserved quantity that depends on the data and the unobserved parameter values. However, suppose the empirical distribution of $\theta_1, \ldots, \theta_m$ is well-represented by some distribution G, absolutely continuous with respect to Lebesgue measure (and so $G(\{0\}) = 0$). We then assume the following model:

$$\theta_1, \ldots, \theta_m \ \sim \ \text{i.i.d. } G. \tag{3.3}$$

Now (3.1) and (3.3) specify a hierarchical model. Under this hierarchical model, the probability of making a sign error for any one experiment, conditional on rejection, can be written as

$$\text{MSER} = \Pr(E_1 = 1 | R_1 = 1) = \frac{\Pr(E_1 = 1, R_1 = 1)}{\Pr(R_1 = 1)}. \tag{3.4}$$

We call the quantity in (3.4) the *marginal sign error rate* (MSER). This quantity does not depend on $\boldsymbol{Y}$ or $\boldsymbol{\theta}$, just on G and α. It also determines the marginal distribution of the SEP:

Lemma 3.2.1. *Under the hierarchical model (3.1) and (3.3), the conditional distribution of $R \cdot SEP$ given $R = r$ is binomial$(r, MSER)$.*

From this lemma it follows that $\text{E[SEP]} = \text{MSER} \cdot \Pr(R > 0) < \text{MSER}$. Thus by controlling MSER to be below a threshold, we bound the expected SEP under this threshold as well. Moreover, by the following Proposition, in scenarios where m is large, controlling MSER gives an accurate control over SEP.

Proposition 3.2.1. *Under the hierarchical model (3.1) and (3.3), SEP converges to MSER in probability as $m \to \infty$.*

In the following subsections, we propose two methods to control the MSER under a prespecified level α_S. The first method is called the *loose control* procedure, which conservatively controls MSER without parametric assumptions. The second method is called tight control, which estimates the distribution of the θ_i's and adaptively chooses an experiment-wise type I error rate α to maximize the number of signs estimated while controlling MSER approximately below level α_S.

3.2.2 Loose Control Procedure

In this subsection, we develop a procedure that conservatively controls MSER. It has a good performance in "spike and slab" scenarios where the sizes of most of the θ_i's are negligible compared to the measurement error, with only a few θ_i's having large values. However, for

other distributions of the θ_i's it can have an MSER substantially below the nominal level, and so we call it the loose control procedure.

The intuition for the loose control procedure is as follows: MSER can be seen as the expected number of sign errors divided by the expected number of signs inferred. With an type I error rate of α, in the extreme case where all the θ_i's are very close to zero, we expect to infer around $\alpha \cdot m$ signs, and expect half of them to be sign errors. Hence the expected number of sign errors will be approximately $\alpha \cdot m/2$. On the other hand, the number of signs we infer is R. Thus intuitively we want $(\alpha m/2)/R$ to be smaller than α_S, which suggests the following procedure:

1. Find the largest α_l such that $\alpha_l \leq 2\alpha_S R(\alpha_l)/m$.

2. Infer the sign for ith experiment if $|Y_i| > z_{1-\alpha_l/2}$.

Here, $R(\alpha_l)$ is the number of rejections made if the rejection threshold is $z_{1-\alpha_l/2}$. We call this procedure the loose control procedure (LC). It controls MSER asymptotically in m:

Proposition 3.2.2. *For the hierarchical model in (3.1) and (3.3) and using the LC procedure, MSER* $\leq \alpha_S + \epsilon$ *with* $\epsilon \to 0$ *in probability as* $m \to \infty$.

This procedure does not provide guaranteed control of MSER for finite m because in particular the significance threshold for each experiment i depends to some extent on Y_i through $R(\alpha_l)$. For small m we suggest using the following procedure that gives exact, non-asymptotic control of MSER:

1. For each experiment i, find the largest α_l^i such that $\alpha_l^i \leq 2\alpha_S((R^{-i}(\alpha_l^i) - 1) \vee 0)/m$.

2. Infer the sign for the ith experiment if $|Y_i| > z_{1-\alpha_l^i/2}$.

Here, $R^{-i}(\alpha_l^i)$ is the number of rejections made among all experiments except experiment i if the significance threshold is $z_{1-\alpha_l^i/2}$. This procedure is slightly more conservative than LC procedure since any α_l^i also satisfies $\alpha_l^i \leq 2\alpha_S R(\alpha_l^i)/m$. We call this procedure the non-asymptotic loose control (NLC) procedure.

Proposition 3.2.3. *For the hierarchical model in (3.1) and (3.3) and using the non-asymptotic loose control procedure, MSER* $\leq \alpha_S$.

These loose control procedures are closely related to the Benjamini Yekutieli (BY) (Benjamini and Yekutieli, 2005) procedure, which is equivalent to finding the maximal α_{by} such that $\alpha_{by} \leq \alpha_S R(\alpha_{by})/m$. It is easy to see that α_{by} is always smaller than α_l. Hence the LC procedure always infers more signs than the BY procedure. The BY procedure was proposed for controlling the unconditional sign error rate SER $= \mathrm{E}[\mathrm{SEP}|\boldsymbol{\theta}]$, which they called the "pure directional FDR". In the case that there are no true nulls, the loose control procedure also controls SER:

Proposition 3.2.4. *Under model (3.1), if* $\theta_i \neq 0$ *for all* $i \in \{1, \ldots, m\}$ *then both the LC and NLC procedures control the SER below* α_S.

3.2.3 Model Based Control Procedure

Although the loose control procedure controls MSER without assumptions on G, it can be conservative in cases where G does not resemble a spike and slab distribution. In this subsection, we propose a model-based MSER control procedure that can be more powerful in terms of the number of sign inferred.

We first discuss the oracle situation where the probability density function G of θ_i's is known. The acceptance region of our test of H_i is $A(\alpha) = \{Y_i : \Phi^{-1}(\alpha/2) < Y_i < \Phi^{-1}(1-\alpha/2)\}$, with Φ being the standard normal cumulative density function. We can write MSER as a function of α as follows:

$$\mathrm{MSER}(\alpha) = \frac{\Pr(E_1 = 1, R_1 = 1)}{\Pr(R_1 = 1)} \tag{3.5}$$

$$= \frac{\mathrm{E}_G[\mathrm{P}_{\theta_1}(E_1 = 1, R_1 = 1)]}{\mathrm{E}_G[\mathrm{P}_{\theta_1}(R_1 = 1)]} \tag{3.6}$$

$$= \frac{\mathrm{E}_G[\mathrm{P}_{\theta_1}(S_1 = -1, R_1 = 1, sign(\theta_1) = 1) + \mathrm{P}_{\theta_1}(S_1 = 1, R_1 = 1, sign(\theta_1) = -1)]}{\mathrm{E}_G[\mathrm{P}_{\theta_1}(Y_1 \notin A(\alpha))]} \tag{3.7}$$

$$= \frac{\mathrm{E}_G[\mathrm{P}_{\theta_1}(Y_1 < 0, Y_1 \notin A(\alpha))\mathbf{1}(\theta_1 > 0) + \mathrm{P}_{\theta_1}(Y_1 > 0, Y_1 \notin A(\alpha))\mathbf{1}(\theta_1 < 0)]}{\mathrm{E}_G[\mathrm{P}_{\theta_1}(Y_1 \notin A(\alpha))]} \quad (3.8)$$

$$= \frac{\mathrm{E}_G[B_1\mathbf{1}(\theta_1 > 0) + B_2\mathbf{1}(\theta_1 < 0)]}{\mathrm{E}_G[B_1 + B_2]}, \quad (3.9)$$

where $B_1 = \Phi(\Phi^{-1}(\alpha/2) - \theta)$ and $B_2 = \Phi(\Phi^{-1}(\alpha/2) + \theta)$. In this case, we need to find the value of α such that $\mathrm{MSER}(\alpha) = \alpha_S$. We denote this α as α_o, and call the resulting procedure the tight control oracle (TCO) procedure. This procedure maximizes the power in inferring signs while keeping MSER at α_S.

In practice, G is unknown and must be estimated from the data. Suppose we have an estimate $\hat{G}$ of G. By replacing G by $\hat{G}$ in (3.9) we can obtain an empirical estimate $\widehat{\mathrm{MSER}}$ for each value of α, and in particular, find an α_e such that $\widehat{\mathrm{MSER}}(\alpha_e) = \alpha_S$. We call the procedure using α_e instead of α_o the tight control empirical (TCE) procedure.

The task of estimating G from $\boldsymbol{Y}$ based on (3.1) and (3.3) is known as deconvolution. Current nonparametric deconvolution techniques are computationally expensive, and converge to the true G slowly in m, yielding unstable results for small m. As an alternative to nonparametric deconvolution, we propose using simple parametric models to facilitate the application of the TCE procedure. The following proposition shows that under certain assumptions, the TCE procedure converges to the optimal TCO procedure when a correct parametric model for the θ_i's is used.

Proposition 3.2.5. *Suppose $\theta_1, \ldots, \theta_m \sim$ i.i.d. G_η where G_η is a member of a parametric family of distributions indexed by a finite-dimensional parameter vector η with density function continuous in η. For each m let $\hat{\eta}$ be an estimate of η, and let $\widehat{MSER}(\alpha)$ be the plug-in estimate of $MSER(\alpha)$ calculated using $G_{\hat{\eta}}$. If $\hat{\eta} \xrightarrow{p} \eta$ as $m \to \infty$, then $\widehat{MSER}(\alpha) \xrightarrow{p} MSER(\alpha)$ and $\alpha_e \xrightarrow{p} \alpha_o$ as $m \to \infty$.*

One useful model for G that we explore in the next section is the family of asymmetric Laplace distributions (Yu and Zhang, 2005), which have probability density functions of the form

$$g(\theta; \mu, \tau, q) = \frac{q(1-q)}{\tau} \exp\big(-\frac{(x-\mu)}{\tau}[q - I(x \le \mu)]\big),$$

where $\mu \in \mathbb{R}$ is the location parameter, $\tau > 0$ is the scale parameter, and $0 < q < 1$ is the skew parameter. Figure 3.1 shows the shape of ALD distributions for $q \in \{0.1, 0.3, 0.5\}$.

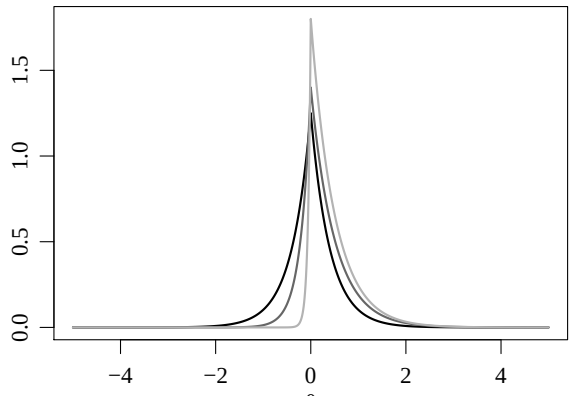

Figure 3.1: Shapes of asymmetric Laplace densities. The black line is the ALD density when $q = 0.5$ and $\tau = 0.2$, the darker grey line is for $q = 0.3$ and $\tau = 0.15$, and the lightest grey line is for $q = 0.1$ and $\tau = 0.05$.

The asymmetric Laplace distribution is a flexible model for unimodal distributions with the Laplace distribution being a special case. It is more peaked at zero than a normal distribution, but also can reflect the potential skewness of the distribution of true effects that often exists in applications, for example, in cases where more θ_i's are positive than negative, or vice versa. For multiple testing problems were we expect that most θ_i's are close to zero, it is natural to consider only submodels where $\mu = 0$. In this case, method of moment estimates for the scale and skew parameters may be obtained from the first and second sample moments of $\boldsymbol{Y}$. Under the hierarchical model, we have

$$\mathrm{E}[Y] = \mathrm{E}[\mathrm{E}[Y|\theta]] = \mathrm{E}[\theta] = \frac{\tau(1-2q)}{q(1-q)},$$

$$\mathrm{Var}[Y] = \mathrm{E}[\mathrm{Var}[Y|\theta]] + \mathrm{Var}[\mathrm{E}[Y|\theta]] = 1 + \mathrm{Var}[\theta] = 1 + \frac{\tau^2(1-2q+2q^2)}{(1-q)^2q^2}.$$

By setting

$$\frac{1}{m}\sum Y_i = \frac{\tau(1-2\hat{q})}{\hat{q}(1-\hat{q})},$$
$$\frac{1}{m-1}\sum (Y_i - \bar{y}) = 1 + \frac{\hat{\tau}^2(1-2\hat{q}+2\hat{q}^2)}{(1-\hat{q})^2\hat{q}^2},$$

we can solve for $\hat{q}$ and $\hat{\tau}$ to obtain moment-based estimates of q and τ.

3.3 Simulation Studies

In this section we use several simulation scenarios to compare the performance of Benjamini and Yekutieli's procedure (BY), the loose control procedure (LC), and a tight control empirical procedure using an asymmetric Laplace model for the θ_i's (TCEA). For each simulation scenario, 1000 datasets were simulated as follows: First, values $\theta_1, \ldots, \theta_m$ were independently simulated from a distribution G, including ALD and "spike and slab" distribution with different parameter settings. Then an observation vector $\boldsymbol{Y}$ was sampled from a $N(\boldsymbol{\theta}, \boldsymbol{I})$ distribution. For each of these datasets, the sign error proportions and the total numbers of signs inferred by each procedure were calculated. For all procedures and simulation scenarios the target level α_S was set to be 10%. Simulations were run for $q \in \{0.1, 0.3, 0.5\}$ and for five different values of τ for each level of q. The ranges of the τ values were chosen so that SEP ranged between 10% to 30% when the experimentwise type I error rate $\alpha = 0.05$.

The results for several simulation scenarios with $m = 5000$ are summarized in Figure 3.2. Overall, the TCEA procedure performs nearly as well as the TCO procedure. Both procedures control SEP at the prespecified level $\alpha_S = 0.1$, and infer many more signs than the BY and LC procedures, with BY being the least powerful of the three. The difference between TCEA and LC or BY becomes larger as τ increases.

When number of experiments is large, the TCEA procedure is very close to the TCO procedure as our asymptotic result predicts. However, when $m = 100$, TCEA and TCO show some differences. The results for several simulations with $m = 100$ are summarized in Figure 3.3. In this situation, TCEA still performs better than BY or LC in terms of the power to infer signs. Also, we see that for some cases, the SEP of the oracle procedure does

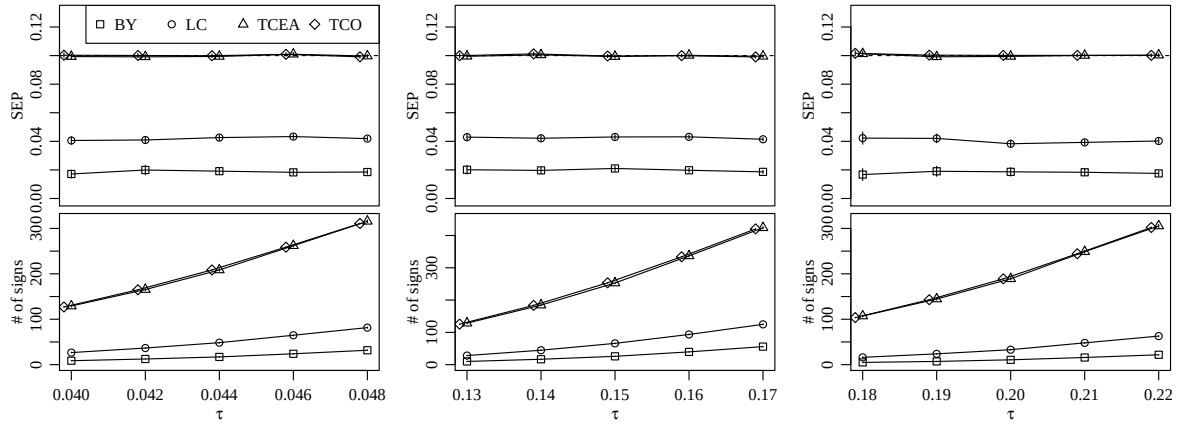

Figure 3.2: Comparison of the three procedures when $m = 5000$ and the θ_i's have an asymmetric Laplace distribution. The skewness parameter q is set to be 0.1 in the left column, 0.3 in the middle column, and 0.5 in the right column. Vertical bars around each plotting character correspond to ± 1.96 Monte Carlo standard errors.

not attain the nominal level of 0.1. This is because tight control procedure is designed to keep MSER under the nominal level α_s. As illustrated before, controlling MSER under α_S gives an accurate control over the expected SEP when m is large. When m is small, the probability of making no rejections across all experiments is non-negligible, and MSER is slightly larger than expectation of SEP. In this case, instead of keeping the average SEP at α_S, TCO keeps it under α_S, making the result slightly conservative.

Finally, we study the situation when G is a spike and slab distribution. The spike is a unimodal distribution with mean zero and small variance, and the slab is a uniform distribution. For two asymmetric cases ($q \in \{0.1, 0.3\}$) the slab is the uniform distribution on (2,4). For the symmetric case ($q = 0.5$), the slab is the uniform distribution on $(-4, -2) \cup (2, 4)$. In each case, the proportion of θ_i's that are sampled from the slab is 1%. Comparisons of the three procedures and TCO are summarized in Figure 3.4. As expected, the LC procedure overall has better performance than the BY and TCEA procedures. As the variance of the spike grows larger, the differences between the θ-values sampled from the spike and the

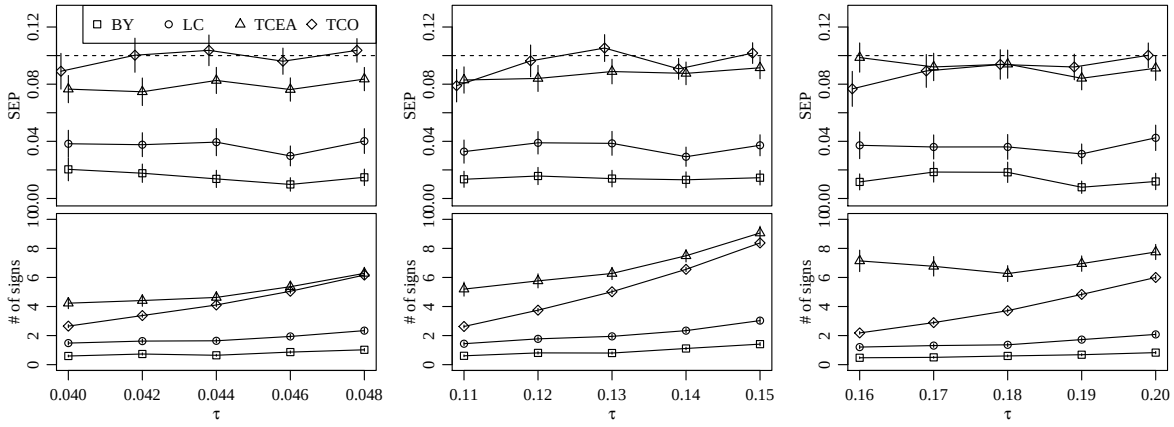

Figure 3.3: Comparison of the four procedures under the same settings as in Figure 3.2 but $m = 100$.

θ-values sampled from the slab becomes smaller, and the multimodal spike and slab distribution becomes closer and closer to a unimodal distribution that can be well-represented by a member of the asymmetric Laplace family. In such scenarios, TCEA does well in terms of maintaining MSER and inferring signs.

3.4 MSER and MSDR Optimization Subject to Type I Error Control

We have discussed controlling MSER under a prespecified level by choosing an appropriate significance threshold. In this section, we study the relationship between MSER and the shape of the acceptance region when the level α for the experimentwise type I error rate is held fixed. We show how to minimize the MSER while maintaining the experimentwise type I error rate. Storey (2007) has proposed a general framework for maximizing the statistical power of a test while maintaining the experimentwise type I error rate. Wasserman and Roeder (2006) and Dobriban et al. (2015) studied a weighted Bonferroni method to control family-wise type I error rate while maximizing the power. As illustrated in Gelman and Carlin (2014) and Owen (2016), a high sign error rate occurs when the error variance is large compared to the true effect size. We show that other than the error variance, the shape of

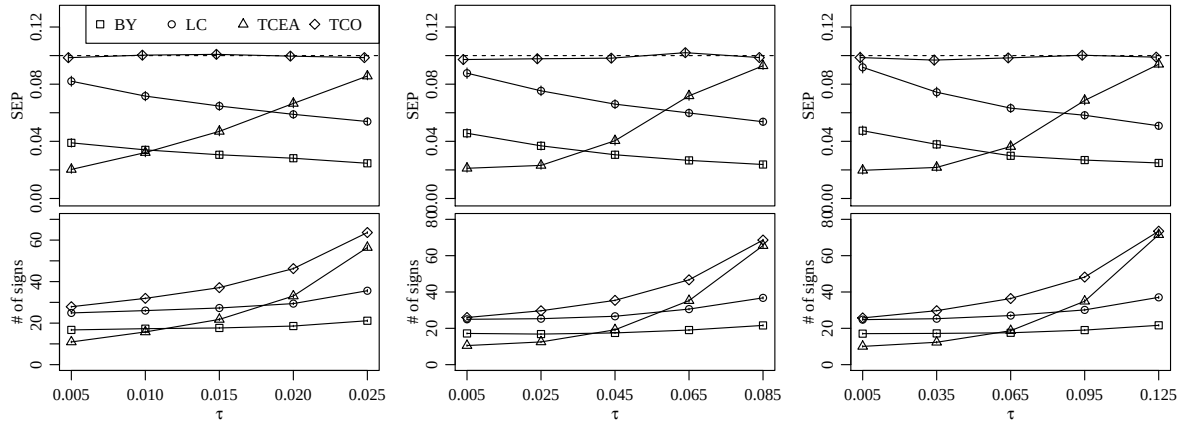

Figure 3.4: Comparisons of the three procedures when $m = 5000$ and under a spike and slab distribution for the θ_i's. From left to right, the spike is sampled from an asymmetric Laplace distribution with $q = 0.1, 0.3, 0.5$, respectively.

the acceptance region is another crucial factor in determining the sign error rate.

In addition to MSER, we define the Marginal Sign Discovery Rate (MSDR) as MSDR $=$ $\Pr(R_1 = 1)$. This quantity measures the expected proportion of the number of experiments with a sign inferred among all of the experiments since

$$\begin{aligned}\text{MSDR} = \Pr(R_1 = 1) = \frac{\sum_{i=1}^{m} \Pr(R_i = 1)}{m} = \frac{\sum_{i=1}^{m} \mathrm{E}[\mathbf{1}(R_i = 1)]}{m} \\ = \frac{\mathrm{E}[\sum_{i=1}^{m} \mathbf{1}(R_i = 1)]}{m} = \mathrm{E}\left[\frac{R}{m}\right].\end{aligned}$$

Both MSER and MSDR are affected by the acceptance region of the test. The usual acceptance region for each H_i is $A = (\Phi^{-1}(\alpha/2), \Phi^{-1}(1 - \alpha/2))$, which corresponds to the uniformly most accurate unbiased (UMAU) test. Following the ideas of Yu and Hoff (2018), we can construct a class of acceptance regions that corresponds to all level α two-sided tests $A(\alpha, s) = \{Y_i : \Phi^{-1}(\alpha s) < Y_i < \Phi^{-1}(1 - \alpha(1 - s))\}$, where $s \in (0, 1)$ is a constant. Thus even if the level α is fixed, we can change the acceptance region by varying its endpoints. When $s < 1/2$, the acceptance region tends to cover more negative observations and less positive observations. When $s > 1/2$, the acceptance region tends to cover more positive

observations and less negative observations. As $s \to 0$ or 1, the two-sided test converges to a one-sided test with an acceptance region of either $(\Phi^{-1}(\alpha), \infty)$ or $(-\infty, \Phi^{-1}(1-\alpha))$. We now examine which s value minimizes MSER and which s value maximizes MSDR when the experimentwise type I error rate α is held fixed. Similar to (3.9), we can express the MSER as

$$\text{MSER}(A(\alpha, s)) = \frac{E[B_1 \mathbf{1}(\theta_1 > 0) + B_2 \mathbf{1}(\theta_1 < 0)]}{E[B_1 + B_2]},$$
$$\text{MSDR}(A(\alpha, s)) = E[B_1 + B_2],$$

where $B_1 = \Phi(\Phi^{-1}(\alpha s) - \theta)$ and $B_2 = \Phi(\Phi^{-1}(\alpha(1-s)) + \theta)$.

If we fix α, MSER and MSDR can be seen as function of s. Under our models, we turn the minimization of MSER and maximization of MSDR into two one-parameter optimization problems: Denote

$$s^D = \arg\max_s \text{MSDR}(s)$$
$$s^E = \arg\min_s \text{MSER}(s).$$

Interestingly the UMAU procedure, where $s = s^U = 1/2$, does not always maximize the expected power, and the s that maximizes the MSDR does not necessarily minimizes the MSER, vice-versa. We use a simple numerical example to illustrate this. Suppose θ_i's are sampled from a shifted chi-square distribution $\chi^2_3 - 3$. By numerical evaluation, the results are summarized in Table 3.1.

	s value	$A(s, 0.05)$	MSER(%)	MSDR
s^U	0.5	(-3.92, 3.92)	3.01	0.189
s^D	0.683	(-3.65, 4.30)	2.79	0.193
s^E	0.829	(-3.45, 4.80)	2.71	0.190

Table 3.1: Comparison of the usual acceptance region, the acceptance region that maximizes MSDR, and the acceptance region that minimizes MSER

On the other hand, Storey (2007) noticed that when $\theta \sim N(0, \sigma_b^2)$, the test that maximizes expected power is the UMAU test. Here we prove a more general theorem that the UMAU test actually both maximizes expected power and minimizes MSER when the distribution of θ is symmetric.

Proposition 3.4.1. *If G is a distribution that is symmetric with respect to 0, the two-sided test that maximizes MSDR and minimizes the MSER is the UMAU test, i.e. $s^D = s^E = 1/2$.*

Thus in applications where α is held fixed, if we believe that the distribution of the θ_i's is symmetric, we should use the usual acceptance region. In situations where we suspect this distribution to be asymmetric, then using either S^D or S^E can lead to a test with either higher MSDR or lower MSER. However, identifying S^D or S^E requires G to be known. Similar to the TCE procedure, in practice we replace G with an estimate $\hat{G}$ and obtain empirical estimates $\widehat{\text{MSDR}}$ and $\widehat{\text{MSER}}$, and then obtain S^D or S^E by maximizing $\widehat{\text{MSDR}}$ or minimizing $\widehat{\text{MSER}}$.

3.5 Discussion

In this project, we use the MSER as a measure of sign errors in multiple testing settings. We proposed two types of procedures to control MSER, loose control procedure and tight control procedure. Loose control procedure can be conservative but is robust to the distribution of the θ_i's, while the tight control procedure is more powerful but assumes the distribution of θ_i's is a member of a known parametric model.

The loose control procedure proposed in this chapter is closely related to the BY procedure. Unlike the derivation for the BY procedure, we derive the LC procedure from the perspective of controlling the MSER, which is a quantity measuring the probability of making a sign error under a hierarchical model. We assume that there are no "true nulls" in this chapter, because in many applications true nulls do not exist. By assuming no true nulls, the loose control procedure we derived is more powerful than the BY procedure in terms of

the number of inferred signs. If it is believed that the true nulls do exist, the loose control procedure can still control the SER, although control over MSER depends on how we define a sign error when $\theta_i = 0$. If we define that when $\theta_i = 0$, either claiming θ is positive or negative is correct, the loose control procedure stays the same as proposed in this chapter. If we define that when $\theta_i = 0$, either claiming θ is positive or negative is wrong, then the BY procedure should be used since it also controls the mixed directional FDR, where any sign declaration of $\theta_i = 0$ is considered as a sign error.

We also discussed varying the endpoints of the acceptance region to reduce MSER and increase MSDR when the type I error rate is fixed. This can be combined with the tight control procedure, leading to a new procedure: Choose α and s such that

$$(\alpha, s) = \arg\max_{(\alpha,s)} \widehat{MSDR}(A(\alpha, s))$$
$$\text{such that} \quad \widehat{\text{MSER}}(A(\alpha, s)) < \alpha_S.$$

Given an estimate $\hat{G}$ of G, the solution for (α, s) can be obtained numerically. This procedure can potentially increase the power in inferring signs. However, the performance of this procedure is more unstable since the optimization task here is more complicated.

Appendix: Proofs

Proof of Lemma 3.2.1. Note that $(Y_1, \theta_1), \ldots, (Y_m, \theta_m)$ are an i.i.d sample from the hierarchical model (3.1) and (3.3). For H_i, $\forall i \in \{1, ..., m\}$, given that it is rejected, the probability of making a sign error is $\Pr(E_i = 1 | R_i = 1)$, which is MSER as specified in (3.4). Given that $R = r$ hypotheses are rejected, the total number of sign errors should follow a binomial distribution, i.e. $E\,|R = r \sim Bi(r, \text{MSER})$. Thus $R \cdot \text{SEP}\,|R = r \sim Bi(r, \text{MSER})$. □

Proof of Proposition 3.2.1. We just need to show that SEP − MSER $\to 0$ in probability, which is to show SEP−E[SEP]+E[SEP]−MSER $\to 0$ in probability. Since E[SEP] = MSER$\cdot$ $\Pr(R > 0) = \text{MSER} \cdot (1 - \Pr(R_1 = 0)^m)$, we have E[SEP] $\to$ MSER in probability as $m \to \infty$ (note $\Pr(R_1 = 0) < 1$ in our setting). Now we just need to show that SEP − E[SEP] $\to 0$ in

probability, which can be done by showing $\mathrm{E}[(\mathrm{SEP} - \mathrm{E}[\mathrm{SEP}])^2] \to 0$. We have

$$\begin{aligned}\mathrm{E}[(\mathrm{SEP} - \mathrm{E}[\mathrm{SEP}])^2] &= \mathrm{Var}[\mathrm{SEP}] = \mathrm{Var}[\mathrm{E}[\mathrm{SEP}|R]] + \mathrm{E}[\mathrm{Var}[\mathrm{SEP}|R]] \\ &= \mathrm{Var}[\mathrm{MSER} \cdot \mathbf{1}(R > 0)] + \mathrm{E}[\frac{R \cdot \mathrm{MSER}(1 - \mathrm{MSER})}{R^2}\mathbf{1}(R > 0)] \\ &= \mathrm{MSER}^2 \cdot \Pr(R > 0)(1 - \Pr(R > 0)) + \mathrm{MSER}(1 - \mathrm{MSER}) \cdot \mathrm{E}[\frac{1}{R}\mathbf{1}(R > 0)].\end{aligned}$$

The first part goes to 0 because $\Pr(R > 0) \to 1$ as $m \to \infty$. The second part goes to 0 because R follows a binomial distribution $Bi(m, \Pr(R_1 = 1))$, and

$$\begin{aligned}\mathrm{E}[\frac{1}{R}\mathbf{1}(R > 0)] &< \mathrm{E}[\frac{2}{R+1}\mathbf{1}(R > 0)] < 2\mathrm{E}[\frac{1}{R+1}\mathbf{1}(R > 0)] \\ &= 2\mathrm{E}[\frac{1}{R+1}] - 2\mathrm{E}[\frac{1}{0+1}\mathbf{1}(R = 0)] \\ &= \frac{2}{(m+1)\Pr(R_1 = 1)} \cdot (1 - (1 - \Pr(R_1 = 1))^{m+1}) - 2\Pr(R = 0) \to 0\end{aligned}$$

as $m \to \infty$. Therefore, $\mathrm{SEP} - \mathrm{MSER} \to 0$ in probability. □

Before proving Proposition 3.2.2 and Proposition 3.2.3, we first prove the Lemma below.

Lemma 3.5.1. *Let* $A(\alpha, s) = \{y : \Phi^{-1}(\alpha s) < y < \Phi^{-1}(1 - \alpha(1 - s))$. *Let*

$$\gamma(A(\alpha, s)) = E_G[B_1(A(\alpha, s))\mathbf{1}(\theta > 0) + B_2(A(\alpha, s))\mathbf{1}(\theta < 0)],$$

we have that $\gamma(A(\alpha, s)) \leq \alpha s\pi_0 + \alpha(1 - s)(1 - \pi_0)$, *where* $\pi_0 = \Pr(\theta > 0)$.

Proof. Under the hierarchical model we have,

$$\gamma(A(\alpha, s)) = \mathrm{E}_G[\Phi(\Phi^{-1}(\alpha s) - \theta)\mathbf{1}(\theta > 0) + \Phi(\Phi^{-1}(\alpha(1 - s)) + \theta)\mathbf{1}(\theta < 0)].$$

Denote $\gamma(A(\alpha, s)) = \gamma_1 + \gamma_2$ where $\gamma_1 = \mathrm{E}_G[\Phi(\Phi^{-1}(\alpha s) - \theta)\mathbf{1}(\theta > 0)]$, and $\gamma_2 = \mathrm{E}_G[\Phi(\Phi^{-1}(\alpha(1-$

$s)) + \theta)\mathbf{1}(\theta < 0)]$. Suppose the probability density function of G is g,for γ_1 we have

$$\begin{aligned}
\gamma_1 =& \mathrm{E}_G[\Phi(\Phi^{-1}(\alpha s) - \theta)\mathbf{1}(\theta \geq 0)] \\
&= \int_0^\infty \Phi(-\theta + \Phi^{-1}(\alpha s))g(\theta)d\theta \\
&= \Phi(-\theta + \Phi^{-1}(\alpha s))G(\theta)|_0^\infty + \int_0^\infty \phi(-\theta + \Phi^{-1}(\alpha s))G(\theta)d\theta \\
&= -\alpha s(1-\pi_0) + \int_0^\infty \phi(-\theta + \Phi^{-1}(\alpha s))G(\theta)d\theta \\
&\leq -\alpha s(1-\pi_0) + \int_0^\infty \phi(-\theta + \Phi^{-1}(\alpha s))d\theta \\
&= -\alpha s(1-\pi_0) + \alpha s = \alpha s \pi_0
\end{aligned} \tag{3.10}$$

For γ_2 we have

$$\begin{aligned}
\gamma_2 =& \mathrm{E}_G[\Phi(\Phi^{-1}(\alpha(1-s)) + \theta)\mathbf{1}(\theta \leq 0)] \\
&= \int_{-\infty}^0 \Phi(\theta + \Phi^{-1}(\alpha(1-s)))g(\theta)d\theta \\
&= \Phi(\theta + \Phi^{-1}(\alpha(1-s)))G(\theta)|_{-\infty}^0 - \int_{-\infty}^0 \phi(\theta + \Phi^{-1}(\alpha(1-s)))G(\theta)d\theta \\
&= \alpha(1-s)(1-\pi_0) - \int_{-\infty}^0 \phi(\theta + \Phi^{-1}(\alpha(1-s)))G(\theta)d\theta \\
&\leq \alpha(1-s)(1-\pi_0)
\end{aligned} \tag{3.11}$$

Therefore $E(A(\alpha, s)) = \gamma_1 + \gamma_2 \leq \alpha s \pi_0 + \alpha(1-s)(1-\pi_0)$. □

Proof of Proposition 3.2.3. Denote R^t as the total number of rejections. We have

$$\mathrm{E}[R^t/m] = \mathrm{E}[\sum \mathbf{1}(R_j = 1)]/m = \sum \Pr(R_j = 1)/m = \Pr(R_i = 1),$$

where the last step is because of the exchangeability of the model. Again, we write MSER = γ/β, where $\gamma = \Pr(E_i = 1, R_i = 1)$ and $\beta = \Pr(R_i = 1)$. Since α_l^i is independent of Y_i, and by Lemma 3.5.1 and letting $s = 1/2$, we have

$$\Pr(E_i = 1, R_i = 1|\alpha_l^i) \leq \alpha_l^i/2 \leq \alpha_S((R(\alpha_l^i) - 1) \vee 0)/m \leq \alpha_S R^t/m.$$

Thus $\gamma = \mathrm{E}[\Pr(E_i = 1, R_i = 1|\alpha_l^i)] \leq \alpha_S \mathrm{E}[R^t/m] = \alpha_S \beta$. Therefore MSER $\leq \alpha_S$. □

Proof of Proposition 3.2.4. This Proposition follows from Benjamini and Yekutieli (2005) Theorem 1 and Corollary 3. To modify the proof for LC procedure, we should replace the kq/m in equation (4) in Benjamini and Yekutieli (2005) with $2kq/m$. Then it is easy to see that the SER can be controlled under q, which is the α_S we have in Chapter 3. Since LC is more conservative than LC, NLC also controls SER below α_S. □

Proof of Proposition 3.2.2. This is implied by Proposition 3.2.4 and Proposition 3.2.1. Let $\epsilon = \text{MSER} - \text{SER}$, thus $\text{MSER} = \text{SER} + \epsilon < \alpha_s + \epsilon$. According to the proof of Proposition 3.2.1, we have when $m \to \infty$, $\epsilon \to 0$ in probability. □

Proof of Proposition 3.2.5. We first show that $\widehat{\text{MSER}} \xrightarrow{p} \text{MSER}$. Since

$$\begin{aligned}\int_{-\infty}^{\infty} B_1 g_\eta(\theta) d\theta &= B_1 G_\eta(\theta)|_{-\infty}^{\infty} + \int_{-\infty}^{\infty} \phi(\Phi^{-1}(\alpha/2) - \theta) g_\eta(\theta) d\theta \\ &= \int_{-\infty}^{\infty} \phi(\Phi^{-1}(\alpha/2) - \theta) g_\eta(\theta) d\theta,\end{aligned}$$

$\text{E}_G[B_1]$ can be seen as $\text{E}_F[g_\eta(\theta)]$, where F has a probability density function $\phi(\Phi^{-1}(\alpha/2) - \theta)$. Since $g_\eta(\theta) < 1$, it is easy to see that $\lim_{\eta' \to \eta} \text{E}_F[g_{\eta'}(\theta)] = \text{E}_F[g_\eta(\theta)]$. Thus $\text{E}_F[g_\eta(\theta)]$ is continuous in η. Hence $\text{E}_G[B_1]$ is continuous in η. Similarly, we can show that $\text{E}_G[B_2]$ is a continuous function of η, and $\text{E}_G[B_1 + B_2]$ it is always nonzero. Similarly, $\text{E}_G[B_1 \mathbf{1}(\theta \geq 0) + B_2 \mathbf{1}(\theta \leq 0)]$ is also a continuous function in η. Therefore, MSER is a continuous function in η. Note that the difference between MSER and $\widehat{\text{MSER}}$ is that the former uses η and the later uses $\hat{\eta}$. If $\hat{\eta} \xrightarrow{p} \eta$, then we have $\widehat{\text{MSER}} \to \text{MSER}$ by Continuous Mapping Theorem.

Since α_o is the unique solution such that $\text{MSER}(A(\alpha_o)) - \alpha_S = 0$, and α^e is the unique solution such that $\widehat{\text{MSER}}(A(\alpha_e)) - \alpha_e = 0$, we have $\alpha_e \xrightarrow{p} \alpha_o$ by M-estimator theory (Lemma 5.10, Van der Vaart (1998)).

□

Proof of Proposition 3.4.1. We first show that $s = 1/2$ maximizes the MSDR. The MSDR

can be written as

$$\begin{aligned}\text{MSDR}(s) &= \int_{-\infty}^{\infty}(B_1(\theta,s)+B_2(\theta,s))g(\theta)d\theta \\ &= \int_{-\infty}^{0}(B_1(\theta,s)+B_2(\theta,s))g(\theta)d\theta + \int_{0}^{\infty}(B_1(\theta,s)+B_2(\theta,s))g(\theta)d\theta\end{aligned}$$

Since g is symmetric,

$$\begin{aligned}\text{MSDR}(s) &= \int_{0}^{\infty}(B_1(-\theta,s)+B_2(-\theta,s))g(\theta)d\theta + \int_{0}^{\infty}(B_1(\theta,s)+B_2(\theta,s))g(\theta)d\theta \\ &= \int_{0}^{\infty}((B_1(\theta,s)+B_2(\theta,s)+B_1(-\theta,s)+B_2(-\theta,s))g(\theta)d\theta\end{aligned}$$

Now we prove that the integrand is maximized when $s = 1/2$, which does not depend on θ. Thus MSDR(s) is maximized when $s = 1/2$. The integrand can be written as $H(s)g(\theta)$ where

$$H(s) = \Phi(\Phi^{-1}(\alpha s)-\theta)+\Phi(\Phi^{-1}(\alpha(1-s))+\theta)+\Phi(\Phi^{-1}(\alpha s)+\theta)+\Phi(\Phi^{-1}(\alpha(1-s))-\theta) \quad (3.12)$$

Taking the derivative with respect to s, we have

$$\begin{aligned}H(s)' &= \frac{\phi(\Phi^{-1}(\alpha s)-\theta)}{\phi(\Phi^{-1}(\alpha s))} + \frac{\phi(\Phi^{-1}(\alpha s)+\theta)}{\phi(\Phi^{-1}(\alpha s))} - \frac{\phi(\Phi^{-1}(\alpha(1-s))-\theta)}{\phi(\Phi^{-1}(\alpha(1-s)))} - \frac{\phi(\Phi^{-1}(\alpha(1-s))+\theta)}{\phi(\Phi^{-1}(\alpha(1-s)))} \\ &= c_1(exp(\Phi^{-1}(\alpha s)\theta) + exp(-\Phi^{-1}(\alpha s)\theta) - exp(\Phi^{-1}(\alpha(1-s))\theta) - exp(-\Phi^{-1}(\alpha(1-s))\theta)),\end{aligned} \quad (3.13)$$

where c_1 is a positive constant. It's easy to see that $s = 1/2$ is one solution to $H(s)' = 0$. Now we show that $H(s)$ is actually concave, hence $s = 1/2$ maximizes $H(s)$ for every $\theta > 0$. Therefore $s = 1/2$ maximizes MSDR(s). By taking derivative of $H(s)'$ with respect to s and rearrange, we obtain

$$\begin{aligned}H(s)'' = c_2(&exp((\Phi^{-1}(\alpha s)+\theta)^2/2) + exp((\Phi^{-1}(\alpha(1-s))+\theta)^2/2) - exp((\Phi^{-1}(\alpha s)-\theta)^2/2) \\ &- exp((\Phi^{-1}(\alpha(1-s))-\theta)^2/2)),\end{aligned}$$

where c_2 is a positive constant. Since $\Phi^{-1}(\alpha s) < 0$ and $\theta > 0$ (the integral is from 0 to ∞), we have

$$|\Phi^{-1}(\alpha s)-\theta| = |\Phi^{-1}(\alpha s)| + |\theta| \geq |\Phi^{-1}(\alpha s)+\theta|.$$

Thus

$$exp((\Phi^{-1}(\alpha s)+\theta)^2/2) - exp((\Phi^{-1}(\alpha s)-\theta)^2/2) < 0.$$

Similarly

$$exp((\Phi^{-1}(\alpha(1-s))+\theta)^2/2) - exp((\Phi^{-1}(\alpha(1-s))-\theta)^2/2) < 0.$$

Therefore $H(s)'' < 0$, and $s = 1/2$ maximizes MSDR(s).

To show MSER is minimized by $1/2$, we can first show that $s = 1/2$ minimizes γ, using the same technique as previous part of this proof. Then by noticing that MSER $= \gamma/$MSDR, we know $s = 1/2$ minimizes MSER. □

Chapter 4

PHYLOGENETIC TREE INFERENCE FOR CONTINUOUS DATA

4.1 Introduction

A phylogenetic tree describes the evolutionary history among a collection of organisms with the belief that these organisms evolved from a common ancestor. The tree is composed of nodes and branches, where leaf nodes represent the current-day organisms, internal nodes represent the divergence events, and branches represent evolutionary lineages changing over time. In phylogenetics, probabilistic models of the evolution of characters are used to infer the evolutionary history of a set of current-day organisms/taxa from a common ancestor. Since the early development of computational molecular biology in the 1960s, inferring phylogenetic trees has been one of the major research problems (Sokal and Sneath, 1961; Camin and Sokal, 1965). Early work considered continuous characters comprised of measurements on morphological traits. Subsequently, focus shifted on genetic information, and a lot of work has been done regarding the evolutionary models as well as the reconstruction methods for discrete-valued characters based on sequence data such as DNA, RNA or amino acid sequences (Friedman et al., 2002). More recently, models for continuous characters have received renewed attention in problems such as analysis of gene frequencies or gene expression levels. In this chapter, we are interested in phylogenetic analysis of such quantitative characters that are measured on a continuous scale.

One widely used model for continuous characters is the Brownian motion model (Felsenstein, 1973, 2004), where the character evolution is assumed to follow a Brownian motion process with variance σ^2 per unit time along the phylogenetic tree. Under this model, the net change along a branch of length t is drawn from a normal distribution with mean 0

and variance $\sigma^2 t$. Thus the length of the branch in the tree describes the time between two divergence events. The displacements in different branches of a tree are independent, because the small steps of which they are composed are all independent. Throughout the chapter, we assume the Brownian motion model for the phylogenetic tree, and as usual in phylogenetics, we focus on a rooted bifurcating tree, meaning that each internal node has exactly two descendants.

Learning the phylogenetic tree structure usually includes the tree topology recovery as well as branch lengths estimation based on observations from the leaf nodes only. One well-known method is the Neighbor-Joining (NJ) approach (Saitou and Nei, 1987), which is a bottom-up clustering method based on distances between each pair of leaf nodes. It is computationally fast and applicable to large data sets, but usually does not provide the most accurate solution (Bravo et al., 2009). In contrast, Felsenstein (1981) proposed the Maximum Likelihood method for phylogenetic tree recovery, which proved superior to other methods in terms of accuracy, but it is based on an exhaustive search making it computationally prohibitive for larger trees. Bravo et al. (2009) proposed a Mixed-Integer Programming approach that recasts the structure learning problem as a tree-structured covariance estimation problem, which amounts to finding the tree-structured covariance matrix nearest to an observed sample covariance matrix in a Mixed-Integer program (MIP). The method can be effective in terms of reliability and scalability, but solving the mixed integer optimization problem is not a simple task. In practice, MIP solvers require careful choice of tuning parameters, which can be difficult. Even with tactics for performance tuning, state-of-the-art hardware and software, some of them can still require hours, or even days, of running time and are not guaranteed to yield an optimal solution, or any solution at all, because of the combinatorial nature of MIPs. Other than reconstructing the tree based on observed data, phylogenetic tree with Brownian motion model is also studied in many other ways. Zwiernik et al. (2017) gave an efficient method for parameter estimation in Brownian motion tree model when the underlying tree structure is known. Pagel (1999) and Blomberg et al. (2003) proposed procedures to test the phylogenetic signal against a null model of no phylogenetic

signal, and a more recent comprehensive review of this topic is done by Münkemüller et al. (2012). O'Meara et al. (2006) proposed a likelihood-based procedure for testing the change of the rates of phenotypic evolution, and a follow-up work is done by Revell et al. (2018). A review of methods for phylogeny-aware analyses of microbiome data sets can be found in Washburne et al. (2018).

In this chapter, we explore three alternative reconstruction approaches for phylogenetic trees under the Brownian motion model. The first approach involves multiple independence tests to recover the tree topology. It is a bottom-up algorithm that iteratively clusters leaf nodes that are more likely to form a subtree together based on p-values estimated from series of independence tests. The second approach is a top-down method based on the sample covariance matrix, where leaf nodes are split into groups iteratively based on magnitudes of the sample covariances.

We then consider Maximum Likelihood tree reconstruction, i.e., we aim to find the tree that maximizes the likelihood based on observations from leaf nodes. Exhaustive search is usually infeasible due to the complexity of the large tree space. Greedy search algorithms have been popular in constructing phylogenetic tree using likelihood as a criterion (Adachi and Hasegawa, 1996), though maximum likelihood is not guaranteed. As an alternative, we propose a Structural Expectation Maximization (EM) algorithm, which is an extension of the classic EM algorithm to problems involving discrete structure learning. Friedman et al. (2002) have successfully developed a Structural EM algorithm for discrete characters, where substitution models were used to describe the evolutionary process. We develop a similar, yet in details very different algorithm for the Brownian motion model. We evaluate the performance of our methods in the simulation section, and we apply our methods to gene expression data. Our results show that the Structural EM algorithm is a reliable, computationally feasible method for state-of-the-art reconstruction of phylogenetic tree models for continuous data.

In the remainder of the chapter, we first review the Brownian motion model and some relevant graph terminologies. In Section 4.3, we propose two fast methods for phylogenetic

tree reconstruction, i.e., the independence test approach and the sample covariance based approach. In Section 4.4, we present the Structural EM algorithm for maximum likelihood phylogenetic tree discovery. Then we use simulations to evaluate the performance of the proposed methods, and use our methods to analyze the gene expression data in Section 4.5. We then discuss the estimation error of the EM algorithm on Brownian motion tree model. A discussion follows in Section 4.8. Proofs are given in Appendix.

4.2 Brownian Motion Model

Phylogenetic tree models are special instances of graphical models (Lauritzen, 1996). Each model is associated with a graph that consists of a vertex set V and an edge set E. In a directed graph, the edge set E comprises ordered pairs of vertices in V. The pair (a, b) represents an edge from a to b and is also denoted as $a \to b$. The set $\mathrm{pa}(b) = \{a \in V : a \to b\}$ is the set of parents of node b. If $a \in \mathrm{pa}(b)$, then b is a direct descendant of a. We write $\mathrm{An}(b) = \{a \in V : a = b \text{ or } a \to \cdots \to b\}$ for the set of ancestors of b, including b itself. If $a \in \mathrm{An}(b)$, then b is a descendant of a. A directed acyclic graph is a directed graphs without directed cycles, that is, subgraphs of the form $a \to \cdots \to a$ do not exist.

A rooted tree is a directed acyclic graph in which every node has at most one parent and precisely one node, the root, has no parents. The nodes with descendants are the tree's internal nodes, and the nodes without descendants are the leaf nodes. We assume the bifurcating case, in which the root has one direct descendant and every other internal node has two direct descendants.

Let $\mathcal{T}$ be a rooted tree, with p leaves indexed by $1, \ldots, p$, a root indexed by 0, and $p-1$ internal nodes indexed by $p+1, \ldots, 2p-1$. Each node and all its descendants together with the edges connecting them form a subtree $\mathcal{T}'$. We use $\mathcal{T}' \subset \mathcal{T}$ to denote that $\mathcal{T}'$ is a subtree of $\mathcal{T}$. Associate each node $i \in V$ with a random variable X_i describing the character state at this node. The Brownian motion model induced by $\mathcal{T}$ assumes that

$$X_i \;=\; X_{\mathrm{pa}(i)} + \epsilon_i, \qquad i = 1, \ldots, 2p-1, \tag{4.1}$$

where the ϵ_i's are independent Gaussian random variables that all have mean 0 but generally different variances $d_i > 0$. The variance d_i represents the branch length between i and $\text{pa}(i)$; see Figure 4.1 for an example with 4 leaves. For the root node, we assume that $X_0 = x_0$ where x_0 is an unknown constant. The equations in (4.1) then imply that the random vector $\boldsymbol{X} = [X_1, \ldots, X_{2p-1}]^T \in \mathbb{R}^{2p-1}$ follows a multivariate normal distribution with mean vector $(x_0, \ldots, x_0)^T$ and a highly structured covariance matrix that reflects the tree topology T and the branch lengths $\boldsymbol{d} = \{d_i\}_{i \in \{1,\ldots,2p-1\}}$. Hence we write the tree model as $\mathcal{T} = (T, \boldsymbol{d})$.

Example 4.2.1. *Under the Brownian motion model for the tree from Figure 4.1, the marginal distribution of the leaves* $[X_1, X_2, X_3, X_4]$ *is a normal distribution with covariance matrix*

$$\Sigma = \begin{bmatrix} d_1 + d_5 + d_7 & d_5 + d_7 & d_7 & d_7 \\ d_5 + d_7 & d_2 + d_5 + d_7 & d_7 & d_7 \\ d_7 & d_7 & d_3 + d_6 + d_7 & d_6 + d_7 \\ d_7 & d_7 & d_6 + d_7 & d_4 + d_6 + d_7 \end{bmatrix}.$$

Each covariance is a linear combination of elements in $\boldsymbol{d}$.

In general, the covariance between node i and node j is the sum of branch lengths for the path starting at the root and ending at the last common ancestor of node i and node j. So the covariance matrix for the leaf nodes is

$$\Sigma = BDB^T \tag{4.2}$$

where the ij^{th} element of matrix B is 1 if $j \in \text{An}(i)$, 0 otherwise, and $D = diag(\boldsymbol{d})$ is the diagonal matrix with $\boldsymbol{d}$ being the diagonal elements. Here the binary matrix B describes the tree topology, and for the tree in Figure 4.1, we have

$$B = \begin{bmatrix} 1 & 0 & 0 & 0 & 1 & 0 & 1 \\ 0 & 1 & 0 & 0 & 1 & 0 & 1 \\ 0 & 0 & 1 & 0 & 0 & 1 & 1 \\ 0 & 0 & 0 & 1 & 0 & 1 & 1 \end{bmatrix}.$$

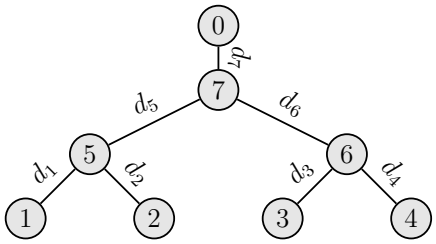

Figure 4.1: A bifurcating tree with four nodes.

Suppose now we have a data set $O = \{x_1^{(k)}, \ldots, x_p^{(k)}\}_{k=1}^n$ that consists of N *i.i.d.* observations, which are assumed to be drawn from the marginal distribution of p leaf nodes. Some algorithms reconstruct phylogenetic tree structure from such a data set O using the correlation matrix. For example, it is popular to use the correlation matrix as a distance matrix that measures how far away two nodes are in the tree, and combining with algorithms like NJ approach the tree structure is inferred. This method works for unrooted tree recovery as an unrooted tree topology is identifiable based on correlation matrix (Shiers et al., 2016). However, for rooted tree case that we discuss here, the correlation matrix is not enough to allow inferences of the root location. Moreover, for small number of leaves, specifically when $p = 3$, the two tree topologies in Figure 4.2 can not be distinguished. Details of the proof and more discussion about using correlation matrix to infer rooted trees are in Section 4.7.

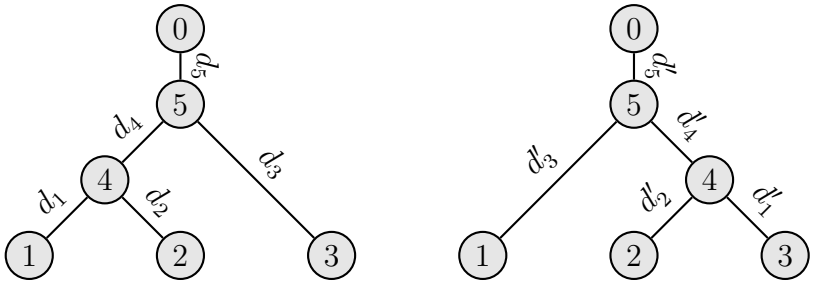

Figure 4.2: Two trees $\mathcal{T} = (T, \boldsymbol{d})$ (left) and $\mathcal{T}' = (T', \boldsymbol{d}')$ (right) that can not be distinguished based only on the correlation matrix.

Likelihood is usually used as a criterion to evaluate how well a model explains the data generating mechanism. In Section 4.4 of this chapter, we will propose an algorithm for maximum likelihood phylogenetic tree discovery. Sample size is crucial in maximum likelihood calculation for graphical models. The minimum number N such that for a sample of size $n \geq N$ the log-likelihood function is almost surely bounded above is called the *maximum likelihood threshold* (Gross and Sullivant, 2018). In fact, the results from Drton et al. (2018) implied that for mixed graph models that are closely related to the phylogenetic tree model we consider here, N needs to be at least as large as the number of nodes in a fully connected mixed graph. Lauritzen et al. (2017) showed that under total positivity condition, i.e. $d_i > 0$ for $i \in V$, Brownian motion tree model only requires $N = 2$ in order to have a bounded maximum likelihood. Hence in order to reconstruct the maximum likelihood tree for the model we consider in this project, we do not require the sample size to be greater than the number of leaf nodes, making it possible to perform such analyses in high-dimensional scenarios.

4.3 Tree Recovery Based on Independence Test or Sample Covariance

4.3.1 Independence Test Approach

In this subsection, we propose a method for phylogenetic tree recovery based on independence tests. This method reconstructs the tree topology based on p-values that evaluates the hypothesis that two sets of nodes form a subtree together. This method is only for tree topology recovery, and the branch lengths are not estimated. It is based on the following properties of phylogenetic Brownian motion model.

Proposition 4.3.1. *Let $\mathcal{T} = (T, \boldsymbol{d})$ be a phylogenetic tree with nonzero branch lengths, let $C_1, C_2 \subset \{1, \ldots, p\}$ be two nonempty and disjoint leaf index sets of two substrees of $\mathcal{T}$. Under the Brownian motion model, we have: $X_i - X_j \perp\!\!\!\perp X_k$, i.e. $X_i - X_j$ is independent of X_k, for $\forall i \in C_1, j \in C_2, k \in \{1, \ldots, p\} \backslash \{C_1 \cup C_2\}$ if and only if the nodes in set C_1 and nodes in set C_2 form a subtree of $\mathcal{T}$.*

Note that the leaf nodes jointly follow a multivariate Gaussian distribution, hence the

correlation matrix contains all the independence information of the distribution. In particular, one can do independence tests of two sets of random variables to learn the independence relations among the leaf nodes and then infer the topology of the tree. In what follows, we will write $p_{C_1-C_2}$ for the p-value calculated from the independence test between $\boldsymbol{X}_{C_1-C_2}$ and $\boldsymbol{X}_{\{1,\ldots,p\}\setminus\{C_1,C_2\}}$, where $\boldsymbol{X}_{C_1-C_2} = (X_s - X_t : s \in C_1, t \in C_2)^T$ is the vector of all possible pairwise differences between C_1 and C_2. The p-value is calculated based on Wilks' Λ-statistic, and the details for such independence test can be found in *chapter 7.4* in Rencher (2003). Note that this independence test only works for cases when $p < n$. Yang et al. (2015) proposed an independence test procedure for high-dimensional scenarios. However, implementing their approach in our algorithm is computationally inefficient, hence here we only focus on low dimensional setting.

We now describe in **Algorithm** 1 the procedure of tree topology reconstruction by independence tests. This algorithm takes as input the observations from the leaf nodes, then iteratively groups two sets of leaf nodes with largest p-value together based on independence tests, until all nodes are combined into one group. Note that the combining of two sets of nodes at each iteration recovers one branch split of the tree, so the whole tree topology with p leaf nodes will be recovered from bottom up in such manner with $p - 1$ iterations. One example of this topology reconstruction process for a tree with $p = 5$ leaf nodes is illustrated in Figure 4.3.

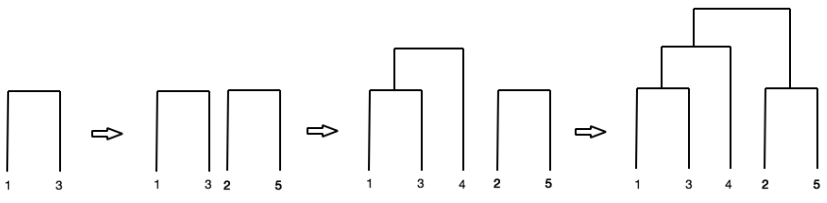

Figure 4.3: An example of the topology reconstruction process for a tree with 5 leaf nodes.

Algorithm 1 Phylogenetic Tree Topology Reconstruction by Independent Test

Require: samples $O = \{x_1^{(k)}, x_2^{(k)}, \ldots, x_p^{(k)}\}_{k=1}^n \in \mathbb{R}^{N\times p}$ from leaf nodes.

1: Set $C_i = \{i\}, \forall i \in \{1, \ldots, p\}$, $nodeSet = \{1, \ldots, p\}$, $iter = 1$, and initialize $split(j), \forall j \in \{1, \ldots, p-1\}$ to $\emptyset$.
2: **repeat**
3: $\forall i, j \in nodeSet$, calculate $p_{C_i - C_j}$ as stated above.
4: Find two node sets $C_{i'}, C_{j'}$ such that $p_{C_{i'} - C_{j'}} = max_{i,j}\, p_{C_i - C_j}$.
5: Combine nodes in $C_{i'}$ and $C_{j'}$.
6: Record the two combining sets of nodes in $split(iter)$.
7: Let $C_{i'} = C_{i'} \cup C_{j'}$, $nodeSet = nodeSet \backslash j'$, $iter = iter + 1$.
8: **until** $|nodeSet| = 1$.
9: **return** The record of two combining sets of nodes at each iteration.

4.3.2 Sample Covariance Approach

We now present another algorithm that achieves the tree structure recovery based directly on the sample covariance matrix of the leaf nodes. Let

$$S = \frac{1}{n-1}\sum_{k=1}^{n}(\boldsymbol{X}^{(k)} - \bar{\boldsymbol{X}})^T(\boldsymbol{X}^{(k)} - \bar{\boldsymbol{X}}) \tag{4.3}$$

be the sample covariance matrix where $\boldsymbol{X}^{(k)}$ represents the k^{th} observation of $\boldsymbol{X}$ and $\bar{\boldsymbol{X}}$ is the sample mean calculated as $\frac{1}{n}\sum_{k=1}^{n}\boldsymbol{X}^{(k)}$. Note that in the Brownian motion model for phylogenetic trees, the marginal distribution of the leaf nodes is a multivariate normal distribution with a common mean x_0. Hence another way to calculate the sample covariance matrix is to use $\hat{\mu}\mathbf{1}_p$ instead of $\bar{\boldsymbol{X}}$ in (4.3), where $\hat{\mu} = \frac{1}{n}\sum_{l=1}^{p}\bar{\boldsymbol{X}}_l$ is the mean of all the entries of $\bar{\boldsymbol{X}}$. According to our numerical evaluations, these two approaches give very close results when used in our algorithm. We choose to use (4.3) to calculate the sample covariance matrix for convenience and to accommodate the general consistency theory framework that will be used later.

Our algorithm is simply based on the following facts:

- In a bifurcating tree, each internal node splits its descendants into two disjoint subtrees.
- For two disjoint subtrees, the covariances between nodes from the same subtree are larger than the covariances between nodes from different subtrees.
- S is an unbiased, consistent estimate for the true covariance matrix Σ of the leaf nodes.

The algorithm is summarized in **Algorithm** 2. If we can correctly specify the covariance matrix, this algorithm returns the true tree topology and branch lengths of the phylogenetic tree under the Brownian motion model. In reality, we use the sample covariance matrix as an estimate of the true covariance matrix and use it as the input of this algorithm. It works well when the sample size is large and the Brownian motion model reflects the true data generating mechanism.

Now we examine the error of the tree estimate obtained by the sample covariance matrix based algorithm. To quantify the parameter estimation error, we first need to find an appropriate norm to quantify it. Notice that the usually used L^2-norm may not be an appropriate norm to measure the distance between the estimated branch length $\hat{\boldsymbol{d}}$ and the true branch length $\boldsymbol{d^*}$ as the L^2-norm weights all the entries in a vector the same, making larger entries in the vector more influential. On the other hand, a ratio is a more natural way to compare the variance parameters than the difference. To quantify the estimation error for the tree model we use in this chapter, we propose to use the weighted norm as a measure. For a vector $\boldsymbol{d}$, we define the weighted 2-norm as

$$||\boldsymbol{d} - \boldsymbol{d^*}||^2_{2,d^*} = \sum_{i \in \{1,\ldots,2p-1\}} |(d_i - d_i^*)/d_i^*|^2. \tag{4.4}$$

We also define the weighted maximum norm as

$$||\boldsymbol{d} - \boldsymbol{d^*}||_{\infty,d^*} = \max_{i \in \{1,\ldots,2p-1\}} |\frac{\hat{d}_i}{d_i} - 1|. \tag{4.5}$$

These two weighted norms are actually measuring the distance between the element-wise ratio of $\boldsymbol{d}$ to $\boldsymbol{d}^*$ to 1. We will work with this type of weighted norms in our later discussions. The following result gives estimation error bounds for phylogenetic tree estimation with the sample covariance based method.

Proposition 4.3.2. *Suppose there is a positive constant d such that $\min_{i\in\{1,\ldots,2p-1\}} d_i > d$. Then using the sample covariance matrix based method, we have a probability of at least $1-2\exp\{\log(p(p-1))-nd^2/(12800\max_j(\Sigma_{jj})^2)\}$ to recover the topology of the tree. Moreover, with the estimated branch length denoted as $\hat{\boldsymbol{d}} = \{\hat{d}_1, \ldots, \hat{d}_{2p-1}\}$, conditional on the true tree topology, we have*

$$\Pr(||\boldsymbol{d} - \boldsymbol{d}^*||_{md^*} < \epsilon) > 1 - 8\exp\Big(\log(2p) - \frac{n\epsilon^2 d^2}{12800\max_j(\Sigma_{jj})^2}\Big).$$

This result quantifies the estimation error of the sample covariance matrix method in recovering the tree topology and estimating the branch length for fixed n and p. For example, when both n and p are large, if $\sqrt{\log(p)/n}$ is small enough, we can still recover the tree topology with high probability.

4.4 Structural EM Algorithm

In this section, we consider the maximum likelihood approach for the reconstruction of the phylogenetic tree based on observations from leaf nodes. In the Brownian motion model we consider here, observations from the internal nodes are needed to compute the complete likelihood. For a given tree structure, the EM algorithm yields one approach to obtain the maximum likelihood estimates for the parameters in the tree model. For unknown tree structure, Friedman et al. (2002) presented a Structural EM algorithm for learning ML phylogenetic tree under substitution model with discrete-valued characters. The Structural EM framework we establish for the phylogenetic Brownian motion model with quantitative characters has two main differences: 1) in the substitution model setting, the distribution

Algorithm 2 Phylogenetic Tree Topology Reconstruction via Sample Covariances

Require: S: Sample covariance matrix of the leaf nodes.

1: Create a root node. Set $\{1, \ldots, p\}$ as the descendants of the root node. Let $C = \{\{1, \ldots, p\}\}$.
2: **loop**
3: $\forall A \in C$ with $|A| > 1$
4: Denote r as the node whose descendant set is A.
5: Find $(i^*, j^*) = \arg\min S_{ij}$ for $i \neq j$ and $i, j \in A$ (if this (i^*, j^*) pair is not unique, we just pick one of them).
6: Let $A_1 = \{i^*\} \cup \{k : k \in A \backslash \{i^*, j^*\} \text{ and } S_{i^*k} > S_{j^*k}\}$, and $A_2 = A \backslash A_1$.
7: **if** $|A_1| = 1$ **then**
8: Set i^* as a immediate descendant of node r.
9: **else**
10: Create a new node, set it as a immediate descendant of node r, and set A_1 as its descendants.
11: **end if**
12: Repeat line 7 - 11 for A_2.
13: Remove A from C, and add A_1, A_2 to C.
14: **end loop** $\forall A \in C$ we have $|A| = 1$.
15: Now, we have the topology T of the tree. Branch length $\boldsymbol{d}$ can be estimated as follows:
16: Let $a_0 = 0$.
17: **loop**
18: For internal node j from top to bottom, left to right:
19: j split its descendants into two subsets C_{j1} and C_{j2}.
20: Let $a_j = \text{mean}(S_{kl} | \forall k \in C_{j1}, l \in C_{j2})$.
21: Let $d_j = a_j - a_i$, where i is the parent node of j.
22: For a leaf node j, let $d_j = S_{jj} - a_i$.
23: Replace all negative branch lengths with a small positive real number.
24: **end loop**
25: **return** T and $\boldsymbol{d}$.

of nodes is invariant to the placement of root, thus it deals with the unrooted tree, while in our case we assume the root node is an unobserved constant and deal with the rooted tree; 2) the maximization procedure and the expected sufficient statistics that are used to summarize the data are different. Our algorithm may be summarized as follows:

We start from an initial tree. For each iteration, we proceed with the following steps. First, we maximize the expected log-likelihood with respect to both the topology and branch lengths. The maximization is done by calculating 'weights' between any pairs of nodes, and then finding the maximum spanning tree. Then we modify the tree into a rooted bifurcating tree without changing the likelihood. Second, we maximize the expected log-likelihood with respect to the root state. The expected log-likelihood is conditional on the tree at the previous iteration, and observed data. We iterate this procedure until the change of the likelihood is under a certain threshold. According to the following theorem, improving the expected log-likelihood forces an improvement of the actual likelihood.

Theorem 4.4.1. *(Friedman et al., 2002) For any $\mathcal{T}$ and $\mathcal{T}^0$, let $Q(\mathcal{T}|\mathcal{T}^0)$ be the expected log-likelihood conditional on $\mathcal{T}^0$, and $l(\mathcal{T})$ be the objective log-likelihood. Then*

$$Q(\mathcal{T}|\mathcal{T}^0) - Q(\mathcal{T}^0|\mathcal{T}^0) \leq l(\mathcal{T}) - l(\mathcal{T}^0).$$

The choice of initial tree is important for the performance of the algorithm. We suggest to use the result of our sample covariance matrix based method as the initialization of the Structural EM algorithm. We show that the Structural EM algorithm is locally contractive and discuss about the estimation error of the Structural EM algorithm in Section 4.6. We now describe other components needed for the Structural EM algorithm in the following subsections.

4.4.1 Log-likelihood of Complete Data

In the complete data scenario, we have samples from the leaf nodes $O = \{x_1^{(k)}, \ldots, x_p^{(k)}\}_{k=1}^n$, and samples from the internal nodes $H = \{x_{p+1}^{(k)}, \ldots, x_{2p-1}^{(k)}\}_{k=1}^n$. The log-likelihood function

of the complete data given $(T, \boldsymbol{d})$ and root value x_0 is

$$l_c(T, \boldsymbol{d}, x_0) = \sum_{k=1}^{n} \log p_1(x_1^{(k)}, x_2^{(k)}, \ldots, x_{2p-1}^{(k)} | T, \boldsymbol{d}, x_0)$$

where $p_1(\cdot|T, \boldsymbol{d}, x_0)$ is the pdf of joint distribution of

$$[X_1, \ldots, X_{2p-1}] \sim N(x_0 \mathbf{1}_{2p-1}, \tilde{\Sigma}), \tag{4.6}$$

where $\tilde{\Sigma}$ is determined by $(T, \boldsymbol{d})$. If we assume that root state is given value x_0^l, then the log-likelihood function of the complete data, denoted as $l_c^{x_0^l}(T, \boldsymbol{d})$, factorizes as

$$\begin{aligned}
l_c^{x_0^l}(T, \boldsymbol{d}) &= \sum_{k=1}^{n} \log p_1(x_1^{(k)}, x_2^{(k)}, \ldots, x_{2p-1}^{(k)} | T, \boldsymbol{d}, x_0 = x_0^l) \\
&= \sum_{k=1}^{n} \log \prod_{i=1}^{2p-1} p_{X_i|X_{\pi(i)}}(x_i^{(k)} | x_{\pi(i)}^{(k)}) \quad [\forall n, x_0^{(k)} = x_0^l] \\
&= \sum_{k=1}^{n} \log \prod_{(i,j)\in E} p_{X_j|X_i}(x_j^{(k)} | x_i^{(k)}) \\
&= \sum_{(i,j)\in E} \sum_{k=1}^{n} \log p_{X_j|X_i}(x_j^{(k)} | x_i^{(k)}).
\end{aligned}$$

Note that for given $T, \boldsymbol{d}$ and x_0^l, if $i = \mathrm{pa}(j)$, then $(X_j | X_i = x_i^{(k)}) \sim N(x_i^{(k)}, d_{ij})$. Here we use d_{ij} instead of d_j to represent the branch length between i and j in this section for a more clear notation purpose, when the tree topology is unknown. Hence the complete log-likelihood with given x_0^l is

$$l_c^{x_0^l}(T, \boldsymbol{d}) = \sum_{(i,j)\in E} \Big[-\frac{n}{2} \log 2\pi - \frac{n}{2} \log d_{ij} - \sum_{k=1}^{n} \frac{1}{2d_{ij}} (x_j^{(k)} - x_i^{(k)})^2 \Big]. \tag{4.7}$$

Now we assume that the root state is unknown but $(T, \boldsymbol{d})$ are given with $(T^l, \boldsymbol{d}^l)$. The

complete log-likelihood with given $(T^l, \boldsymbol{d}^l)$ becomes

$$\begin{aligned} l_c^{(T^l,\boldsymbol{d}^l)}(x_0) &= \sum_{k=1}^{n} \log p_1(x_1^{(k)}, x_2^{(k)}, \ldots, x_{2p-1}^{(k)} | T = T^l, \boldsymbol{d} = \boldsymbol{d}^l, x_0) \\ &= \sum_{k=1}^{n} \log \prod_{i=1}^{2p-1} p_{X_i|X_{\pi(i)}}(x_i^{(k)} | x_{\pi(i)}^{(k)}) \\ &= \sum_{k=1}^{n} \log p_{X_a|X_0}(x_a^{(k)} | x_0) \\ &= \sum_{k=1}^{n} -\frac{1}{2d_{0,a}^l}(x_a^{(k)} - x_0)^2 \end{aligned} \tag{4.8}$$

where a is the immediate descendant of root x_0, and irrelevant constants are omitted.

4.4.2 Maximizing Expected Log-likelihood

In phylogenetic applications, we usually only have observations from leaf nodes while observations from internal nodes are unobserved. Here we use Structural EM algorithm to reconstruct the maximum likelihood tree iteratively. At each maximization step, instead of maximizing the complete log-likelihood directly, we maximize the expected log-likelihood. Assuming that the tree we obtain from the last iteration is $\mathcal{T}^l = (T^l, \boldsymbol{d}^l, x_0^l)$. By (4.7), when the root state is given with value x_0^l, the expected log-likelihood conditional on observed data and (T^l, d^l) is

$$Q_1(T, \boldsymbol{d}) = \mathrm{E}[l_c^{x_0^l}(T, \boldsymbol{d}) | O, \mathcal{T}^l] = \sum_{(i,j)\in E} \Big[-\frac{n}{2}\log 2\pi - \frac{n}{2}\log d_{ij} - \sum_{k=1}^{n} \frac{1}{2d_{ij}} \mathrm{E}[(X_j^{(k)} - X_i^{(k)})^2 | O, \mathcal{T}^l] \Big]. \tag{4.9}$$

Let

$$\mathcal{W}_{ij}(d_{ij}) = -\frac{n}{2}\log 2\pi - \frac{n}{2}\log d_{ij} - \sum_{k=1}^{n} \frac{1}{2d_{ij}} \mathrm{E}[(X_j^{(k)} - X_i^{(k)})^2 | O, \mathcal{T}^l]. \tag{4.10}$$

Maximizing $\mathcal{W}_{ij}(d_{ij})$ with respect to d_{ij}, the solution is

$$\begin{aligned} \hat{d}_{ij} &= \arg\max_{d_{ij}} \mathcal{W}_{ij}(d_{ij}) = \frac{1}{n}\sum_{k=1}^{n} \mathrm{E}[(X_j^{(k)} - X_i^{(k)})^2 | O, \mathcal{T}^l], \\ w_{ij} &= \max_{d_{ij}} \mathcal{W}_{ij}(d_{ij}) = \mathcal{W}_{ij}(\hat{d}_{ij}) = -\frac{n}{2}(\log 2\pi + \log \hat{d}_{ij} + 1). \end{aligned} \tag{4.11}$$

It is easily seen that $w_{ij} = w_{ji}$. Hence,

$$\begin{aligned}\max_{T,\boldsymbol{d}} Q_1(T,\boldsymbol{d}) &= \max_{T,\boldsymbol{d}} \sum_{(i,j)\in T} \mathcal{W}_{ij}(d_{ij}) = \max_{T} \sum_{(i,j)\in T} \max_{\boldsymbol{d}} \mathcal{W}_{ij}(d_{ij}) \\ &= \max_{T} \sum_{(i,j)\in T} \max_{d_{ij}} \mathcal{W}_{ij}(d_{ij}) = \max_{T} \sum_{(i,j)\in T} w_{ij}.\end{aligned} \tag{4.12}$$

Therefore, in the case of complete data with given root value, the problem of finding ML estimates of $(T, \boldsymbol{d})$ has been reduced to the problem of finding the tree topology with the maximum value of $\sum_{(i,j)\in T} w_{ij}$, which is equivalent to finding a maximum spanning tree. This maximization problem can then be solved by applying e.g. Kruskal's algorithm (Kruskal, 1956a). However, the constructed spanning tree is not necessarily a bifurcating tree. We will show how to transform the spanning tree to a bifurcating tree without changing the likelihood in next subsection.

Suppose the current tree we have is $\mathcal{T}^l$. By equation (4.8), the expected log-likelihood when the root state is unknown is

$$Q_2(x_0) = \sum_{k=1}^{n} -\frac{1}{2d_{0,a}^l} \mathrm{E}[(X_a^{(k)} - x_0)^2 | D, \mathcal{T}^l].$$

Hence the $\hat{x}_0$ that maximizes $Q_2(x_0)$ is

$$\hat{x}_0 = \frac{\sum_{k=1}^{n} \mathrm{E}[X_a^{(k)} | O, \mathcal{T}^l]}{n}. \tag{4.13}$$

By equation (4.11) and (4.13), to complete the maximization step, all we need is to compute the conditional expectation of X_{i+p} and $X_{i+p}X_{j+p}$ for $i, j \in \{1, \ldots, p-1\}$ given observed data O and current tree state $\mathcal{T}^l$. By (4.6), we have

$$\begin{aligned} & [X_{p+1}, \ldots, X_{2p-1}]^T | [X_1, \ldots, X_p]^T \sim N(\boldsymbol{\mu}_c^l, \Sigma_c^l), \\ \text{where} \quad & \boldsymbol{\mu}^c = x_0^l \mathbf{1}_{p-1} + \Sigma_{HO}^l (\Sigma_{OO}^l)^{-1} ([X_1, \ldots, X_p]^T - x_0^l \mathbf{1}_p), \\ & \Sigma^c = \Sigma_{HH}^l - \Sigma_{HO}^l (\Sigma_{OO}^l)^{-1} \Sigma_{OH}^l. \end{aligned} \tag{4.14}$$

Here Σ^l is the current estimate of covariance matrix and we divide it into blocks

$$\Sigma^l = \begin{pmatrix} \Sigma_{OO}^l & \Sigma_{OH}^l \\ \Sigma_{HO}^l & \Sigma_{HH}^l \end{pmatrix},$$

where we use O to index the observed variables and H to index the hidden variables. Hence it is easy to obtain the first and second moments of the unobserved random variables conditional on the observed random variables: $\forall i, j \in \{1, \ldots, p-1\}$

$$\begin{aligned}
&\mathrm{E}[X_{i+p}|[X_1, \ldots, X_p]] = \mu_i^c, \\
&\mathrm{E}[X_{i+p}^2|[X_1, \ldots, X_p]] = (\mu_i^c)^2 + \Sigma_{ii}^c, \\
&\mathrm{E}[(X_{i+p} - X_{j+p})^2|[X_1, \ldots, X_p]] = (\mu_i^c - \mu_j^c)^2 + (\Sigma_{ii} + \Sigma_{jj} - 2\Sigma_{ij}).
\end{aligned}$$

Maximum spanning bifurcating tree

The maximum spanning tree is found using Kruskal's algorithm. However, the obtained maximum spanning tree may not be bifurcating. Fortunately we can obtain an equivalent likelihood bifurcating tree by using the following procedure which is an adaptation of the procedure in Friedman et al. (2002).

Let $de(i)$ be the degree of the ith node, which is the number of nodes connected to the ith node. For a phylogenetic tree, in our setting, each leaf node has degree 1 (connecting to one internal node), each internal node has degree 3 (connecting to two immediate descendants and one parent), and the root node has degree 1. Let $De(i)$ be the degree of the ith node in a bifurcating tree, we have $De(i) = 1$ when $0 \leq i \leq p$, and $De(i) = 3$ when $p+1 \leq i \leq 2p-1$.

Given a maximum spanning tree, we do the following transformations to turn it into a likelihood-equivalent bifurcating tree:

- For each internal node j:
  - If $de(j) = 1$, we remove this internal node.
  - If $de(j) = 2$ and $(i, j), (j, k) \in E$, then remove node j and branches (i, j), (j, k). Then add branch (i, k).
- For each node i with $de(i) > De(i)$, denote $i_1, \ldots, i_{(de(i))}$ as its neighbors:
  - Add a new node i'.

- Replace the edges $(i, i_{De(i)}), \ldots, (i, i_{de(i)})$ with $(i', i_{De(i)}), \ldots, (i', i_{de(i)})$
- Add edge (i, i') with length 0.

Note that in practice, in the last step we usually use a very small number ϵ instead of 0 as a "pseudo branch" so that in the E-step, the relevant internal nodes are appropriately updated with "imputed" observations and allow us to distinguish the nodes that are connected by the "pseudo branch" in the next iteration. By applying this procedure, we can transform the spanning tree into a bifurcating tree with a negligible change in the likelihood.

4.5 Numerical Results

4.5.1 Simulation Studies

In this subsection, we evaluate the performance of our methods using simulated examples. We first compare the performance of the independence test method, the sample covariance matrix based method, and the Structural EM algorithm. Each time, we randomly simulate a tree. Then data are simulated from the tree. We then use the three methods mentioned to recover the topology of the tree and compare the recovered topology with the true topology. The performance is evaluated using *topological distance* (Penny and Hendy, 1985), which is the number of edges for which there is no edge in the other tree that gives the same partitions of leaf nodes. We try both smaller trees ($p = 10$) and larger trees ($p = 50$), for different sample sizes ($n = 100, 1000, 10000$). All experiments and analyses were carried out in R (R Core Team, 2018).

The results are summarized in Figure 4.4. As we can see, all three approaches have good performance in recovering the tree topology (note that the average topological distance between two random trees is 14 for $p = 10$, and 94 for $p = 50$). As the sample size grows, the accuracy increases. The independence test approach and sample covariance matrix based approach are faster than the Structural EM algorithm. For the case of $p = 50$, average running time is 10 seconds for independence test approach, 0.02 seconds for sample covariance matrix based approach, and 84 seconds for Structural EM algorithm. However, Structural

EM algorithm has a noticeable improvement over the other two methods, especially when sample size n is relatively small. When n is large, all three methods perform well in terms of recovering the true tree topology.

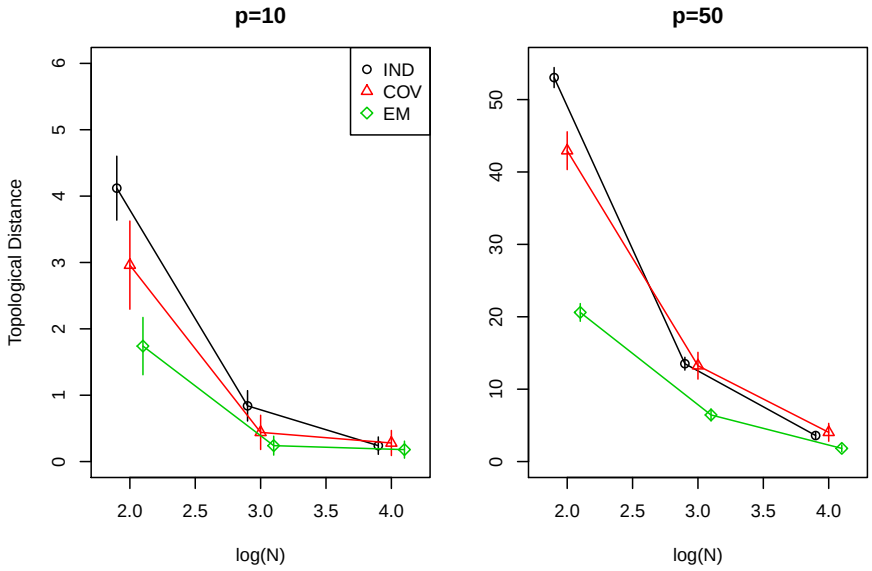

Figure 4.4: Comparison of the independence test method (IND), the sample covariance matrix based method (COV), and the Structural EM algorithm based on simulations. This figure summarizes the results from 100 replications, where the points represent a mean and the bars are the Monte Carlo error bars. Points are shifted slightly horizontally to avoid overlaps.

We then look at the likelihood of the estimated trees from the sample covariance matrix based method and the Structural EM algorithm. We compare the likelihood calculated on a training data set and a test data set ($n = 100, 1000, 10000$ for both training and test sets), and the results are shown in Figure 4.5 and Figure 4.6. As we can see, when sample size n is relatively small, Structural EM algorithm has a noticeable improvement over the sample covariance based method. When n is large, both methods perform well in terms of getting a likelihood that is close to that of the true tree. This is because Structure EM algorithm

is guaranteed to give higher or equal likelihood than the initial tree on the one hand so we expect an improvement from using Structural EM algorithm; on the other hand the sample covariance based method is a consistent method, so when the sample size is large, we expect it to perform well and the Structural EM algorithm has not much space to improve.

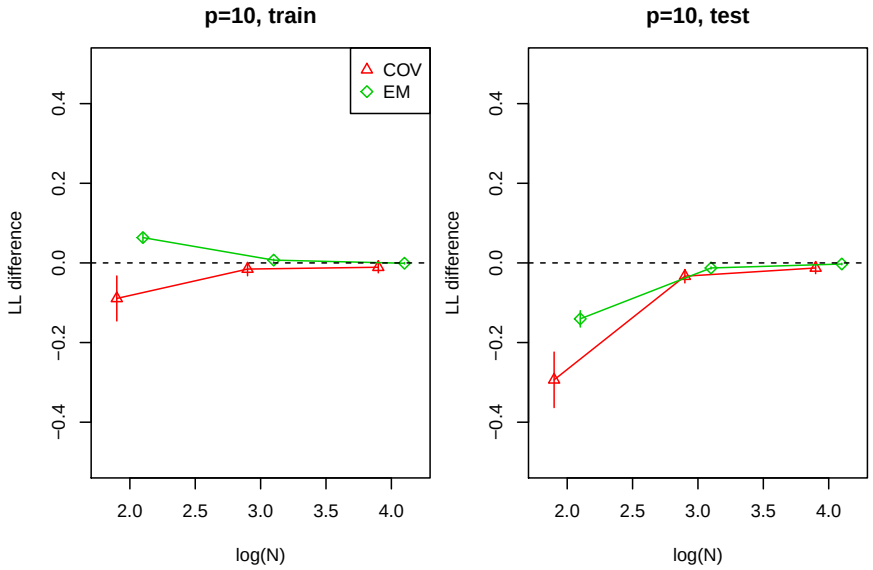

Figure 4.5: Comparison in log-likelihood between the sample covariance matrix based method (COV) and the Structural EM algorithm based on simulations (the difference between the log-likelihood of the estimated tree and the true tree) when $p = 10$. This figure summarizes the results from 100 replications, where the points represent the mean value and the bars are the Monte Carlo error bars. Note that points are shifted slightly horizontally to avoid overlaps.

4.5.2 Gene Expression Data Analysis

In this subsection, we analyze the gene expression data from Brawand et al. (2011). In order to understand the dynamics of mammalian transcriptome evolution, the sequencing of polyadenylated RNA from six organs of ten species was completed. These species represent

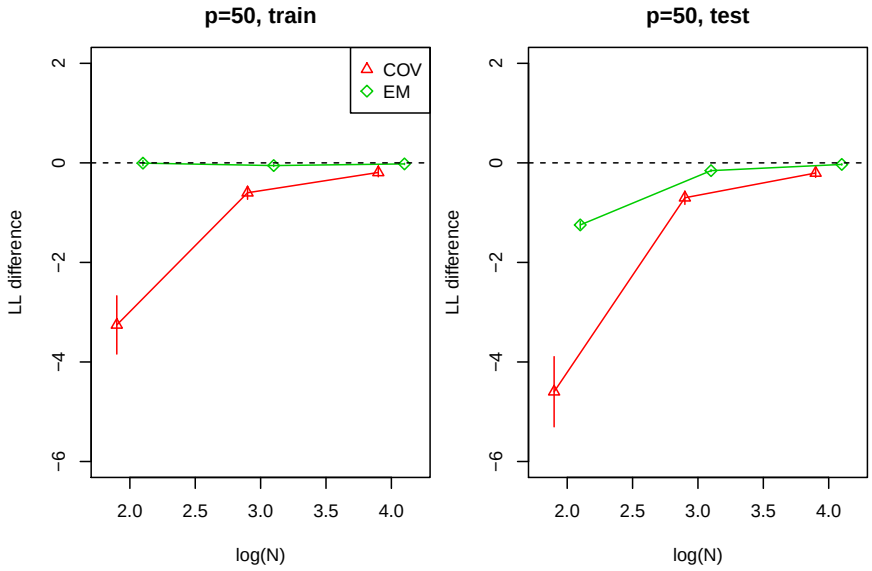

Figure 4.6: Comparison between the sample covariance matrix based method (COV) and Structural EM algorithm based on simulations in log-likelihood (the difference between the log-likelihood of the estimated tree and the true tree) when $p = 50$. This figure summarizes the results from 100 replications, where the points represent the mean value and the bars are the Monte Carlo error bars. Note that points are shifted slightly horizontally to avoid overlaps.

all major mammalian lineages and birds, and the organs include brain, cerebellum, heart, kidney, liver and testis. Each species may include several individuals. They obtained standard expression values (RPKM), which were normalized across species and tissues with a median-scaling procedure on the basis of rank-conserved genes. Details about the data can be found in Brawand et al. (2011).

We first reconstruct the phylogenetic tree using gene expression data from the brain. There are 31 individuals included in this study, with each individual has 5,636 gene expression levels. We compare the tree constructed by independence test method, sample covariance matrix based method, Structural EM algorithm and neighbor-joining. Note that neighbor-

joining is the method that was used in Brawand et al. (2011), where the author used the Spearman's correlation matrix as the distance matrix. The results are summarized in Figure 4.7. As we can see, all four methods successfully separate primates from non-primates. The independence test method and the sample covariance based method can basically cluster individuals from the same species together, but both cluster some individuals from different species together. The Structural EM algorithm improves upon them. The tree estimated from the Structural EM algorithm is consistent with the known phylogeny of these species (Brawand et al., 2011): First, the only bird (chicken) is separated from the mammals in the first divergence event. Then monotreme (platypus) and marsupial (opossum) are separated from eutherians (the rest) in the second event. Then the two eutherian lineages (primates and rodents) are separated, and great apes are clustered together with an exclusion of the macaque, which is a type of monkey from the Old World. Although the neighbor-joining method gives similar results in terms of clustering the species together, the interpretation is not as clear as the results from the Structural EM as the tree obtained from the neighbor-joining method is actually unrooted and the placement of the internal nodes is arbitrary in the plot.

While usually different phylogenetic trees are reconstructed for different organs, we experiment here with pooling the data of the other five organs together and applying the Structural EM algorithm to construct the phylogenetic tree. Only primates are included in this analysis to make the tree smaller. In total, 56 individuals are included. The results are summarized in Figure 4.8. According to the tree, the events that separates the organs occur earlier than the divergence events. The testis is separated first, and then the cerebellum. After that, kidney, heart and liver are separated. Within each subtree of organ, species are well separated similar to the results from previous analysis.

4.6 Estimation Error of the EM Algorithm

In this section, we discuss the estimation error of the EM algorithm on Brownian motion tree model. We focus on the case where the true tree topology is known, and the EM algorithm

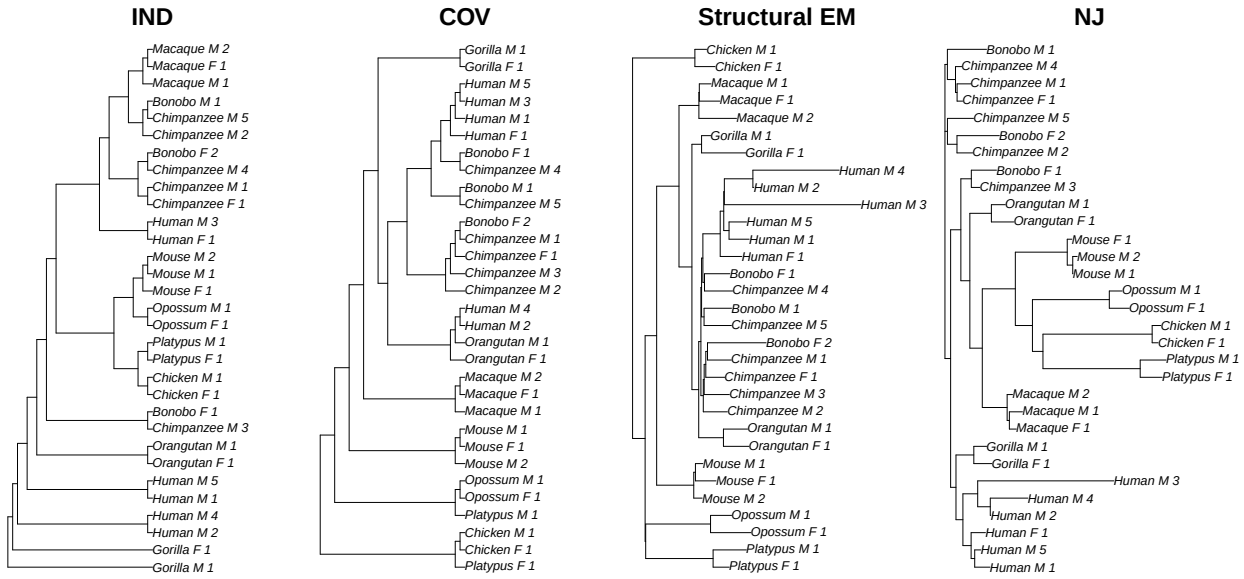

Figure 4.7: Comparison of independence test (IND), sample covariance matrix based method (COV), Structural EM algorithm (Structural EM) and neighbor joining method (NJ) in reconstructing phylogenetic tree using gene expression data from the brain.

is used to estimate the parameter values. We assume that $x_0 = 0$ to simplify the problem. We study the error bound of the estimate of branch length given by EM algorithm.

The EM algorithm has been one of the most popular approaches in obtaining the maximum likelihood estimate for models with latent variables. However, despite its wide application and simple implementation, the convergence of the EM algorithm can be complicated to study. Wu (1983) first established general sufficient conditions for the convergence of the EM algorithm, where he showed the EM algorithm converges if the likelihood is unimodal and some regularity conditions hold. Other significant work about the convergence of the EM algorithm includes Tseng (2004); McLachlan and Krishnan (2007); Wang et al. (2014); Balakrishnan et al. (2017), to name a few. In particular, a recent work from Balakrishnan et al. (2017) established results about the contractiveness in the steps of the EM algorithm and about its estimation error, and applied their results on several latent models, including

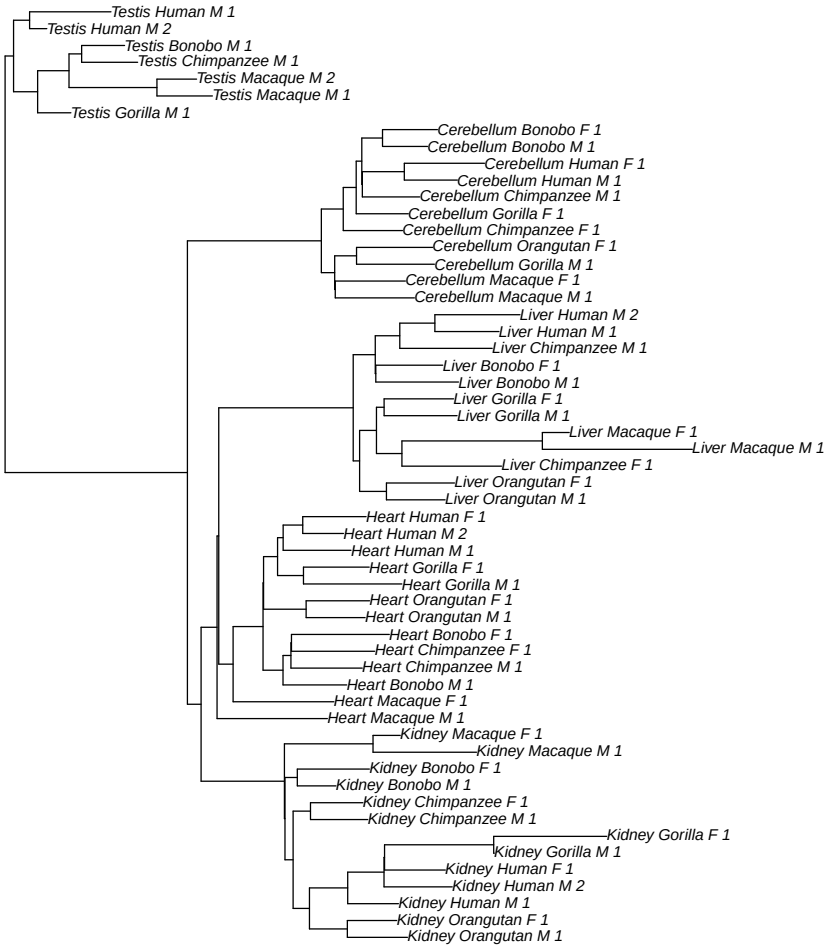

Figure 4.8: Phylogenetic tree of different organs of different species.

the symmetric mixture of two Gaussians, the symmetric mixture of two regressions and the linear regression with covariates missing completely at random. Later in Wang et al. (2014), the theory is extended to high-dimensional situations with a sparse parameters vector.

Here we first review the framework of Balakrishnan et al. (2017) and then consider the problem of extending the framework to phylogenetic tree estimation when both the number of leaf nodes p and the number of observations n are large. Let Y and Z be random variables in sample spaces $\mathcal{Y}$ and $\mathcal{Z}$, respectively, with a joint density $f_{\theta^*}(y, z)$ that belongs to a

parametric family $\{f_\theta(y,z)|\theta \in \Omega\}$. Suppose Z is the latent variable, and let $k_\theta(z|y)$ denote the conditional density of Z given Y. Our goal is to estimate θ^*. The sample conditional likelihood that we maximize in the EM algorithm is

$$Q_n(\theta|\theta') = \frac{1}{n}\sum_{i=1}^{n}(\int_{\mathcal{Z}} k_{\theta'}(z|y_i)\log(f_\theta(y_i,z))dz)$$

where θ is the parameter that we maximize over and θ' is the parameter that is conditional on, i.e., the parameter estimate from last iteration of EM algorithm. We then denote the population conditional likelihood as

$$Q(\theta|\theta') = \int_{\mathcal{Y}}(\int_{\mathcal{Z}} k_{\theta'}(z|y)\log(f_\theta(y,z))dz)g_{\theta^*}(y)dy.$$

This can be seen as the conditional likelihood we would have for an infinite number of samples. Define the operator $M : \Omega \to \Omega$,

$$M(\theta) = \arg\max_{\theta\in\Omega} Q(\theta|\theta').$$

This operator is the update function for the population version of EM iteration when inputting θ, i.e. in the t^{th} step we have $\theta^{t+1} = M(\theta^t)$. Balakrishnan et al. (2017) gave conditions under which the population EM operator moves the input parameter closer to the true parameter.

Theorem 4.6.1. *(Balakrishnan et al., 2017) Suppose that the function $Q(.|\theta^*)$ is globally $\lambda-$strongly concave, and the first-order stability condition holds with parameter γ in a ball $\mathbb{B}_2(\theta^*, r)$ with $r > 0$, i.e.,*

$$||\nabla Q(M(\theta)|\theta^*) - \nabla Q(M(\theta)|\theta)||_2 \leq \gamma||\theta - \theta^*||.$$

If $0 \leq \gamma < \lambda$, we have

$$||M(\theta) - \theta^*||_2 \leq \kappa||\theta - \theta^*||_2, \tag{4.15}$$

with $\kappa = \frac{\gamma}{\lambda}$, and we call M a contractive operator with parameter κ.

In applications, we use the sample EM algorithm instead of its population version, and the sample-based operator is defined as

$$M_n(\theta) = \arg\max_{\theta \in \Omega} Q_n(\theta|\theta').$$

Ideally we want this sample-based operator to be as close as possible to the population-based operator. Given $\delta \in (0,1)$, let $\epsilon(n,\delta)$ be the smallest scalar such that with probability at least $1-\delta$,

$$\sup_{\theta \in \mathbb{B}_2(r,\theta^*)} ||M_n(\theta) - M(\theta)||_2 \leq \epsilon(n,\delta).$$

The following theorem gives an error bound for the sample-based EM algorithm.

Theorem 4.6.2. *(Balakrishnan et al., 2017) Suppose that the population EM operator is contractive with parameter $\kappa \in (0,1)$ on the ball $\mathbb{B}_2(\theta^*, r)$, and the initial parameter θ^0 is in $\mathbb{B}_2(\theta^*, r)$. If*

$$\epsilon(n,\delta) \leq (1-\kappa)r,$$

then the estimate θ^t from the t^{th} iteration satisfies

$$||\theta^t - \theta^*||_2 \leq \kappa^t ||\theta^0 - \theta^*||_2 + \frac{1}{1-\kappa}\epsilon(n,\delta)$$

with probability at least $1-\delta$.

This theorem tells us that if the number of iterations is large enough, the EM algorithm achieves a statistical convergence rate of the order of $\epsilon(n,\delta)$. For the symmetric mixture of two Gaussians, the symmetric mixture of two regressions and the linear regression with covariates missing completely at random, the estimation error are of order $\sqrt{p/n}$.

Ideally we would hope that these results apply to Brownian motion models for phylogenetic trees. However, our tree model setting is quiet different from the examples given in Balakrishnan et al. (2017) in the sense that we are estimating variance parameters in our tree model while in their examples mean parameter estimation is the goal. Numerical examples show that the contractive condition in (4.15) does not hold for the Brownian motion model

for phylogenetic trees. First we notice that L_2 norm may not be a appropriate norm to measure the distance between the estimated branch length $\hat{\boldsymbol{d}}$ and the true branch length $\boldsymbol{d}^*$ since the L_2 norm weights all the entries in a vector the same, making larger entries in the vector more influential. We suggest to use the weighted norm proposed in Section 4.3.2 as the measure to quantify the error.

The sample-based EM operator in the EM algorithm for phylogenetic tree is

$$M_n(d_{ij}) = \frac{1}{n}\sum_{k=1}^{n} \mathrm{E}[(X_i^{(k)} - X_j^{(k)})^2 | O, T, \boldsymbol{d}] \tag{4.16}$$

for any $(i,j) \in E$. Note that generally M_n is an operator on the entire $\boldsymbol{d}$ vector, here we write $M_n(d_{ij})$ as the updated element that replaces d_{ij} after an EM iteration. Similarly, the population-based EM operator is

$$M(d_{ij}) = \mathrm{E}[\mathrm{E}[(X_i - X_j)^2 | O, T, \boldsymbol{d}] | \boldsymbol{d}^*] \tag{4.17}$$

for any $(i,j) \in E$.

We first examine whether the population-based EM operator is contractive or not. Our goal is to show that

$$||M(\boldsymbol{d}) - \boldsymbol{d}^*||_{2,d^*} < \kappa ||\boldsymbol{d} - \boldsymbol{d}^*||_{2,d^*} \tag{4.18}$$

for $\kappa \in (0,1)$ when $\boldsymbol{d} \in \mathbb{B}_{2,d^*}(\boldsymbol{d}^*, r)$. Note that $M(\boldsymbol{d}^*) = \boldsymbol{d}^*$, hence by Mean-value Theorem we have

$$\begin{aligned} ||M(\boldsymbol{d}) - \boldsymbol{d}^*||_{2,d^*} &= ||M(\boldsymbol{d}) - M(\boldsymbol{d}^*)||_{2,d^*} \\ &\leq \sqrt{\lambda}||\boldsymbol{d} - \boldsymbol{d}^*||_{2,d^*}. \end{aligned} \tag{4.19}$$

Here λ is the maximum eigenvalue of $D^2 H^T (D^2)^{-1} H$ with H being the Jacobian matrix $\frac{\partial M(\boldsymbol{d}')}{\partial \boldsymbol{d}'}$, for $\boldsymbol{d}'$ in the interval between $\boldsymbol{d}$ and $\boldsymbol{d}^*$. Meanwhile, $D = diag(\boldsymbol{d}^*)$ is the diagonal matrix with $\boldsymbol{d}^*$ being the diagonal elements. Hence if $\lambda < 1$ at the point $\boldsymbol{d} = \boldsymbol{d}^*$, then $\exists r > 0$ and $\kappa \in (\sqrt{\lambda}, 1)$ such that (4.18) holds in a ball $\mathbb{B}_{2,d^*}(\boldsymbol{d}^*, r)$.

Now we derive the form of the H matrix. First, by some simplification we have

$$\begin{aligned} M(d_{ij}) &= d_{ij} B_{.j}^T \Sigma_{oo}^{-1}(\Sigma_{oo}^* \Sigma_{oo}^{-1} - I) B_{.j} + d_{ij}, \qquad j \text{ is an internal node,} \\ M(d_{ij}) &= \Sigma_{io}\Sigma_{oo}^{-1}(\Sigma_{oo}^*\Sigma_{oo}^{-1} - I)\Sigma_{oi} + \Sigma_{jj}^* + \Sigma_{ii} - 2\Sigma_{io}\Sigma_{oo}^{-1}\Sigma_{oj}^*, \qquad j \text{ is a leaf node.} \end{aligned} \tag{4.20}$$

Lemma 4.6.3. *The Jacobian matrix $H = \frac{\partial M(\boldsymbol{d})}{\partial \boldsymbol{d}}$ for $M(\boldsymbol{d})$ that is defined in (4.20) is*

$$H = I_{2p-1} - A \circ A \tag{4.21}$$

with $A = DB^T(BDB^T)^{-1}B$, when $\boldsymbol{d} = \boldsymbol{d}^*$. Here $\circ$ is the Hadamard product operator.

We now show that the algorithm is contractive at the point $\boldsymbol{d}^*$. To do this, we need to analyze the spectral norm of the following matrix

$$\begin{aligned} D^2 H^T (D^2)^{-1} H &= D^2 (I - A \circ A)^T (D^2)^{-1} (I - A \circ A) \\ &= (I - A \circ A)(I - D^2 (A \circ A)^T (D^2)^{-1}) \\ &= (I - A \circ A)(I - D^2 (A^T \circ A^T)(D^2)^{-1}) \\ &= (I - A \circ A)(I - (D^2 B^T (BDB^T)^{-1} BD) \circ (B^T (BDB^T)^{-1} BD (D^2)^{-1})) \\ &= (I - A \circ A)(I - A \circ A). \end{aligned} \tag{4.22}$$

The follows result shows that the spectral norm of (4.22) is smaller than 1.

Proposition 4.6.1. *The spectral norm of $D^2 H^T (D^2)^{-1} H$ is in $[0, 1)$, i.e. the EM algorithm with Brownian motion model is contractive at $\boldsymbol{d}^*$. Here H is defined in Lemma 4.6.3 and D is the diagonal matrix with $\boldsymbol{d}$ being the diagonal elements.*

Note that the results in Lemma 4.6.3 and Proposition 4.6.1 do not assume the tree to be bifurcating. Here in Example 4.6.1, we show that for a special case of a star tree, we can derive the contractive coefficient and verify that for any $p > 1$, this contractive coefficient is smaller than 1 at $\boldsymbol{d}^*$.

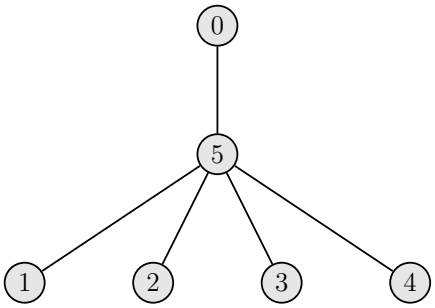

Figure 4.9: A star tree with four nodes.

Example 4.6.1. *A star tree is a tree that has one internal node, and all the leaf nodes are its descendants. Figure 4.9 shows an example of a star tree with four leaf nodes. The B matrix for star tree topology is $B = (I_p, \mathbf{1}_p)$. Now using Lemma 4.6.3, and assume the branch lengths are all equal, we have*

$$H = \frac{1}{(p+1)^2}((2p+2)I_{p+1} - \mathbf{1}_{p+1}\mathbf{1}_{p+1}^T).$$

Therefore

$$H^T H = \frac{1}{(p+1)^4}(I_{p+1} - \frac{3}{4(p+1)}\mathbf{1}_{p+1}\mathbf{1}_{p+1}^T).$$

The eigenvalues for this matrix are $\frac{4}{(p+1)^2}$ and $\frac{1}{(p+1)^2}$. Therefore, when $p > 1$, the largest eigenvalue is smaller than 1.

We now discuss the statistical error of the sample-based operator as compared to the population-based operator as p and n change. For $\epsilon > 0$ we have

$$\begin{aligned}
\Pr(||M_n(\boldsymbol{d}) - M(\boldsymbol{d})||_{2,d^*} > \epsilon) &= \Pr\left(\sum_{(i,j)\in E} |M_n(d_{ij}) - M(d_{ij})|^2/(d_{ij}^*)^2 > \epsilon^2\right) \\
&\leq \Pr\left(\bigcup_{(i,j)\in E} \{|M_n(d_{ij}) - M(d_{ij})|^2/(d_{ij}^*)^2 > \epsilon^2/p\}\right) \\
&\leq \sum_{(i,j)\in E} \Pr\left(|M_n(d_{ij}) - M(d_{ij})|^2/(d_{ij}^*)^2 > \epsilon^2/p\right) \\
&\leq \sum_{(i,j)\in E} \Pr\left(|M_n(d_{ij}) - M(d_{ij})| > \epsilon d_{ij}^*/\sqrt{p}\right).
\end{aligned} \tag{4.23}$$

Now examine the error bound for $|M_n(d_{ij}) - M(d_{ij})|$. Similar to (4.20), we have

$$\begin{aligned} M_n(d_{ij}) &= d_{ij} B_{\cdot j}^T \Sigma_{oo}^{-1}(S_{oo}\Sigma_{oo}^{-1} - I)B_{\cdot j} + d_{ij}, \qquad j \text{ is an internal node,} \\ M_n(d_{ij}) &= \Sigma_{io}\Sigma_{oo}^{-1}(S_{oo}\Sigma_{oo}^{-1} - I)\Sigma_{oi} + S_{jj} + \Sigma_{ii} - 2\Sigma_{io}\Sigma_{oo}^{-1}S_{oj}, \qquad j \text{ is a leaf node.} \end{aligned} \tag{4.24}$$

where S is the sample covariance matrix. Hence by (4.20) and (4.24), $M_n(d_{ij}) - M(d_{ij})$ can be written as a linear combination of elements in $\Sigma_{oo}^* - S_{oo}$. We can write

$$|M_n(d_{ij}) - M(d_{ij})| = \left| \sum_{a,b \in \{1,\ldots,p\}} c_{ab}\Sigma_{ab}^* - S_{ab} \right|,$$

with c_{ab}'s being constants that depends on $\boldsymbol{d}$. Therefore by Lemma 4.8.1 we have, for $\epsilon_2 > 0$

$$\begin{aligned} \Pr\left(|M_n(d_{ij}) - M(d_{ij})| > \epsilon_2\right) &\le \Pr\left(\sum_{a,b \in \{1,\ldots,p\}} |c_{ab}||\Sigma_{ab}^* - S_{ab}| > \epsilon_2 \right) \\ &\le p^2 \Pr\left(|\Sigma_{ab}^* - S_{ab}| > \epsilon_2/(cp^2)\right) \\ &\le p^2 \exp\left(-\frac{n\epsilon_2^2}{3200p^4c^2 \max_i(\Sigma_{ii}^*)^2} \right), \end{aligned} \tag{4.25}$$

where $c = \max_{a,b \in \{1,\ldots,p\}} c_{ab}$. Thus by (4.23) we have

$$\begin{aligned} \Pr\left(||M_n(\boldsymbol{d}) - M(\boldsymbol{d})||_{2,d^*} > \epsilon\right) &\le \sum_{(i,j)\in E} 4p^2 \exp\left(-\frac{n\epsilon^2 d_{ij}^2}{3200p^4c^2 \max_i(\Sigma_{ii}^*)^2} \right) \\ &\le 4p^4 \exp\left(-\frac{n\epsilon^2 d^2}{3200p^4c^2 \max_i(\Sigma_{ii}^*)^2} \right) \\ &\le 4 \exp\left(4\log(p) - \frac{n\epsilon^2 d^2}{3200p^4c^2 \max_i(\Sigma_{ii}^*)^2} \right). \end{aligned} \tag{4.26}$$

Note that this error bound is derived using the $||\cdot||_{2,d^*}$ norm, which sums all the entry-wise statistical estimation errors together, while the $||\cdot||_{\infty,d^*}$ norm characterizes the entry-wise statistical estimation error. Hence for the same n, using $||\cdot||_{2,d^*}$ norm requires larger p to have the same probability of controlling the error under ϵ as compared to using $||\cdot||_{\infty,d^*}$ norm. However, according to our numerical evaluation, the contractive condition does not hold for the $||\cdot||_{\infty,d^*}$ norm. A potential reason is that the weighted maximum norm ball is too big, making it harder for the algorithm to be contractive in such a big ball.

Although we have shown that the EM algorithm is contractive in a ball $\mathbb{B}_{2,d^*}(\boldsymbol{\theta}^*, r)$ for fixed n and p, we have not calculated r as a function of p and n. Meanwhile, in order to further establish the consistency of the EM algorithm in the scenario when both p and n are growing, more work on how the tree structure changes as p changes is needed. We hope our discussion here can provide more insight about the consistency of the structural EM algorithm in high-dimensional scenarios.

4.7 Non-distinguishable Tree Topologies Based on Correlation Matrix

When there are three leaf nodes, the only two possible types of tree topology, $\mathcal{T}$ and $\mathcal{T}'$, under our settings are shown in Figure 4.2. Given a correlation matrix

$$R = \begin{bmatrix} 1 & \rho_{12} & \rho_{13} \\ \rho_{12} & 1 & \rho_{23} \\ \rho_{13} & \rho_{23} & 1 \end{bmatrix}$$

with $0 < \rho_{ij} < 1 \ \forall i < j, i, j \in \{1, 2, 3\}$, and the product of any two in $\{\rho_{ij} | i < j, i, j \in \{1, 2, 3\}\}$ smaller than the third, we can obtain two sets of parameters $\boldsymbol{d} = \{d_1, d_2, d_3, d_4, d_5\}$ and $\boldsymbol{d}' = \{d'_1, d'_2, d'_3, d'_4, d'_5\}$ such that both $\mathcal{T} = (T, \boldsymbol{d})$ and $\mathcal{T}' = (T', \boldsymbol{d}')$ give the same correlation matrix R. To achieve this, we simply pick the two sets of parameterization according to the following rules

$$\begin{aligned}
d_1 &= (d_4 + d_5)(\frac{\rho_{23}}{\rho_{12}\rho_{13}} - 1), & d'_1 &= d'_5(\frac{d'_5}{d'_5 + d'_4}\frac{\rho_{23}}{\rho_{12}\rho_{13}} - 1), \\
d_2 &= (d_4 + d_5)(\frac{\rho_{13}}{\rho_{12}\rho_{23}} - 1), & d'_2 &= (d'_4 + d'_5)(\frac{\rho_{13}}{\rho_{12}\rho_{23}} - 1), \\
d_3 &= d_5(\frac{d_5}{d_5 + d_4}\frac{\rho_{12}}{\rho_{13}\rho_{23}} - 1), & d'_3 &= (d'_4 + d'_5)(\frac{\rho_{12}}{\rho_{13}\rho_{23}} - 1), \\
\frac{d_5}{d_5 + d_4} &\geq \frac{\rho_{13}\rho_{23}}{\rho_{12}}, & \frac{d'_5}{d'_5 + d'_4} &\geq \frac{\rho_{13}\rho_{23}}{\rho_{12}},
\end{aligned}$$

and then $\mathcal{T}$ and $\mathcal{T}'$ will imply the same correlation matrix R. To be specific, for $\boldsymbol{d}$, we only need to pick $d_4 > 0$ and $d_5 > 0$ such that $\frac{d_5}{d_5+d_4} \geq \frac{\rho_{13}\rho_{23}}{\rho_{12}}$, and then use these two values to calculate d_1, d_2 and d_3. A similar calculation can be done for $\boldsymbol{d}'$. For tree with more leaf

nodes, we can use similar method to show that we can not distinguish different types of tree topology based on correlation matrix only. For example, when $p = 4$, the two trees in Example 4.10 can not be distinguished using correlation matrix only.

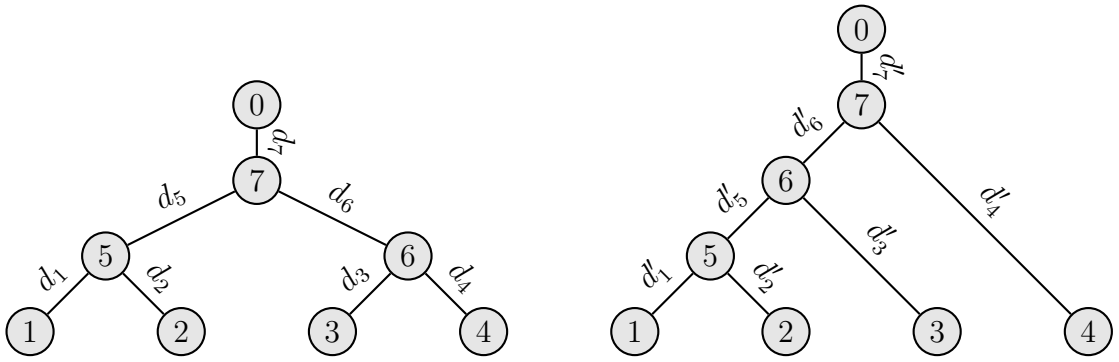

Figure 4.10: Two trees $\mathcal{T} = (T, \boldsymbol{d})$ (left) and $\mathcal{T}' = (T', \boldsymbol{d}')$ (right) that are not distinguishable only based on correlation matrix.

4.8 Discussion

In this chapter, we present several approaches for phylogenetic tree inferences under the Brownian motion model. The first approach is based on independence tests. We test the independence relationships among leaf nodes and then infer the tree topologies based on that. The second method is based on sample covariances. We use sample covariance matrix as a distance matrix and use a top-down procedure to reconstruct the tree. These two methods are simple and fast, especially the sample covariance matrix based method. They are developed specifically under the Brownian motion model, making them good initial starts for the Structural EM algorithm.

Our Structural EM algorithm for maximum likelihood phylogenetic tree inferences builds on the existing algorithm but modifies it for continuous data. Each step of our algorithm includes maximization over the topology, the branch length and the root state. For applications, this methods shows good performance in both simulated data and real data. Com-

paring to our other two simple methods, Structural EM algorithm shows improvements over them in terms of topological distance to the true tree and likelihood in simulation studies.

The EM algorithm is a popular approach in obtaining the maximum likelihood estimate for models with latent variables. However, the convergence property of the EM algorithm is complicated to study. Recently, Balakrishnan et al. (2017) first established the results about the convergence rate of the EM algorithm and applied their results on several latent models. We speculate that our Structural EM algorithm is also consistent in the scenario when both p and n are growing. A simpler but still complicated objective is to show our algorithm gives consistent estimates for the branch lengths when the tree structure is given. The estimation error of the EM algorithm output can be decomposed into two major source, the optimization error and the statistical error. The optimization error will diminish as the number of iterations grow if the algorithm is contractive, i.e. at each step, the algorithm improves the accuracy of the estimate. The statistical error measures the stochastic difference between the output of the algorithm using observations from a finite sample and the output of the algorithm using the population, which diminishes as a function of p and n. We discussed this topic in Section 4.6 and established the contractiveness of the EM algorithm in Brownian tree parameter estimation, which helps us characterize the behavior the EM algorithm for fixed p and n. We leave the characterization of the behavior of the EM algorithm when both p and n are growing as future work.

Appendix: Proofs

Proof of Proposition 4.3.1. Suppose nodes in $C_1 \cup C_2$ form a subtree of $\mathcal{T}$. We have that $\forall k \in \{1, \ldots, p\} \backslash \{C_1 \cup C_2\}$, $Cov(X_i, X_k) = Cov(X_j, X_k)$ since both of them are equal to the covariance between X_k and the root of the subtree. Thus $Cov(X_i - X_j, X_k) = 0$, and therefore $X_i - X_j \perp\!\!\!\perp X_k$ since $(X_1, \ldots, X_p)$ follows a multivariate Gaussian distribution.

Suppose we have $X_i - X_j \perp\!\!\!\perp X_k$ for $\forall i \in C_1, j \in C_2, k \in \{1, \ldots, p\} \backslash \{C_1 \cup C_2\}$. Suppose $\mathcal{T}^*$ is the smallest subtree that contains $C_1 \cup C_2$. We now show that $\forall k \in \{1, \ldots, p\} \backslash \{C_1 \cup C_2\}$, $k \notin \mathcal{T}^*$. If there is one node $k \in \{1, \ldots, p\} \backslash \{C_1 \cup C_2\}$ in $\mathcal{T}^*$, k must be in some subtree

$\mathcal{T}^{**}$ of $\mathcal{T}^*$ together with some node $i \in C_1 \cup C_2$, while some node $j \in C_1 \cup C_2$ is not in $\mathcal{T}^{**}$, or $\mathcal{T}^*$ will not be the smallest subtree that contains $C_1 \cup C_2$. This implies that $Cov(X_i, X_k) \neq Cov(X_j, X_k)$. Thus $\forall k \in \{1, \ldots, p\} \backslash \{C_1 \cup C_2\}$, $k \notin \mathcal{T}^*$. Therefore nodes in set C_1 and nodes in set C_2 form a subtree of $\mathcal{T}$.

□

We need the following lemma from Ravikumar et al. (2011) before proving Proposition 4.3.2.

Lemma 4.8.1. *(Ravikumar et al., 2011) Consider a zero-mean random vector $(X_1, \ldots, X_p)$ with covariance Σ^* such that each $X_i/\sqrt{\Sigma^*_{ii}}$ is sub-Gaussian with parameter σ. Given n i.i.d. samples, the associated sample covariance $\hat{\Sigma}^n$ satisfies the tail bound*

$$\Pr\left(|\hat{\Sigma}^n_{ij} - \Sigma^*_{ij}| > \epsilon\right) \leq 4 \exp\left(-\frac{n\epsilon^2}{128(1+4\sigma^2)^2 \max_i(\Sigma^*_{ii})^2}\right), \tag{4.27}$$

for all $\epsilon \in (0, \max_i(\Sigma^*_{ii})8(1+4\sigma^2))$.

Proof of Proposition 4.3.2. We first show that, if we use the population covariance matrix as input in algorithm 2, we can recover the true tree topology. Essentially, we want to show that at each iteration, when given a set of leaf nodes A of a subtree $\mathcal{T}_A$, algorithm 2 correctly split them into two disjoint subsets A_1 and A_2 that belong to two disjoint subtrees $\mathcal{T}_{A_1}$ and $\mathcal{T}_{A_2}$. Since in the Brownian motion tree model, the more common ancestors two nodes share, the larger the covariance between them is. Hence among leaf nodes in A, the smallest covariance is the covariance between two nodes from two separate subtrees. Therefore for $(i^*, j^*) = \arg\min \Sigma_{ij}$, i^* and j^* must from two subtrees. For any other node in A, if its covariance with i^* is larger than that with j^*, then it is in the same subtree as i^*, otherwise it is in the same subtree as j^*. Hence by this method, we can split leaf nodes in A into two disjoint subsets A_1 and A_2 that belong to two disjoint subtrees.

In order to recover the tree topology using algorithm 2 with sample covariance components, we need the order of the sample covariance components to be the same as that of the true covariance components, i.e. $\Sigma_{ij} \leq \Sigma_{i'j'}$ if and only if $S_{ij} \leq S_{i'j'}$. Since the smallest

branch length is greater than d, if we have $|S_{ij} - \Sigma_{ij}| < d/2$ for all $i, j \in \{1, \ldots, p\}$, then the order of S_{ij}'s and the order of the Σ_{ij}s are the same. Therefore

$$\begin{aligned}
\Pr(\hat{T} \neq T) &\leq \Pr\left(\cup_{i \neq j \in \{1,\ldots,p\}} \{|S_{ij} - \Sigma_{ij}| \geq d/2\}\right) \\
&\leq \sum_{i \neq j \in \{1,\ldots,p\}} \Pr\left(|S_{ij} - \Sigma_{ij}| \geq d/2\right) \\
&\leq 2p(p-1) \exp\left(-\frac{n\epsilon^2 d^2}{12800 \max_i (\Sigma^*_{ii})^2}\right) \\
&= 2 \exp\left(\log(p(p-1)) - \frac{n\epsilon^2 d^2}{12800 \max_i (\Sigma^*_{ii})^2}\right).
\end{aligned}$$

Similarly, given true tree topology, we have

$$\begin{aligned}
\Pr\left(\max_{i \in \{1,\ldots,2p-1\}} |\frac{\hat{d}_i}{d_i} - 1| > \epsilon\right) &= \Pr\left(\cup_{i \in \{1,\ldots,2p-1\}} \{|\frac{\hat{d}_i}{d_i} - 1| > \epsilon\}\right) \\
&\leq \sum_{i \in \{1,\ldots,2p-1\}} \Pr\left(|\hat{d}_i - d_i| \geq \epsilon d_i\right) \\
&\leq \sum_{i \in \{1,\ldots,2p-1\}} \Pr\left(|\hat{d}_i - d_i| \geq \epsilon d\right) \\
&\leq (2p-1) \Pr\left(\{|(S_{ii} - S_{ki}) - (\Sigma_{ii} - \Sigma_{ki})| \geq \epsilon d\}\right) \\
&\leq (2p-1) \Pr\left(\{|S_{ii} - \Sigma_{ii}| + |S_{ki} - \Sigma_{ki}| \geq \epsilon d\}\right) \\
&\leq (2p-1) \Pr\left(\{|S_{ii} - \Sigma_{ii}| \geq \epsilon d/2\} \cup \{|\hat{S}_{ki} - \Sigma_{ki}| \geq \epsilon d/2\}\right) \\
&\leq 8(2p-1) \exp\left(-\frac{n\epsilon^2 d^2}{12800 \max_i (\Sigma^*_{ii})^2}\right) \\
&\leq 8 \exp\left(\log(2p) - \frac{n\epsilon^2 d^2}{12800 \max_i (\Sigma^*_{ii})^2}\right),
\end{aligned}$$

where the third inequality is obtained by considering the worst case of estimation where we estimate the branch $k \to i$ when i is a leaf node.

□

Proof of Lemma 4.6.3. When j is an internal node,

$$\begin{aligned}\frac{\partial M(d_{ij})}{\partial d_{ij}} =&2d_{ij}B_{.j}^T\Sigma_{oo}^{-1}(\Sigma_{oo}^*\Sigma_{oo}^{-1}-I)B_{.j}+d_{ij}^2B_{.j}^T\frac{\partial\Sigma_{oo}^{-1}}{\partial d_{ij}}(\Sigma_{oo}^*\Sigma_{oo}^{-1}-I)B_{.j}+\\ &d_{ij}^2B_{.j}^T\Sigma_{oo}^{-1}(\Sigma_{oo}^*\frac{\partial\Sigma_{oo}^{-1}}{\partial d_{ij}})B_{.j}+1,\\ \frac{\partial M(d_{ij})}{\partial d_{kl}} =&d_{ij}^2B_{.j}^T\frac{\partial\Sigma_{oo}^{-1}}{\partial d_{kl}}(\Sigma_{oo}^*\Sigma_{oo}^{-1}-I)B_{.j}+d_{ij}^2B_{.j}^T\Sigma_{oo}^{-1}(\Sigma_{oo}^*\frac{\partial\Sigma_{oo}^{-1}}{\partial d_{kl}})B_{.j}.\end{aligned} \tag{4.28}$$

When j is an leaf node,

$$\begin{aligned}\frac{\partial M(d_{ij})}{\partial d_{ij}} =&\Sigma_{io}\frac{\partial\Sigma_{oo}^{-1}}{\partial d_{ij}}(\Sigma_{oo}^*\Sigma_{oo}^{-1}-I)\Sigma_{oi}+\Sigma_{io}\Sigma_{oo}^{-1}(\Sigma_{oo}^*\frac{\partial\Sigma_{oo}^{-1}}{\partial d_{ij}})\Sigma_{oi}-2\Sigma_{io}\frac{\partial\Sigma_{oo}^{-1}}{\partial d_{ij}}\Sigma_{oj}^*,\\ \frac{\partial M(d_{ij})}{\partial d_{kl}} =&\frac{\partial\Sigma_{io}\Sigma_{oo}^{-1}}{\partial d_{kl}}(\Sigma_{oo}^*\Sigma_{oo}^{-1}-I)\Sigma_{oi}+\Sigma_{io}\Sigma_{oo}^{-1}(\Sigma_{oo}^*\frac{\partial\Sigma_{oo}^{-1}}{\partial d_{kl}})\Sigma_{oi}+\frac{\partial\Sigma_{ii}}{\partial d_{kl}}\\ &-2\frac{\partial\Sigma_{io}}{\partial d_{kl}}\Sigma_{oo}^{-1}\Sigma_{oj}^*-2\Sigma_{io}\frac{\partial\Sigma_{oo}^{-1}}{\partial d_{ij}}\Sigma_{oj}^*.\end{aligned} \tag{4.29}$$

Now if $\Sigma=\Sigma^*$, we have that (4.28) is simplified to

$$\begin{aligned}\frac{\partial M(d_{ij})}{\partial d_{ij}} &= 1+d_{ij}^2B_{.j}^T\frac{\partial\Sigma_{oo}^{-1}}{\partial d_{ij}}B_{.j},\\ \frac{\partial M(d_{ij})}{\partial d_{kl}} &= d_{ij}^2B_{.j}^T\frac{\partial\Sigma_{oo}^{-1}}{\partial d_{ij}}B_{.j},\end{aligned} \tag{4.30}$$

and (4.29) is simplified to

$$\begin{aligned}\frac{\partial M(d_{ij})}{\partial d_{ij}} &= \Sigma_{io}\frac{\partial\Sigma_{oo}^{-1}}{\partial d_{ij}}-2\Sigma_{io}\frac{\partial\Sigma_{oo}^{-1}}{\partial d_{ij}}\Sigma_{oj},\\ \frac{\partial M(d_{ij})}{\partial d_{kl}} &= \Sigma_{io}\frac{\partial\Sigma_{oo}^{-1}}{\partial d_{kl}}-2\Sigma_{io}\frac{\partial\Sigma_{oo}^{-1}}{\partial d_{kl}}\Sigma_{oj}-2\frac{\partial\Sigma_{io}}{\partial d_{kl}}\Sigma_{oo}^{-1}\Sigma_{oj}+1(l\in\mathrm{An}(i)).\end{aligned} \tag{4.31}$$

Note that $\Sigma_{oo}=BDB^T$, we have

$$\frac{\partial\Sigma_{oo}^{-1}}{\partial d_{kl}}=-\Sigma_{oo}^{-1}\frac{\partial\Sigma_{oo}}{\partial d_{kl}}\Sigma_{oo}^{-1}=-\Sigma_{oo}^{-1}(B\frac{\partial D}{\partial d_{kl}}B^T)\Sigma_{oo}^{-1}=-\Sigma_{oo}^{-1}(B_{.l}B_{.l}^T)\Sigma_{oo}^{-1}. \tag{4.32}$$

By using (4.32) and (4.30), we have that when j is an internal node

$$\begin{aligned}\frac{\partial M(d_{ij})}{\partial d_{ij}} &= 1-d_{ij}^2(B_{.j}^T\Sigma_{oo}^{-1}B_{.j})^2,\\ \frac{\partial M(d_{ij})}{\partial d_{kl}} &= -d_{ij}^2(B_{.j}^T\Sigma_{oo}^{-1}B_{.l})^2,\end{aligned} \tag{4.33}$$

and by (4.32) and (4.31), when j is a leaf node

$$\begin{aligned}
\frac{\partial M(d_{ij})}{\partial d_{ij}} &= -(\Sigma_{io}\Sigma_{oo}^{-1}B_{\cdot j})^2 + 2\Sigma_{io}\Sigma_{oo}^{-1}(B_{\cdot j}B_{\cdot j}^T)\Sigma_{oo}^{-1}\Sigma_{oj}, \\
\frac{\partial M(d_{ij})}{\partial d_{kl}} &= -(\Sigma_{io}\Sigma_{oo}^{-1}B_{\cdot l})^2 + 2\Sigma_{io}\Sigma_{oo}^{-1}(B_{\cdot l}B_{\cdot l}^T)\Sigma_{oo}^{-1}\Sigma_{oj} - 2\frac{\partial \Sigma_{io}}{\partial d_{kl}}\Sigma_{oo}^{-1}\Sigma_{oj} + 1(l \in \mathrm{An}(i)).
\end{aligned} \tag{4.34}$$

Furthermore, by observing that

$$\Sigma_{io} = \Sigma_{jo} - d_{ij}B_{\cdot j} = B_{\cdot j}^T\Sigma_{oo} - d_{ij}B_{\cdot j}^T, \tag{4.35}$$

we have that when j is a leaf node

$$\begin{aligned}
\frac{\partial M(d_{ij})}{\partial d_{ij}} &= -((B_{\cdot j}^T\Sigma_{oo} - d_{ij}B_{\cdot j}^T)\Sigma_{oo}^{-1}B_{\cdot j})^2 + 2(B_{\cdot j}^T\Sigma_{oo} - d_{ij}B_{\cdot j}^T)\Sigma_{oo}^{-1}(B_{\cdot j}B_{\cdot j}^T)\Sigma_{oo}^{-1}\Sigma_{oo}B_{\cdot j} \\
&= -(1 - d_{ij}B_{\cdot j}^T\Sigma_{oo}^{-1}B_{\cdot j})^2 + 2(1 - d_{ij}B_{\cdot j}^T\Sigma_{oo}^{-1}B_{\cdot j}) \\
&= 1 - d_{ij}^2(B_{\cdot j}^T\Sigma_{oo}^{-1}B_{\cdot j})^2.
\end{aligned} \tag{4.36}$$

Meanwhile, we have

$$\begin{aligned}
&\frac{\partial \Sigma_{io}}{\partial d_{kl}}\Sigma_{oo}^{-1}\Sigma_{oj} = B_{\cdot l}\Sigma_{oo}^{-1}(\Sigma_{oo}B_{\cdot j}) = 1 \quad \text{and} \quad B_{\cdot j}^TB_{\cdot l} = 1 \qquad \text{when } l \in \mathrm{An}(i), \\
&\frac{\partial \Sigma_{io}}{\partial d_{kl}}\Sigma_{oo}^{-1}\Sigma_{oj} = 0 \quad \text{and} \quad B_{\cdot j}^TB_{\cdot l} = 0 \qquad \text{when } l \notin \mathrm{An}(i).
\end{aligned} \tag{4.37}$$

Hence, when j is a leaf node and $l \in \mathrm{An}(i)$, we have

$$\begin{aligned}
\frac{\partial M(d_{ij})}{\partial d_{kl}} &= -((B_{\cdot j}^T\Sigma_{oo} - d_{ij}B_{\cdot j}^T)\Sigma_{oo}^{-1}B_{\cdot l})^2 + 2(B_{\cdot j}^T\Sigma_{oo} - d_{ij}B_{\cdot j}^T)\Sigma_{oo}^{-1}(B_{\cdot l}B_{\cdot l}^T)\Sigma_{oo}^{-1}\Sigma_{oo}B_{\cdot j} - 1 \\
&= -(1 - d_{ij}B_{\cdot j}^T\Sigma_{oo}^{-1}B_{\cdot l})^2 + 2(1 - d_{ij}B_{\cdot j}^T\Sigma_{oo}^{-1}B_{\cdot l}) - 1 \\
&= -d_{ij}^2(B_{\cdot j}^T\Sigma_{oo}^{-1}B_{\cdot l})^2.
\end{aligned} \tag{4.38}$$

Similarly, when j is a leaf node and $l \notin \mathrm{An}(i)$, we have

$$\begin{aligned}
\frac{\partial M(d_{ij})}{\partial d_{kl}} &= -((B_{\cdot j}^T\Sigma_{oo} - d_{ij}B_{\cdot j}^T)\Sigma_{oo}^{-1}B_{\cdot l})^2 + 2(B_{\cdot j}^T\Sigma_{oo} - d_{ij}B_{\cdot j}^T)\Sigma_{oo}^{-1}(B_{\cdot l}B_{\cdot l}^T)\Sigma_{oo}^{-1}\Sigma_{oo}B_{\cdot j} \\
&= -d_{ij}^2(B_{\cdot j}^T\Sigma_{oo}^{-1}B_{\cdot l})^2.
\end{aligned} \tag{4.39}$$

Therefore, by (4.33), (4.38) and (4.39), $H = I - A \circ A$ with $A = DB^T\Sigma_{oo}^{-1}B = DB^T(BDB^T)^{-1}B$.

□

Proof of Proposition 4.6.1. By (4.22), we only need to show the eigenvalues of $I - A \circ A$ are in $[0, 1)$, which is equivalent to show the eigenvalues of $A \circ A$ are in $(0, 1]$. We have

$$\begin{aligned} A \circ A &= (DB^T(BDB^T)^{-1}B) \circ (DB^T(BDB^T)^{-1}B) \\ &= DD^{-1}((DB^T(BDB^T)^{-1}B) \circ (DB^T(BDB^T)^{-1}B))DD^{-1} \\ &= D((D^{1/2}B^T(BDB^T)^{-1}BD^{1/2}) \circ (D^{1/2}B^T(BDB^T)^{-1}BD^{1/2}))D^{-1} \\ &= D(P \circ P)D^{-1}, \end{aligned} \tag{4.40}$$

where $P = D^{1/2}B^T(BDB^T)^{-1}BD^{1/2}$. Hence, in order to show that the eigenvalues of $A \circ A$ are in $(0, 1]$, we only need to show that the eigenvalues of $(P \circ P)$ are in $(0, 1]$.

We first show that $\rho(P \circ P) \leq 1$ with ρ being the spectral norm operator. Note that P is a symmetric projection matrix, hence it is positive semidefinite with eigenvalues either 0 or 1. By Theorem 5.3.4 from Hom and Johnson (1991), we have

$$\rho(P \circ P) \leq \rho(P)\rho(P) \leq 1.$$

In order to show the eigenvalues of $P \circ P$ are positive, we need to show $P \circ P$ is positive definite. Suppose there is a nonzero vector $(x_1, \ldots, x_{2p-1})$ such that

$$(x_1, \ldots, x_{2p-1})^T(P \circ P)(x_1, \ldots, x_{2p-1}) = 0.$$

This is equivalent to

$$\mathrm{tr}(XPXP) = 0,$$

with X being a diagonal matrix with $(x_1, \ldots, x_{2p-1})$ being the diagonal elements. Since P is a projection matrix, we have

$$0 = \mathrm{tr}(XPXP) = \mathrm{tr}(XPPXPP) = \mathrm{tr}(PXPPXP) = \mathrm{tr}(CC),$$

with $C = PXP$. Let C_j be the j^{th} column vector of C, and the j^{th} diagonal element of CC is actually $C_j^T C_j$, which is non-negative. Since the sum of all the diagonal elements of CC is 0, this implies that all the diagonal elements of CC are 0 and hence all the entries of C_j are 0, and therefore $C = \mathbf{0}$. Since X is a diagonal matrix, it can be decomposed as Y^TY with Y

being a diagonal matrix. Hence $PY^TYP = \mathbf{0}$, which implies $PY = \mathbf{0}$ by similar arguments we just used in showing $C = \mathbf{0}$. This implies that all the vectors e_i with the i^{th} element being 1 and all the other elements being 0 is orthogonal to all the row vectors of $BD^{1/2}$, which does not hold by inspecting the structure of B for a bifurcating tree. Therefore, $P \circ P$ is positive definite. Hence the eigenvalues of $P \circ P$ are in $(0, 1]$ and thus $\rho(D^{-1}H^TDH) \in [0, 1)$.

□

BIBLIOGRAPHY

Adachi, J. and M. Hasegawa (1996). *MOLPHY version 2.3: programs for molecular phylogenetics based on maximum likelihood*. Number 28. Institute of Statistical Mathematics Tokyo.

Amenta, N., M. Datar, A. Dirksen, M. de Bruijne, A. Feragen, X. Ge, J. H. Pedersen, M. Howard, M. Owen, J. Petersen, et al. (2015). Quantification and visualization of variation in anatomical trees. In *Research in Shape Modeling*, pp. 57–79. Springer.

Balakrishnan, S., M. J. Wainwright, B. Yu, et al. (2017). Statistical guarantees for the em algorithm: From population to sample-based analysis. *The Annals of Statistics 45*(1), 77–120.

Benjamini, Y. (2010). Discovering the false discovery rate. *Journal of the Royal Statistical Society: Series B (Statistical Methodology) 72*(4), 405–416.

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

Benjamini, Y. and D. Yekutieli (2005). False discovery rate-adjusted multiple confidence intervals for selected parameters. *J. Amer. Statist. Assoc. 100*(469), 71–93. With comments and a rejoinder by the authors.

Berger, J. (1980). A robust generalized Bayes estimator and confidence region for a multivariate normal mean. *Ann. Statist. 8*(4), 716–761.

Blomberg, S. P., T. Garland Jr, and A. R. Ives (2003). Testing for phylogenetic signal in comparative data: behavioral traits are more labile. *Evolution 57*(4), 717–745.

Bravo, H. C., S. Wright, K. Eng, S. Keles, and G. Wahba (2009). Estimating tree-structured covariance matrices via mixed-integer programming. In *Artificial Intelligence and Statistics*, pp. 41–48.

Brawand, D., M. Soumillon, A. Necsulea, P. Julien, G. Csárdi, P. Harrigan, M. Weier, A. Liechti, A. Aximu-Petri, M. Kircher, et al. (2011). The evolution of gene expression levels in mammalian organs. *Nature 478*(7369), 343.

Camin, J. H. and R. R. Sokal (1965). A method for deducing branching sequences in phylogeny. *Evolution 19*(3), 311–326.

Casella, G. and J. T. Hwang (1986). Confidence sets and the Stein effect. *Comm. Statist. A—Theory Methods 15*(7), 2043–2063.

Chatterjee, S., P. Lahiri, and H. Li (2008). Parametric bootstrap approximation to the distribution of EBLUP and related prediction intervals in linear mixed models. *Ann. Statist. 36*(3), 1221–1245.

Dobriban, E., K. Fortney, S. K. Kim, and A. B. Owen (2015). Optimal multiple testing under a gaussian prior on the effect sizes. *Biometrika 102*(4), 753–766.

Drton, M., C. Fox, A. Käufl, and G. Pouliot (2018, May). The Maximum Likelihood Threshold of a Path Diagram. *ArXiv e-prints*.

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

Efron, B. (2016). Empirical Bayes deconvolution estimates. *Biometrika 103*(1), 1–20.

Efron, B. and C. Morris (1973). Stein's estimation rule and its competitors—an empirical Bayes approach. *J. Amer. Statist. Assoc. 68*, 117–130.

Evans, S. N., B. B. Hansen, and P. B. Stark (2005). Minimax expected measure confidence sets for restricted location parameters. *Bernoulli*, 571–590.

Farchione, D. and P. Kabaila (2008). Confidence intervals for the normal mean utilizing prior information. *Statist. Probab. Lett. 78*(9), 1094–1100.

Felsenstein, J. (1973). Maximum-likelihood estimation of evolutionary trees from continuous characters. *American journal of human genetics 25*(5), 471.

Felsenstein, J. (1981). Evolutionary trees from DNA sequences: a maximum likelihood approach. *Journal of molecular evolution 17*(6), 368–376.

Felsenstein, J. (2004). *Inferring phylogenies*, Volume 2. Sinauer associates Sunderland, MA.

Ferguson, T. S. (1967). *Mathematical statistics: A decision theoretic approach*. Probability and Mathematical Statistics, Vol. 1. Academic Press, New York-London.

Friedman, N., M. Ninio, I. Pe'er, and T. Pupko (2002). A structural EM algorithm for phylogenetic inference. *Journal of Computational Biology 9*(2), 331–353.

Gelman, A. and J. Carlin (2014). Beyond power calculations: Assessing type s (sign) and type m (magnitude) errors. *Perspectives on Psychological Science 9*(6), 641–651.

Gelman, A. and F. Tuerlinckx (2000). Type s error rates for classical and bayesian single and multiple comparison procedures. *Computational Statistics 15*(3), 373–390.

Genovese, C. and L. Wasserman (2004). A stochastic process approach to false discovery control. *Annals of Statistics*, 1035–1061.

Gross, E. and S. Sullivant (2018, 02). The maximum likelihood threshold of a graph. *Bernoulli 24*(1), 386–407.

Guo, W. and J. P. Romano (2015). On stepwise control of directional errors under independence and some dependence. *J. Statist. Plann. Inference 163*, 21–33.

Guo, W., S. K. Sarkar, and S. D. Peddada (2010). Controlling false discoveries in multidimensional directional decisions, with applications to gene expression data on ordered categories. *Biometrics 66*(2), 485–492.

He, K. (1992). Parametric empirical Bayes confidence intervals based on James-Stein estimator. *Statist. Decisions 10*(1-2), 121–132.

Hoff, P. D. and C. Yu (2017). Exact adaptive confidence intervals for linear regression coefficients. *arXiv preprint arXiv:1705.08331*.

Hom, R. A. and C. R. Johnson (1991). Topics in matrix analysis. *Cambridge UP, New York*.

Hwang, J., J. Qiu, and Z. Zhao (2009). Empirical Bayes confidence intervals shrinking both means and variances. *Journal of the Royal Statistical Society: Series B (Statistical Methodology) 71*(1), 265–285.

Hwang, J. T. G. and Z. Zhao (2013). Empirical Bayes confidence intervals for selected parameters in high-dimensional data. *J. Amer. Statist. Assoc. 108*(502), 607–618.

James, W. and C. Stein (1961a). Estimation with quadratic loss. In *Proceedings of the fourth Berkeley symposium on mathematical statistics and probability*, Volume 1, pp. 361–379.

James, W. and C. Stein (1961b). Estimation with quadratic loss. In *Proc. 4th Berkeley Sympos. Math. Statist. and Prob., Vol. I*, pp. 361–379. Berkeley, Calif.: Univ. California Press.

Joshi, V. (1967). Inadmissibility of the usual confidence sets for the mean of a multivariate normal population. *The Annals of Mathematical Statistics 38*(6), 1868–1875.

Joshi, V. (1969). Admissibility of the usual confidence sets for the mean of a univariate or bivariate normal population. *The Annals of Mathematical Statistics 40*(3), 1042–1067.

Kabaila, P. and D. Tissera (2014). Confidence intervals in regression that utilize uncertain prior information about a vector parameter. *Australian & New Zealand Journal of Statistics 56*(4), 371–383.

Koenker, R. and I. Mizera (2014). Convex optimization, shape constraints, compound decisions, and empirical Bayes rules. *J. Amer. Statist. Assoc. 109*(506), 674–685.

Kozubowski, T. J. and K. Podgorski (2000). Asymmetric laplace distributions. *Mathematical Scientist 25*(1), 37–46.

Kruskal, J. B. (1956a). On the shortest spanning subtree of a graph and the traveling salesman problem. *Proceedings of the American Mathematical Society 7*(1), 48–50.

Kruskal, J. B. (1956b). On the shortest spanning subtree of a graph and the traveling salesman problem. *Proceedings of the American Mathematical society 7*(1), 48–50.

Laird, N. M. and T. A. Louis (1987). Empirical Bayes confidence intervals based on bootstrap samples. *J. Amer. Statist. Assoc. 82*(399), 739–757. With discussion and with a reply by the authors.

Lauritzen, S., C. Uhler, and P. Zwiernik (2017, February). Maximum likelihood estimation in Gaussian models under total positivity. *ArXiv e-prints*.

Lauritzen, S. L. (1996). *Graphical models*, Volume 17. Clarendon Press.

Lee, J. D., D. L. Sun, Y. Sun, J. E. Taylor, et al. (2016). Exact post-selection inference, with application to the lasso. *The Annals of Statistics 44*(3), 907–927.

Lee, J. D. and J. E. Taylor (2014). Exact post model selection inference for marginal screening. In *Advances in Neural Information Processing Systems*, pp. 136–144.

McLachlan, G. and T. Krishnan (2007). *The EM algorithm and extensions*, Volume 382. John Wiley & Sons.

Morris, C. N. (1983a). Parametric empirical Bayes confidence intervals. In *Scientific inference, data analysis, and robustness (Madison, Wis., 1981)*, Volume 48 of *Publ. Math. Res. Center Univ. Wisconsin*, pp. 25–50. Academic Press, Orlando, FL.

Morris, C. N. (1983b). Parametric empirical bayes inference: theory and applications. *Journal of the American Statistical Association 78*(381), 47–55.

Münkemüller, T., S. Lavergne, B. Bzeznik, S. Dray, T. Jombart, K. Schiffers, and W. Thuiller (2012). How to measure and test phylogenetic signal. *Methods in Ecology and Evolution 3*(4), 743–756.

O'Meara, B. C., C. Ané, M. J. Sanderson, and P. C. Wainwright (2006). Testing for different rates of continuous trait evolution using likelihood. *Evolution 60*(5), 922–933.

Ou, C.-Y., C. A. Ciesielski, G. Myers, C. I. Bandea, C.-C. Luo, B. T. Korber, J. I. Mullins, G. Schochetman, R. L. Berkelman, A. N. Economou, et al. (1992). Molecular epidemiology of hiv transmission in a dental practice. *Science 256*(5060), 1165–1171.

Owen, A. B. (2016). Confidence intervals with control of the sign error in low power settings. *arXiv preprint arXiv:1610.10028*.

Pagel, M. (1999). Inferring the historical patterns of biological evolution. *Nature 401*(6756), 877.

Penny, D. and M. Hendy (1985). The use of tree comparison metrics. *Systematic Zoology 34*(1), 75–82.

Pratt, J. W. (1963). Shorter confidence intervals for the mean of a normal distribution with known variance. *The Annals of Mathematical Statistics 34*(2), 574–586.

Price, P. N., A. V. Nero, and A. Gelman (1996). Bayesian prediction of mean indoor radon concentrations for Minnesota counties. *Health Physics 71*(6), 922–936.

R Core Team (2018). *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing.

Ravikumar, P., M. J. Wainwright, G. Raskutti, B. Yu, et al. (2011). High-dimensional covariance estimation by minimizing l1-penalized log-determinant divergence. *Electronic Journal of Statistics 5*, 935–980.

Rencher, A. C. (2003). *Methods of multivariate analysis*, Volume 492. John Wiley & Sons.

Revell, L. J., L. E. González-Valenzuela, A. Alfonso, L. A. Castellanos-García, C. E. Guarnizo, and A. J. Crawford (2018). Comparing evolutionary rates between trees, clades and traits. *Methods in Ecology and Evolution 9*(4), 994–1005.

Roch, S. (2006). A short proof that phylogenetic tree reconstruction by maximum likelihood is hard. *IEEE/ACM Transactions on Computational Biology and Bioinformatics 3*(1), 92–94.

Saitou, N. and M. Nei (1987). The neighbor-joining method: a new method for reconstructing phylogenetic trees. *Molecular biology and evolution 4*(4), 406–425.

Scaduto, D. I., J. M. Brown, W. C. Haaland, D. J. Zwickl, D. M. Hillis, and M. L. Metzker (2010). Source identification in two criminal cases using phylogenetic analysis of hiv-1 dna sequences. *Proceedings of the National Academy of Sciences 107*(50), 21242–21247.

Shao, J. (2003). *Mathematical statistics* (Second ed.). Springer Texts in Statistics. Springer-Verlag, New York.

Shiers, N., P. Zwiernik, J. A. Aston, and J. Q. Smith (2016). The correlation space of gaussian latent tree models and model selection without fitting. *Biometrika 103*(3), 531–545.

Snijders, T. A. B. and R. J. Bosker (2012). *Multilevel analysis* (Second ed.). Sage Publications, Los Angeles, CA. An introduction to basic and advanced multilevel modeling.

Sokal, R. R. and P. H. Sneath (1961). Principles of numerical taxonomy.

Stein, C. (1956). Inadmissibility of the usual estimator for the mean of a multivariate normal distribution. Technical report, STANFORD UNIVERSITY STANFORD United States.

Stephens, M. (2016). False discovery rates: a new deal. *Biostatistics*, kxw041.

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

Storey, J. D. (2007). The optimal discovery procedure: a new approach to simultaneous significance testing. *Journal of the Royal Statistical Society: Series B (Statistical Methodology) 69*(3), 347–368.

Storey, J. D., J. E. Taylor, and D. Siegmund (2004). Strong control, conservative point estimation and simultaneous conservative consistency of false discovery rates: a unified approach. *Journal of the Royal Statistical Society: Series B (Statistical Methodology) 66*(1), 187–205.

Tseng, P. (2004). An analysis of the EM algorithm and entropy-like proximal point methods. *Mathematics of Operations Research 29*(1), 27–44.

Tseng, Y.-L. and L. D. Brown (1997). Good exact confidence sets for a multivariate normal mean. *Ann. Statist. 25*(5), 2228–2258.

Tukey, J. W. (1962, 03). The future of data analysis. *Ann. Math. Statist. 33*(1), 1–67.

Tukey, J. W. (1991). The philosophy of multiple comparisons. *Statistical Science 6*(1), 100–116.

University of California Museum of Paleontology (2018). Understanding evolution. `http://evolution.berkeley.edu/`, Last accessed on 2018-06-16.

Van der Vaart, A. W. (1998). *Asymptotic statistics*, Volume 3. Cambridge university press.

Wang, Z., Q. Gu, Y. Ning, and H. Liu (2014). High dimensional expectation-maximization algorithm: Statistical optimization and asymptotic normality. *arXiv preprint arXiv:1412.8729*.

Washburne, A. D., J. T. Morton, J. Sanders, D. McDonald, Q. Zhu, A. M. Oliverio, and R. Knight (2018). Methods for phylogenetic analysis of microbiome data. *Nature Microbiology 3*(6), 652.

Wasserman, L. and K. Roeder (2006). Weighted hypothesis testing. *arXiv preprint math/0604172*.

Weinstein, A., W. Fithian, and Y. Benjamini (2013). Selection adjusted confidence intervals with more power to determine the sign. *Journal of the American Statistical Association 108*(501), 165–176.

Wu, C. J. (1983). On the convergence properties of the EM algorithm. *The Annals of statistics*, 95–103.

Yang, Y., G. Pan, et al. (2015). Independence test for high dimensional data based on regularized canonical correlation coefficients. *The Annals of Statistics 43*(2), 467–500.

Yu, C. and P. D. Hoff (2017). Adaptive sign error control. *arXiv preprint arXiv:1801.00152*.

Yu, C. and P. D. Hoff (2018). Adaptive multigroup confidence intervals with constant coverage. *Biometrika 105*(2), 319–335.

Yu, K. and J. Zhang (2005). A three-parameter asymmetric laplace distribution and its extension. *Communications in Statistics Theory and Methods 34*(9-10), 1867–1879.

Zhao, H., S. D. Peddada, and X. Cui (2015). Mixed directional false discovery rate control in multiple pairwise comparisons using weighted p-values. *Biom. J. 57*(1), 144–158.

Zwiernik, P., C. Uhler, and D. Richards (2017). Maximum likelihood estimation for linear gaussian covariance models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology) 79*(4), 1269–1292.